APA Handbook of

Testing and Assessment in Psychology

APA Handbook of Testing and Assessment in Psychology, Vol. 3: Testing and Assessment in School Psychology and Education, edited by K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, and M. C. Rodriguez

APA Handbooks in Psychology

APA Handbook of Testing and Assessment in Psychology

VOLUME 3

Testing and Assessment in School Psychology and Education

Kurt F. Geisinger, *Editor-in-Chief*

Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodriguez, *Associate Editors*

American Psychological Association • Washington, DC

Published by
American Psychological Association
750 First Street, NE
Washington, DC 20002-4242
www.apa.org

To order
APA Order Department
P.O. Box 92984
Washington, DC 20090-2984
Tel: (800) 374-2721; Direct: (202) 336-5510
Fax: (202) 336-5502; TDD/TTY: (202) 336-6123
Online: www.apa.org/pubs/books/
E-mail: order@apa.org

In the U.K., Europe, Africa, and the Middle East, copies may be ordered from
American Psychological Association
3 Henrietta Street
Covent Garden, London
WC2E 8LU England

American Psychological Association Staff
Gary R. VandenBos, PhD, *Publisher*
Julia Frank-McNeil, *Senior Director, APA Books*
Theodore J. Baroody, *Director, Reference, APA Books*
Lisa T. Corry, *Project Editor, APA Books*

Typeset in Berkeley by Cenveo Publisher Services, Columbia, MD

Printer: United Book Press, Baltimore, MD
Cover Designer: Naylor Design, Washington, DC

Library of Congress Cataloging-in-Publication Data

APA handbook of testing and assessment in psychology / Kurt F. Geisinger, editor-in-chief ; Bruce A. Bracken . . . [et al.], associate editors.
v. cm. — (APA handbooks in psychology)
Includes bibliographical references and index.
Contents: v. 1. Test theory and testing and assessment in industrial and organizational psychology — v. 2. Testing and assessment in clinical and counseling psychology — v. 3. Testing and assessment in school psychology and education.
ISBN 978-1-4338-1227-9 — ISBN 1-4338-1227-4
1. Psychological tests. 2. Psychometrics. 3. Educational tests and measurements. I. Geisinger, Kurt F., 1951- II. Bracken, Bruce A. III. American Psychological Association. IV. Title: Handbook of testing and assessment in psychology.
BF176.A63 2013
150.28′7—dc23
2012025015

British Library Cataloguing-in-Publication Data
A CIP record is available from the British Library.

Printed in the United States of America
First Edition

DOI: 10.1037/14049-000

Contents

Volume 3: Testing and Assessment in School Psychology and Education

Editorial Board

PART I

SCHOOL PSYCHOLOGY

CHAPTER 1

PSYCHOLOGICAL ASSESSMENT BY SCHOOL PSYCHOLOGISTS: OPPORTUNITIES AND CHALLENGES OF A CHANGING LANDSCAPE

Jack A. Naglieri

The reliability and validity of information obtained from any psychological test is dependent on the scope and psychometric attributes of the instrument used. As in all areas of science, what psychologists discover depends on the quality of the instruments used and the information they provide as well as skillful interpretation of the test results. Better conceptualized instruments yield more accurate and informative data than do weaker instruments. Instruments that uncover more useful information about the individual being examined are more valid and ultimately better inform both researchers and clinicians. The tools school psychologists choose for diagnostic decision making substantially influence the reliability and validity of the information they obtain and the decisions they make. Simply put, the better the tool is, the more valid and reliable the decisions; the more useful the information obtained is, the better the services provided. In this chapter, some important issues regarding quality and effectiveness of the tools used in school psychology are discussed.

The purpose of this chapter is to discuss some important issues in school-based psychological and educational assessment. To capture the essence of the major changes occurring in the schools, the chapter is organized into three sections. The first section involves the role of intelligence tests in determining learning disability eligibility. Next, some changes in achievement testing are reviewed. Third, evaluation of social–emotional status is examined. Each of these areas has been influenced by a combination of federal legislation and changes in school psychological practice, as described by the National Association of School Psychologists (2010). The goal of this chapter is not to summarize all the changes that have recently occurred or to predict the outcomes of these changes but rather to summarize a few important issues related to the current state of the field and the apparent strengths and weaknesses of the various options.

INTELLIGENCE AND SPECIFIC LEARNING DISABILITIES

Controversy is not new to the construct of intelligence and its measurement (see Jensen, 1998). Arguments have raged about the nature of intelligence—is it one factor or multiple factors, are intelligence tests biased or not, what are the best ways to interpret test results, do children with specific disabilities have distinctive ability profiles, and do intelligence test scores have relevance beyond diagnostic classification (e.g., implications for instruction and treatment)? In recent years, the most important questions have centered on the utility of intelligence tests for evaluation and treatment of children suspected of having a specific learning disability (SLD). More important, although the construct of general intelligence has considerable empirical support (see Jensen, 1998, for a review), especially when measured by tests such as the Wechsler scales and the Stanford–Binet, the value of traditional intelligence tests for evaluation of children with SLD is less clear.

There is little doubt that the psychometric characteristics of the Wechsler and Binet tests, the oldest

DOI: 10.1037/14049-001
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

intelligence tests, have advanced considerably over the past 30 years (see O'Donnell, 2009, and Roid & Tippin, 2009, for summaries). The hallmark of their advancement has been improved psychometric qualities, including improved reliability, more representative normative samples, more attractive physical materials, and computer-assisted scoring and interpretive analysis. These improvements have provided clear advantages to traditional intelligence tests over their predecessors. Despite excellent psychometric qualities, the limitations of these traditional tests have been noted by many, particularly those related to the evaluation and classification of children with SLD.

One of the most important and hotly debated limitations, particularly relevant for school psychologists, is the diagnostic value and stability of Wechsler subtest profiles. The interpretation of subtest profiles is widely accepted by many practitioners and was encouraged in many influential textbooks (e.g., Kaufman, 1979). Over time, however, a series of compelling articles have been published that have questioned the stability of subtest profiles and strongly suggested that Wechsler subtest variability is ineffective for diagnosis (see McDermott, Fantuzzo, & Glutting, 1990). It is quite clear that traditional intelligence tests that were developed in the early 1900s as measures of general ability may not meet more modern purposes, especially for evaluation of SLDs. To clarify the role intelligence tests play in the evaluation of SLD, the definitions of this disorder and the limitations and strengths of the tests that are used should be understood. This chapter, therefore, provides a brief summary of the Individuals With Disabilities Education Improvement Act of 2004 (IDEA), assessment issues related to learning disabilities in school psychology, and some summative data on current test profiles.

Learning Disabilities Defined

In the schools, IDEA (2004) and related state laws define SLD. For those in the medical profession and many psychologists in independent practice, the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text revision; *DSM–IV–TR;* American Psychiatric Association, 2000) is often used. Both of these definitions involve evaluation of a child's cognitive abilities. The *DSM–IV–TR* bases diagnosis on an inconsistency between assessed ability and achievement (i.e., reading, math, or written expression or a nonspecified academic area) when that difference is not better accounted for by other life conditions, such as inadequate education, cultural or ethnic differences, impaired vision or hearing, or mental retardation. The *DSM–IV–TR* definition is based on documenting achievement scores on individually administered, standardized tests in reading, mathematics, or written expression that are substantially below that which would be expected for peers of comparable age, schooling, and level of intelligence. The size of the discrepancy should be at least 1 standard deviation if the intelligence test score might have been adversely influenced by an associated disorder in cognitive processing, a mental disorder, or the ethnic or cultural background of the individual and 2 standard deviations if not. More important, the learning disorder should significantly interfere with the student's reading, math, or writing (which can be quantified with a variety of achievement tests) or daily living (which is often more difficult to quantify). The *DSM–IV–TR* also recognizes or assumes that problems with cognitive processing (e.g., deficits in visual perception, linguistic processes, attention, or memory) may have preceded or be associated with the learning disorder (i.e., the underlying cause of the disorder).

An SLD as defined in IDEA (2004) has similarities to and differences from the definition used in the *DSM–IV–TR*. The similarities include academic failure not explained by inadequate education, cultural or ethnic differences, or impaired vision or hearing, mental retardation, or other disability. The differences between IDEA and the *DSM–IV–TR* include (a) the age range for which the definition applies, (b) the disability being described as a specific disability, and (c) the definition of the disability as a disorder in basic psychological processes. The IDEA definition is as follows:

> Specific learning disability means a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, which may manifest itself

> in an imperfect ability to listen, think, speak, read, write, spell, or to do mathematical calculations. The term includes such conditions as perceptual handicaps, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia. The term does not include children who have problems that are primarily the result of visual, hearing, or motor disabilities, or mental retardation, emotional disturbance, or of environmental, cultural, or economic disadvantage. (pp. 11–12)

The measurement of intelligence plays a key role in both approaches to SLD determination, and these differences have important implications for the use of intelligence. Both the *DSM–IV–TR* and IDEA (2004) definitions involve a comparison of ability with achievement, the so-called ability–achievement discrepancy model. This approach has been widely criticized for some time and is no longer considered effective (see Fletcher, Denton, & Francis, 2005; Meyer, 2000; Stanovich, 1994) and is no longer required under IDEA (Kavale, Kaufman, Naglieri, & Hale, 2005).

To go beyond the ability–achievement discrepancy, practitioners were encouraged to examine subtest profiles, expecting that this information could aid in eligibility determination and instructional planning. Kaufman (1979) was among the first to recognize the limitation of global ability scores and suggested that useful information about a child could be obtained by a careful, psychometrically defined examination of subtest scores. Over time, the idea of going beyond the Full Scale IQ and the difference between that global score and achievement has gained favor, but using subtest analysis has not. More recently, greater emphasis has been placed on theoretically guided interpretations as described by Naglieri (1999); Flanagan, Ortiz, Alfonso, and Mascolo (2002); and Hale and Fiorello (2004). Before these methods are described, an examination of profiles for intelligence test scales rather than subtests is provided.

Because intelligence tests play such an important role in SLD eligibility determination, it is important to ask the question, "Do intelligence tests yield scale profiles that are distinctive for children with SLDs?" Naglieri (1999, 2000) suggested that subtest profile analysis should be replaced by scale profile analysis so that diagnostic reliability could be increased and, more important, so that each scale should be clearly related to some theoretical ability construct. To examine this method of profile analysis, Naglieri and Goldstein (2011) provided an examination of intelligence test profiles for adolescents and adults with SLD on the basis of information provided in the respective test manuals or book chapters of Naglieri and Goldstein (2009). They found that traditional intelligence tests did not yield a pattern of scores on scales encompassing these tests that was distinct to any one type of disability. This chapter describes a broader analysis of scores based on samples of children ages 5 to 18 years.

The research on intelligence test scale profiles is summarized next with the goal of examining mean score patterns of the scales for children with reading failure. This review helps to determine whether ability tests show particular patterns for children with a SLD in reading decoding. This information could have important implications for understanding the cognitive characteristics of that clinical group, which might allow for possible diagnostic and intervention considerations (Naglieri, 1999). To compile data from various intelligence tests, several different sources were used. Reports in the technical manuals were used for the Wechsler Intelligence Scale for Children—Fourth Edition (Wechsler, 2003), Stanford–Binet—Fifth Edition (Roid, 2003), Differential Ability Scales—Second Edition (Elliott, 2007), Kaufman Assessment Battery for Children—Second Edition (Kaufman & Kaufman, 2004), and Cognitive Assessment System (CAS; Naglieri & Das, 1997). (The CAS data also included findings from Naglieri, Otero, DeLauder, & Matto, 2007.) The findings, however, must be taken with recognition that the samples were not matched across the various studies, the accuracy of the diagnosis may not have been verified, and some of the sample sizes were small. Notwithstanding these limitations, the findings provide important insights into the extent to which these various tests can be used for assessing adolescents and adults suspected of having a specific learning disorder.

The results of this analysis are provided in Figure 1.1, which includes the standard scores obtained on these various intelligence tests for students with a specific reading disability. The comparison of scale profiles across the various ability tests suggests that some tests are more sensitive to the cognitive characteristics of individuals with specific reading disabilities than others. The Differential Ability Scales—Second Edition, Stanford–Binet—Fifth Edition, and Kaufman Assessment Battery for Children—Second Edition showed relatively little variability among the scales; the differences between the lowest and the highest scale within each test were 3.2, 3.8, and 3.8, respectively. That is, the pattern of scores on the separate scales making up these tests did not suggest that a specific cognitive disorder was uncovered. The scales on the Wechsler Intelligence Scale for Children—Fourth Edition showed more variability (range = 7.4), followed by the Woodcock–Johnson III Tests of Achievement (Woodcock, McGrew, & Mather, 2007; range = 10) and the CAS (range = 10.3). More important, the

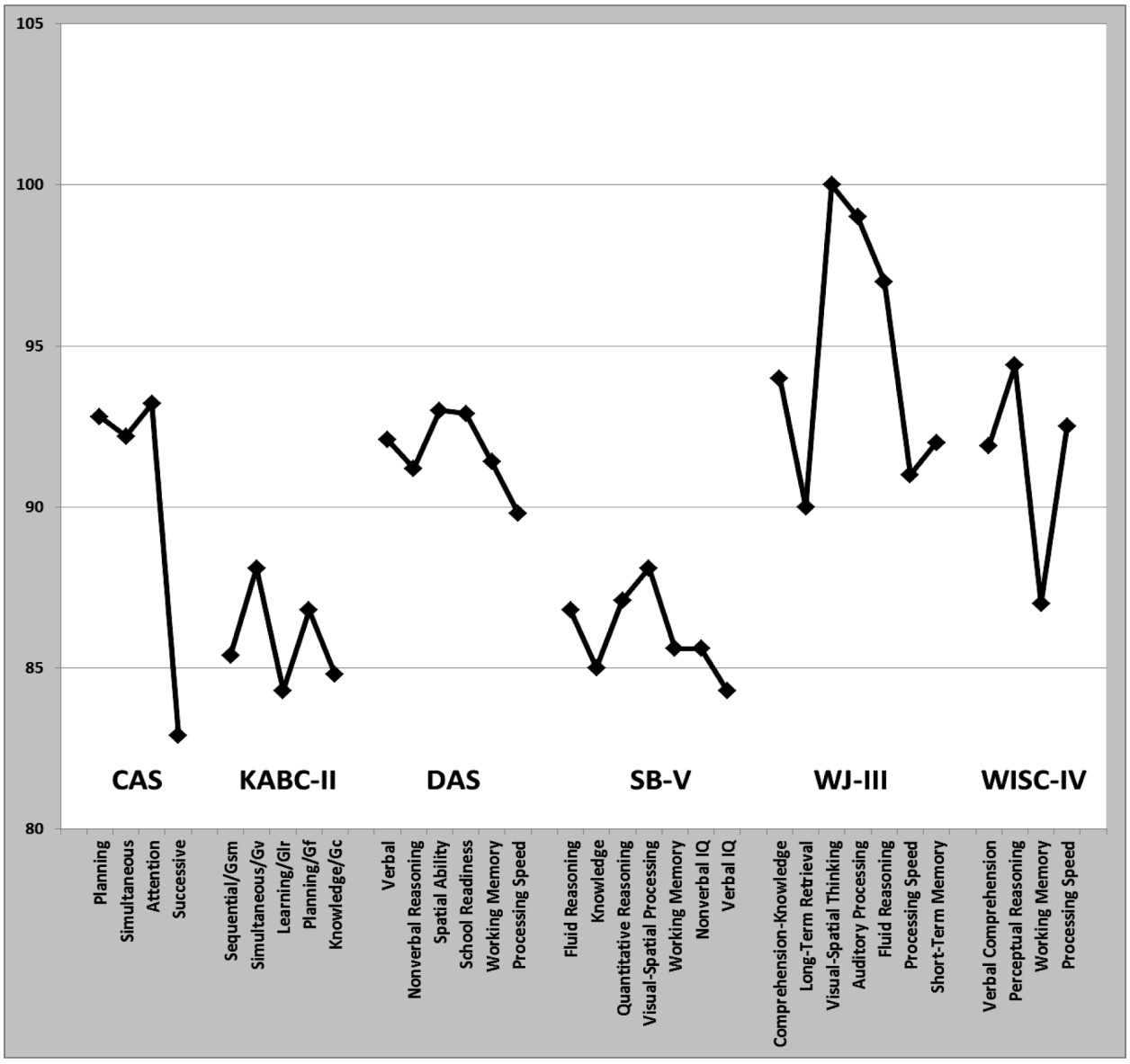


FIGURE 1.1. Mean scores earned by samples of students with reading decoding disorders by ability test. CAS = Cognitive Assessment System; KABC–II = Kaufman Assessment Battery for Children—Second Edition; DAS = Differential Ability Scales—Second Edition; SB–V = Stanford–Binet—Fifth Edition; WJ–III = Woodcock–Johnson III Tests of Cognitive Abilities; WISC–IV = Wechsler Intelligence Scale for Children—Fourth Edition.

lowest score (90) on the Woodcock–Johnson III was for Long Term Retrieval, which measures associative memory (Wendling, Mather, & Shrank, 2009). The Wechsler Intelligence Scale for Children—Fourth Edition profile suggests that the sample was low in Working Memory and the remaining scales were in the low end of the average range, such as for the CAS Planning, Simultaneous, and Attention Scales. Interestingly, the Working Memory tests on the Wechsler Intelligence Scale for Children—Fourth Edition require repetition of numbers in the order provided by the examiner (Digit Span Forward) or in the reverse order (Digit Span Backward) and recitation of numbers in ascending sequential order and letters in alphabetical order (Letter–Number Sequencing), both of which require sequencing. Schofield and Ashman (1986) showed that Digit Span Forward and Digit Span Backward correlated significantly with measures of successive processing as measured in the CAS.

The CAS showed the most variability (range = 10.3) even though three of the four scale means were within 1 point of each other. The exception was successive processing ability, on which the sample earned a very low score (82.9). The CAS profile for the sample with SLD suggested that this group had a specific academic (reading decoding) and a specific cognitive weakness (successive), meaning that as a group, these individuals had difficulty working with stimuli that are arranged in serial order, as in the sequence of sounds that make words, the sequence of letters to spell words, and the sequence of groups of sounds and letters that make words. Taken as a whole, these findings suggest that the tool with which practitioners choose to evaluate children suspected of having a SLD may or may not uncover a disorder in one or more of the basic psychological processes required in the IDEA (2004) definition.

Next Steps

The evaluation of children with a SLD is among the most complex and contentious issues facing the field of school psychology. Because IDEA (2004) specifies that children with SLD have a disorder in one or more of the basic psychological processes, cognitive processes must be measured (Kavale et al., 2005). A comprehensive evaluation of the basic psychological processes unites the statutory and regulatory components of IDEA and ensures that the methods used for identification more closely reflect the definition. Any defensible eligibility system would demand continuity between the statutory and regulatory definitions, and for this reason alone SLD determination requires the documentation of a basic psychological processing disorder (Hale, Kaufman, Naglieri, & Kavale, 2006). Moreover, the tools used for this assessment must meet the technical criteria included in IDEA, and well-validated measures of cognitive and neuropsychological measures are available that can be used to document SLD (Hale & Fiorello, 2004; Kaufman & Kaufman, 2001; Naglieri & Otero, 2011). To use a cognitive processing approach to SLD identification, three main components are needed. First, the child must have significant intraindividual differences among the basic psychological processes, with the lowest processing score substantially below average. Second, average processing scores and some specific area of achievement need to differ significantly. Third, consistency between poor processing scores and a specific academic deficit or deficits is essential (Hale & Fiorello, 2004; Naglieri, 1999, 2011). These systematic requirements are collectively referred to as a *discrepancy–consistency model* by Naglieri (1999, 2011) and as the *concordance–discordance model* by Hale and Fiorello (2004).

Naglieri (1999, 2011) described the discrepancy–consistency model for the identification of SLDs on the basis of finding a cognitive processing disorder (see Figure 1.2). The method involves a systematic examination of variability of basic psychological processes and academic achievement test scores. Determining whether cognitive processing scores differ significantly is accomplished using the method originally proposed by Davis (1959), popularized by Kaufman (1979), and modified by Silverstein (1993). This so-called ipsative method determines when the child's scores are reliably different from the child's average score. It is important to note that in the discrepancy–consistency model described by Naglieri (1999), the ipsative approach is applied to the scales that represent four neuropsychologically defined constructs, not subtests from a

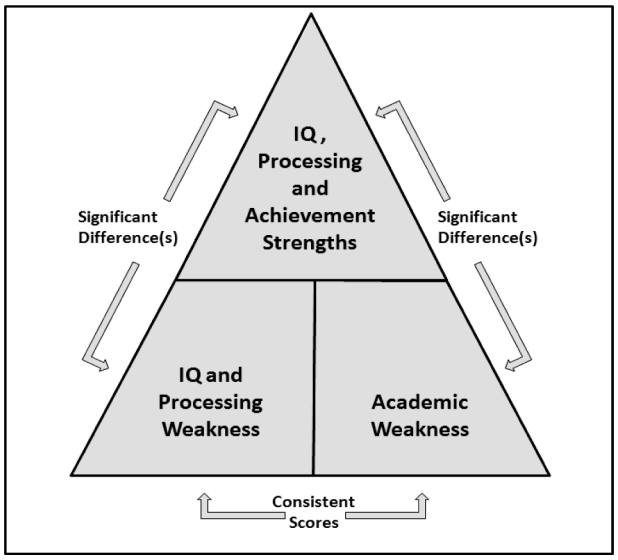


FIGURE 1.2. Naglieri's (1999, 2011) discrepancy–consistency model for determination of specific learning disability.

larger test of ability. This distinction is important because the criticisms of the ipsative method (McDermott et al., 1990) have centered on subtests, not scale-level analysis. In contrast, good evidence for the utility of using scales (from CAS) for diagnosis has been reported (see Canivez & Gaboury, 2010; Huang, Bardos, & D'Amato, 2010), and Huang et al. (2010) concluded that their study "substantiate[d] the usefulness of profiles analysis on composite scores as a critical element in LD [learning disability] determination" (p. 19).

Naglieri (1999) and Flanagan and Kaufman (2004) recognized that because a low score in relation to the child's scale mean could still be within the population's average range, adding the requirement that the weakness in a processing test score also be well below the average range is important. For example, Naglieri (2000) found that those students who had a low processing score relative to their personal mean and the normative group were likely to have significantly lower achievement scores and were more likely to have been identified as having an academic disability. That study was described by Carroll (2000) as one that illustrated a more successful profile methodology. Davison and Kuang (2000) suggested that "adding information about the absolute level of the lowest score improves identification over what can be achieved using ipsative profile pattern information alone" (p. 462). More important, this method has been shown to have implications for instruction for children with SLD (Naglieri & Gottling, 1995, 1997; Naglieri & Johnson, 2000) and attention deficit/hyperactivity disorder (Iseman & Naglieri, 2011) and to be tied to many instructional methods used in the classroom (Naglieri & Pickering, 2010).

Hale and Fiorello's (2004) proposed method, the concordance–discordance model, is based on the cognitive hypothesis testing methodology that relies on multiple assessment tools and data sources to maximize validity of assessment findings. Hale and Fiorello used cognitive and neuropsychological assessment data for both diagnostic and intervention purposes. When cognitive hypothesis testing results suggest that a child may have a SLD, differences among the scores is determined using the standard error of difference (Anastasi & Urbina, 1997) to test differences among the three components of the model: cognitive assets, cognitive deficits, and achievement deficits in standardized test scores. This approach has been advocated for use in school psychology by Hale and colleagues (Hale & Fiorello, 2004; Hale et al., 2006), who also cautioned that the method not be rigidly applied. They argued that practitioners follow the literature to ensure that the apparent cognitive strength is not typically related to the deficit achievement area and that the apparent cognitive weakness could explain the achievement deficit. This method ensures that children identified as having a SLD meet both IDEA (2004) requirements and, more important, has been shown to be relevant to intervention (Fiorello, Hale, & Snyder, 2006; Hale & Fiorello, 2004).

These two methods for identifying children with SLDs provide a means of uniting the definition found in IDEA (2004) with well-standardized tests that practitioners use on a regular basis. As the field of SLD evolves within the context of federal law and federal and state regulations, the applicability of these methods will become more apparent. Although initial research on the effectiveness of these methods for both eligibility determination and remediation of academic deficiencies is encouraging, additional studies are warranted.

ASSESSMENT OF ACHIEVEMENT

Achievement tests used by school psychologists are comprehensive, well-developed, and psychometrically refined tools. For example, tests such as the Kaufman Test of Educational Achievement—Second Edition (Kaufman & Kaufman, 2004; see Lichtenberger, Sotelo-Dynega, & Kaufman, 2009), the Wechsler Individual Achievement Test—Second Edition (Wechsler, 2005; see Choate, 2009), and the Woodcock–Johnson III Tests of Achievement (Woodcock et al., 2007; see Mather & Wendling, 2009) are high-quality instruments (see Naglieri & Goldstein, 2009, for descriptions of these and other tests of academic skills). These individually administered tests offer many content-dependent subtests with ample coverage of various aspects of academic achievement, excellent normative samples, and strong psychometric documentation. All of these tests provide age-corrected standard scores that calibrate a student's standing relative to his or her respective standardization groups. Some, for example, the Kaufman Test of Educational Achievement—Second Edition, offer the added advantage of item-level analysis so that the student's score can be more completely described on the basis of which academic skills have been acquired or are in need of instruction. Additionally, some offer excellent psychometric scores that can be used to monitor progress over time (e.g., Wide Range Achievement Test—Fourth Edition; see Roid & Bos, 2009).

Traditional measures of academic skills have been challenged by proponents of the curriculum-based measurement (CBM) field, including two tests representing CBM methodology. Traditional and CBM assessments differ in that CBM measures are used more as universal screening tools for identifying poor readers in early elementary grades and for monitoring academic progress when evaluating the effectiveness of instructional methods. In school psychology, these very brief alternative measures are used for evaluating reading for universal screening and monitoring student progress and as part of the evaluation process for SLD eligibility determination, sometimes in lieu of a comprehensive assessment of academic skills (e.g., Koehler-Hak & Bardos, 2009). This alternative approach to academic assessment is perhaps best illustrated by brief fluency tests and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002).

The approach to testing and monitoring of progress exemplified by the DIBELS differs from that of the traditional achievement tests mentioned earlier as well as from that of group-administered measures of achievement (e.g., the Stanford Achievement Test—10th Edition (Pearson, 2006) that can be used for universal screening. Tests such as the Stanford Achievement Test—10th Edition and many individually administered achievement tests (a) are nationally normed on a representative sample of students, (b) cover many different aspects of reading and math curriculum, (c) are based on appropriate learning standards, and (d) yield age-corrected standard scores. These tests can be used to identify students at risk of academic failure on the basis of a comparison to national norms as well as ranking within the classroom, school, or school district. The greatest difference between a test such as the DIBELS and more traditional achievement tests is the brevity of CBM assessments and the shift toward measures that come from the CBM field.

CBM procedures are intended to give educators tests that are reliable, valid, inexpensive, and efficient estimates of student achievement. Researchers have generally found consistency in the relationship between CBM scores and standardized measures across samples and various achievement tests as well as acceptable levels of reliability and validity (Reschly, Busch, Betts, Deno, & Long, 2009). For this reason, these brief tests (e.g., correct words read per minute) have been used to identify children at risk of reading failure and to assess student progress over time. The main differences between tests from the CBM field and traditional achievement tests rest on the CBM assumption that a brief measure of achievement is as effective as a comprehensive, standardized measure of current and future academic performance. So instead of measuring reading comprehension, for example, a 1-minute CBM reading fluency test is used because it correlates moderately with reading comprehension as measured by tests such as the Stanford Achievement Test—10th Edition or the Wechsler Individual Achievement Test—Second Edition. Another important difference is that

the CBM measures do not yield scores that are calibrated against a national norm and are not corrected for age effects.

Those who advocate for the use of CBM place emphasis on the goals of identifying children at risk for academic failure and monitoring academic progress over time to determine instructional effectiveness. The psychometric methods used, however, raise several important concerns that have been largely ignored by CBM advocates. These issues include the publication and use of tests without technical manuals that explicate the psychometric quality of the scores the tests yield (e.g., reliability and validity) and, perhaps most important, norms. The use of raw scores as measures of current status and as a means of calibrating current standing and response to intervention is another important difference between CBM measures and traditional normed measures of achievement. In this chapter, I ill focus on issues related to the use of raw scores from the DIBELS Oral Reading Fluency (ORF) test. The first topic concerns the use of raw scores to identify which students may be at risk of academic failure; the second concerns the use of raw scores to monitor progress over time for an individual child; and the third concerns the use of raw scores for the purpose of examining changes in groups of students as a function of some intervention.

Identifying At-Risk Students

To illustrate some of the problems with using raw scores instead of age-corrected standard scores, I show a simple examination. Figure 1.3 shows raw scores that are used as benchmarks (Koehler-Hak & Bardos, 2009) for making decisions about students' academic standing in a classroom. These values are approximately associated with the 40th percentile for the DIBELS ORF. This figure was developed by finding which raw scores were associated with the 40th-percentile scores during the fall, winter, and spring of second grade according to Table 7 in the DIBELS Technical Report Number 9 (Good, Wallin, Simmons, Kame'enui, & Kaminski, 2002). According to the technical report, the transformation of raw scores to "system-wide percentile ranks . . . [was] based on all participating students in the DIBELS Web data system as of May 20, 2002" (Good et al., 2002, p. 2). The test authors used raw scores from the 2001 to 2002 academic year to obtain percentile scores for children tested in the beginning, middle, and end of the year. Good et al.'s (2002) description of the sample used to obtain the raw-score-to-percentile-score conversion is very limited, and they do not indicate whether their sample is representative of the U.S. population. Although this sample description does not meet commonly accepted standards for reporting reference groups for commonly used standardized achievement tests, I used the conversion tables to approximate normal maturation rates in ORF scores. The results provided in Figure 1.3 show that DIBELS ORF raw scores can increase dramatically over the course of a school year while the percentile score associated with the score remains the same. This implies that even though a student can read more words per minute, his or her relative standing has not improved. This situation leaves the user in a quandary: How exactly should examiners interpret raw scores on the DIBELS?

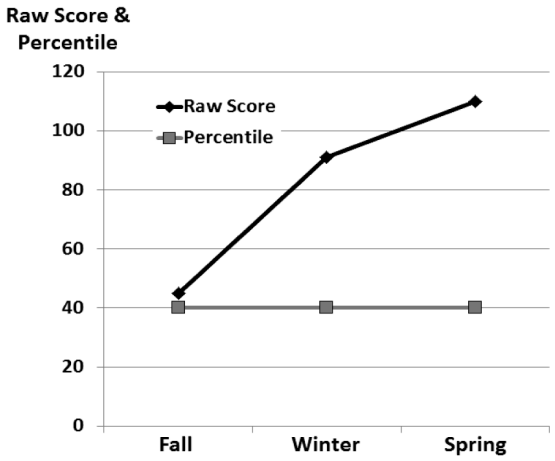


FIGURE 1.3. Relationships between Dynamic Indicators of Basic Early Literacy Skills raw scores and percentile ranks over three points in time.

Some researchers (e.g., Shinn, Tindal, & Stein, 1988) have advocated for the use of local norms to determine whether a child's academic needs are being met in the classroom or whether a referral for special services is appropriate. The apparent expectation is that local norms can help school psychologists make sound data-based decisions and more accurately identify students at risk of academic

failure. As an example, I present constructed local norms for DIBELS ORF scores for nine schools (N = 620) and the results across schools and in comparison to the contrast group provided by Good et al. (2002). The data used for this illustration came from a medium-sized city in the mid-South region of the United States. Local norms were constructed by transforming raw scores to z scores and then to standard scores of an IQ metric (M = 100, SD = 15) on the basis of raw-score DIBELS ORF means and standard deviations for each school. In addition to local norms, I converted DIBELS raw scores to standard scores (M = 100, SD = 15) via the percentile ranks provided by Good et al. That is, raw scores were converted to percentile ranks on the basis of conversion Table 7 in Good et al. Next, I converted percentiles to standard scores (M = 100, SD = 15) using the statistical function NORMINV from Microsoft Excel. This procedure provided a means of comparing local norms with those of a national comparison group (assuming, however, that there is no evidence that this group represents the U.S. population). The findings are quite revealing.

As seen in Table 1.1, the nine schools' mean ORF scores varied considerably, as did minority representation and percentages of students on free or reduced lunch programs. The mean number of words per minute was highest for the school with the least percentage of students receiving free or reduced lunch and the smallest number of minority students. The raw scores corresponding to standard scores are provided in Table 1.2 and show that the standard score a child would earn for the same raw score varies considerably across the nine schools. For example, a raw score of 20 words per minute yields a standard score of 101 for a child in School 1 but a standard score of 86 for a student in School 9. This considerable difference would ensure inequity of assessment and faulty interpretations within the same school district. The problem is that those students in schools in which the raw score mean is lowest will earn scores that are average, implying that no deficiency was found. Even more concerning is that the students who earn a raw score of 20 on the basis of the local norm are actually well below the national reference group, which earned a standard score of 84. In fact, the students in Schools 1, 2, and 3 earned local standard scores of 101, 98, and 98, respectively, but when compared with the national reference group would have earned a standard score of 84 (more than 1 standard deviation below the mean).

The only logical conclusion drawn from this analysis is that local norms mislead the user into thinking that students (as well as teachers and curricula) are doing well when in fact they may be well below what would be considered normal or expected on a national basis. In this illustration, the schools with the lowest scores were those with the highest percentage of minority children. Because these students earn high scores when local norms

TABLE 1.1

Demographic Characteristics of Nine Schools Used to Create Local Norms for DIBELS Measure of Oral Reading Fluency

	Schools								
Characteristic	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
M	19.17	23.47	23.70	26.55	27.61	33.61	35.30	43.08	60.59
SD	20.37	22.88	22.97	25.30	29.08	30.15	26.35	34.14	44.04
N	63	72	90	77	57	83	61	84	96
% Black	83	91	16	21	12	56	87	51	22
% Hispanic	2	3	42	29	54	8	4	4	1
% White	14	6	40	48	27	34	8	39	73
% Other	1	0	2	2	7	2	1	6	4
% Free or reduced lunch	99	99	99	65	99	56	99	56	25

Note. DIBELS = Dynamic Indicators of Basic Early Literacy Skills.

TABLE 1.2

Calibration of Standard Scores Using Local Norms by School and Using DIBELS Reference Group for Fall of Second Grade

	Standard scores for each school ($M = 100$, $SD = 15$)									National
Raw score	1	2	3	4	5	6	7	8	9	reference
60	130	124	124	120	117	113	114	107	100	102
58	129	123	122	119	116	112	113	107	99	101
56	127	121	121	117	115	111	112	106	98	100
54	126	120	120	116	114	110	111	105	98	100
52	124	119	118	115	113	109	110	104	97	99
50	123	117	117	114	112	108	108	103	96	98
48	121	116	116	113	111	107	107	102	96	98
46	120	115	115	112	109	106	106	101	95	97
44	118	113	113	110	108	105	105	100	94	96
42	117	112	112	109	107	104	104	100	94	95
40	115	111	111	108	106	103	103	99	93	94
38	114	110	109	107	105	102	102	98	92	93
36	112	108	108	106	104	101	100	97	92	93
34	111	107	107	104	103	100	99	96	91	92
32	109	106	105	103	102	99	98	95	90	91
30	108	104	104	102	101	98	97	94	90	90
28	106	103	103	101	100	97	96	93	89	89
26	105	102	102	100	99	96	95	92	88	87
24	104	100	100	98	98	95	94	92	88	86
22	102	99	99	97	97	94	92	91	87	85
20	101	98	98	96	96	93	91	90	86	84
18	99	96	96	95	95	92	90	89	85	82
16	98	95	95	94	94	91	89	88	85	82
14	96	94	94	93	93	90	88	87	84	80
12	95	92	92	91	92	89	87	86	83	78
10	93	91	91	90	91	88	86	85	83	75
8	92	90	90	89	90	87	84	85	82	74
6	90	89	88	88	89	86	83	84	81	72
4	89	88	87	87	88	85	82	83	81	69
2	87	87	86	85	87	84	81	82	80	65

Note. DIBELS = Dynamic Indicators of Basic Early Literacy Skills.

are used, fewer would be identified as being in need of instruction and fewer would be provided good instruction as a result. This approach would, therefore, result in considerable educational inequality.

Monitoring Progress

An essential goal of the CBM approach is the examination of changes over time. Research studies that seek to evaluate the effects of educational interventions need to consider the potential confound resulting from the issue of natural maturation in numbers of words a student can read per minute. That is, evaluation research often involves comparison of pretest and posttest scores, but with raw scores, such as those for reading fluency, it is unclear how much of the pretest–posttest change is associated with the intervention and how much is attributable to normal growth and learning. The value of the normalizing raw score to percentile rank conversion is that age-related changes are controlled because the student's score is calibrated in relation to a similarly aged comparison group. However, to avoid the problems associated with analyzing percentile ranks (see Anastasi & Urbina, 1997), converting percentiles to standard scores retains rank and improves the psychometric qualities of the scores. To better

understand the influence raw scores can have on evaluation of pretest–posttest treatment, I examined scores analysis of a data set containing DIBELS ORF scores in an intervention study.

The data used included 352 second-grade boys and girls enrolled in 12 public elementary schools located in the southern Atlantic region of the United States. The experimental ($N = 136$) and the control ($N = 216$) groups consisted of students from six schools whose reading skills were tested at the start of school and in the middle of the school year. The students in the experimental group participated in an online reading program called Ramps to Reading (see Naglieri & Pickering, 2010, for a description), and the control group was not exposed to Ramps to Reading at all. The demographic characteristics of the schools and students in the experimental and control groups were similar, but the schools that made up the experimental group were represented by high percentages of individuals receiving free and reduced lunch and African Americans. All schools in the experimental group met criteria for Title I funding. What is most important, however, is that in this study both ORF raw scores and standard scores (obtained by converting the percentile scores to standard scores having a mean of 100 and a standard deviation of 15) were obtained for the groups. Additionally, each group was further divided on the basis of having initial DIBELS ORF scores that were described as low risk, some risk, or at risk. The results were quite informative.

Table 1.3 provides the means, standard deviations, sample sizes, and effect sizes for the various groups using ORF raw scores and standard scores. For the experimental group, one sees effect sizes based on raw scores of 1.5, 2.8, and 1.7 for the low-risk, some-risk, and at-risk groups. These values are very large, as are the effect sizes for the control group (1.1, 2.2, and 1.2). Because growth in raw scores of words read per minute over the course of time is considerable, as previously discussed (see Figure 1.2), these effect sizes can be considered to be inflated. Examination of the standard scores, which calibrate standing relative to a comparison group, suggest a far different result. The control group's pretest–posttest differences were essentially zero, but the experimental group showed small to medium effect-size changes. Thus, the method of calibration of the raw scores had a direct impact on the interpretation of the intervention's effectiveness. Using scores that take into account developmental changes that occur over time inflated the effect sizes for both groups, and only when age-related changes were controlled did a more realistic finding result.

What Now?

The current state of achievement testing in school psychology can be described as having too much variance. The psychometric quality of measures used today ranges from marginal to excellent. As with assessment of other constructs, disorders, and

TABLE 1.3

Effectiveness of a Reading Intervention on the Basis of Comparisons of Raw Scores and Standard Scores

		Raw scores					Standard scores				
		Pretest		Posttest			Pretest		Posttest		
Group	*N*	*M*	*SD*	*M*	*SD*	Effect size	*M*	*SD*	*M*	*SD*	Effect size
Experimental											
Low risk	61	67.7	19.6	98.4	21.8	1.5	104.5	6.8	106.8	7.7	0.3
Some risk	37	33.8	4.7	60.3	12.7	2.8	91.5	2.3	93.4	4.4	0.6
At risk	38	16.1	6.8	33.2	12.3	1.7	80.4	5.7	83.4	5.3	0.5
Control											
Low risk	145	71.4	19.8	93.0	20.7	1.1	105.9	7.0	105.0	7.4	–0.1
Some risk	43	33.5	5.2	53.3	11.8	2.2	91.3	2.5	91.1	4.1	–0.1
At risk	28	15.4	8.7	27.5	11.9	1.2	79.3	7.6	80.6	5.8	0.2

abilities, for example, autism spectrum disorders (see Naglieri & Chambers, 2009), the options range considerably, and practitioners have to choose wisely between tools to obtain scores they can use with confidence. As far as the information provided in this section is concerned, practitioners must be particularly cautious when using very short measures of skills, such as using reading fluency as a predictor of reading, and when using raw scores for (a) evaluating current status and (b) evaluating changes over time. The best option remains using well-normed tests that assess academic skills directly, especially those that provide strong psychometric quality and norms for calibrating growth. Put simply, the use of raw scores for the calibration of academic skills and progress monitoring is not good science.

SOCIAL–EMOTIONAL STATUS

Evaluation of emotional status has been dominated by projective tests and rating scales. As with the assessment of achievement and intelligence, the evaluation of emotional well-being has also been evolving in the areas of both individual and universal assessment. This evolution has been driven by efforts to focus on positive attributes (so-called social–emotional strengths related to resilience) instead of, or in addition to, emotional or behavioral disorders and psychopathology. Emphasis on social–emotional strengths that are related to resilience and particularly on universal screening has come from governmental agencies, professional organizations, and practitioners in fields such as psychology, sociology, and education. For example, in 2003 the President's New Freedom Commission on Mental Health urged that early mental health screening and assessment services be routinely conducted and that school district personnel ensure the mental health care of children. In 2010, the National Association of School Psychologists published its model for comprehensive and integrated school psychological services, which addressed the delivery of school psychological services within the context of educational programs and educational settings. The model states that school psychologists should have knowledge of principles and research related to resilience and risk factors that are important for learning and mental health. Additionally, the National Association of School Psychologists' position contended that school psychologists should be involved in universal screening programs to identify students in need of support services to ensure learning and promote social–emotional skills and resilience. Clearly, the field is moving toward assessment of social–emotional strengths as well as psychopathology.

The emphasis on assessment and interventions for social–emotional competence is important for several reasons. First, at any given time about 20% of children and adolescents are estimated to have a diagnosable emotional or behavioral disorder that interferes with learning (Doll, 1996). Second, emerging research has suggested that social–emotional competence underlies school success (Payton et al., 2008). Third, state departments of education have adopted or are in the process of developing social–emotional learning standards that could lead to (a) universal screening of social–emotional skills and (b) social–emotional skills instruction within the regular education curriculum. This approach, as with any assessment and intervention approach, requires reliable and valid tools for assessing and monitoring social–emotional competencies (see Goldstein & Brooks, 2005).

Progress has been made in recent years as evidenced by the availability of published rating scales to measure protective factors that measure children's social–emotional strengths related to resilience. Sometimes social–emotional strengths have been integrated into scales that also include problem behaviors related to emotional or behavioral disturbance. For example, Bracken and Keith (2004) included specific scales related to serious emotional disturbance and social maladjustment as well as both clinical and adaptive (e.g., social skills) scales using items that are designed to identify children and adolescents in need of behavioral, educational, or psychiatric treatments. Similarly, the Behavior Assessment System for Children, Second Edition (Reynolds & Kamphaus, 2004), measures adaptive and maladaptive behavior. Using all positively worded items for assessment of social–emotional strengths and behavioral needs, LeBuffe and Naglieri (2003) published the Devereux Early Childhood

Assessment—Clinical Form. These three scales illustrate how measures of social–emotional problems and strengths can be combined into one rating scale. The authors of other scales, however, conceptualized evaluation of mental health using a different approach—assessment of social–emotional factors related to resilience.

The Resiliency Scales for Children and Adolescents (Prince-Embury, 2005) measure areas of perceived strength and vulnerability related to psychological resilience along three dimensions (sense of mastery, sense of relatedness, and emotional reactivity). Using a similar approach, researchers at the Devereux Center for Resilient Children have published a comprehensive system made up of several measures of factors related to resilience that vary across ages and purposes. For example, the Devereux Early Childhood Assessment for Infants and Toddlers (Mackrain, LeBuffe, & Powell, 2007) and the Devereux Early Childhood Assessment (LeBuffe & Naglieri, 1999) are designed to measure social–emotional strengths of young children. The Devereux Student Strengths Assessment (LeBuffe, Shapiro, & Naglieri, 2009) was developed for children from kindergarten through eighth grade; each of these is a thorough measure with many items. In contrast, the Devereux Student Strengths Assessment—Mini (Naglieri, LeBuffe, & Shapiro, 2011) is an eight-item scale of social–emotional strengths for universal screening. The availability of carefully developed measures of protective factors related to resilience offers the opportunity to examine validity questions related to these new instruments, especially as they may be used for universal screening.

The availability of new scales built on the concept of social–emotional strengths using a perspective described as strengths based is clearly an important development in the assessment of mental health. An evolution has also occurred in the assessment of psychological and behavioral disorders, especially as it relates to the use of raw scores and comparison groups, as discussed earlier in this chapter for achievement tests. More specifically, in some contexts, for example, identification of specific psychological disorders such as autism, researchers are using a specific reference group for calibration of scores instead of using a nationally representative reference group.

Naglieri and Chambers (2009) summarized the characteristics of rating scales used to assess behaviors associated with autism and examined the psychometric qualities that such measures possess. They concluded that the methods used to develop rating scales differed considerably in their approaches to instrument development. For example, some of the scales are very short (e.g., 15 items), and others contain many items (e.g., about 90 items). Some authors provided only raw scores, which makes interpretation difficult, and only two scales provided standard scores (*T* scores). Although some rating scales provide derived scores, the samples on which they were based were the particular group the scale was intended to identify. Raters obtained a score that tells how similar the individual being assessed is to those the scale is intended to identify, for example, those with an autism spectrum disorder (ASD). Of all the scales Naglieri and Chambers summarized, only one used a national comparison sample; all the others used samples of individuals who had or were referred for autism. The question of the utility of a comparison group consisting of children referred for or having the disorder of interest needs to be addressed. I consider this issue next using data from a recent project involving the Autism Spectrum Rating Scale (Goldstein & Naglieri, 2009).

An understanding of the differences between using a nationally representative sample and a sample of children identified as having autism as a reference group is best examined empirically. To do so, Goldstein and Naglieri (2009) constructed a raw-score-to-standard-score (*T*-scores) conversion table on the basis of a sample of children with ASD ($N = 243$) who were diagnosed with autism ($n = 137$), Asperger syndrome ($n = 80$), or pervasive developmental disorder—not otherwise specified ($n = 26$). This sample was made up of individuals with a single primary diagnosis made by a qualified professional (e.g., psychiatrist, psychologist) according to the *DSM–IV–TR* (American Psychiatric Association, 2000) or the *International Statistical Classification of Diseases and Related Health Problems*, *10th Revision* (World Health Organization, 2007) using appropriate

methods (e.g., record review, rating scales, observation, and interview). The sample, representative of the U.S. population, included boys and girls from each of the four geographic regions of the United States and four racial–ethnic groups (Asian, Black, White–not Hispanic, and Hispanic origin) ages 6 to 18 years. The sample size was 1,828. (See Goldstein & Naglieri, 2009, for more details about the normative sample of the Autism Spectrum Rating Scale and those identified with ASD.)

Table 1.4 provides a raw-score-to-*T*-score conversion table based on the descriptive statistics for the ASD (N = 243, M = 129.1, SD = 46.9) and national (N = 1,828, M = 53.1, SD = 36.1) reference groups. It is clear from an examination of this table that a raw score of 130 yielded very different scores for the two samples. A raw score of 130 yielded a *T* score of 50 for the ASD comparison group and a *T* score of 71 for the national comparison group. A raw score of 80 yielded a *T* score of 40 (1 standard deviation below the mean) for the ASD group and a *T* score of 57 (nearly 1 standard deviation above the mean) for the national comparison group. These results illustrate how different conclusions may be reached when the same rating scale is calibrated against two different samples.

TABLE 1.4

Comparison of *T* Scores Based on a Sample of Individuals With Autism (N = 243) and a National Comparison Group (N = 1,828)

Raw score	ASD comparison	National comparison
170	59	82
165	58	81
160	57	80
155	56	78
150	54	77
145	53	75
140	52	74
135	51	73
130	50	71
125	49	70
120	48	69
115	47	67
110	46	66
105	45	64
100	44	63
95	43	62
90	42	60
85	41	59
80	40	57
75	38	56
70	37	55
65	36	53
60	35	52
55	34	51
50	33	49
45	32	48
40	31	46
35	30	45
30	29	44
25	28	42

Note. ASD = autism spectrum disorder.

CONCLUSIONS

The field of assessment in school psychology, as in other areas of psychology, is changing. This chapter has focused on three main issues related to measurement, test development, and norming of scores. These issues are important at both theoretical and practical levels. Theoretically, the need for intelligence tests to be firmly grounded in a theory of intelligence, preferably one that is multidimensional, is increasingly apparent. These separate cognitive abilities need to be well examined and, insofar as identification of special populations is concerned, different ability profiles should be related to different academic performance patterns. In the field of skills assessment, in which tests are structured according to academic content rather than some underlying theoretical concept, it is clear that the validity of CBM measures warrants considerable research, particularly in regard to the validity of test score interpretation and normative versus true academic growth. Finally, in the area of emotional status, assessment of social–emotional strengths offers important advantages to traditional methods based on behavioral manifestations of psychopathology. The validity of this change in perspective also warrants more research. In summary, practitioners and researchers alike need to be mindful of the need to take a scientific perspective on the strengths and weaknesses of these various approaches to assessment, ask the important reliability and validity questions, and follow the research to make good decisions about which tests to use and for what purposes.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing.* Upper Saddle River, NJ: Prentice Hall.

Bracken, B. A., & Keith, L. (2004). *Clinical assessment of behavior.* Lutz, FL: Psychological Assessment Resources.

Canivez, G. L., & Gaboury, A. R. (2010, August). *Cognitive assessment system construct and diagnostic utility in assessing ADHD.* Paper presented at the 118th Annual Convention of the American Psychological Association, San Diego, California.

Carroll, J. B. (2000). Commentary on profile analysis. *School Psychology Quarterly, 15*, 449–456. doi:10.1037/h0088800

Choate, K. T. (2009). Wechsler Individual Achievement Test—Second Edition. In J. A. Naglieri & S. Goldstein (Eds.), *Assessment of intelligence and achievement: A practitioner's guide* (pp. 479–502). New York, NY: Wiley.

Davis, F. B. (1959). Interpretation of differences among averages and individual test scores. *Journal of Educational Psychology, 50*, 162–170. doi:10.1037/h0044024

Davison, M. L., & Kuang, H. (2000). Profile patterns: Research and professional interpretation. *School Psychology Quarterly, 15*, 457–464. doi:10.1037/h0088801

Doll, B. (1996). Prevalence of psychiatric disorders in children and youth: An agenda for advocacy by school psychology. *School Psychology Quarterly, 11*, 20–467. doi:10.1037/h0088919

Elliott, C. (2007). *Differential Ability Scales—Second Edition.* San Antonio, TX: Pearson.

Fiorello, C. A., Hale, J. B., & Snyder, L. E. (2006). Cognitive hypothesis testing and response to intervention for children with reading disabilities. *Psychology in the Schools, 43*, 835–853. doi:10.1002/pits.20192

Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of WISC–IV assessment.* Hoboken, NJ: Wiley.

Flanagan, D. P., Ortiz, S. O., Alfonso, V. C., & Mascolo, J. (2002). *The achievement test desk reference (ATDR): Comprehensive assessment and learning disabilities.* Boston, MA: Allyn & Bacon.

Fletcher, J. M., Denton, C., & Francis, D. J. (2005). Validity of alternative approaches for the identification of learning disabilities: Operationalizing unexpected underachievement. *Journal of Learning Disabilities, 38*, 545–552. doi:10.1177/00222194050380061101

Goldstein, S., & Brooks, R. (Eds.). (2005). *Handbook of resilience in children.* New York, NY: Kluwer/Academic Press. doi:10.1007/b107978

Goldstein, S., & Naglieri, J. A. (2009). *Autism Spectrum Rating Scale.* Toronto, Ontario, Canada: Multi-Health Systems.

Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Good, R. H., Wallin, J., Simmons, D. C., Kame'enui, E. J., & Kaminski, R. A. (2002). *System-wide percentile ranks for DIBELS Benchmark Assessment* (Tech. Rep. 9.). Eugene, OR: University of Oregon.

Hale, J. B., & Fiorello, C. A. (2004). *School neuropsychology: A practitioner's handbook.* New York, NY: Guilford Press.

Hale, J. B., Kaufman, A. S., Naglieri, J. A., & Kavale, K. A. (2006). Implementation of IDEA: Using RTI and cognitive assessment methods. *Psychology in the Schools, 43*, 753–770. doi:10.1002/pits.20186

Huang, L. V., Bardos, A. N., & D'Amato, R. C. (2010). Identifying students with learning disabilities: Composite profile analysis using the Cognitive Assessment System. *Journal of Psychoeducational Assessment, 28*, 19–30. doi:10.1177/0734282909333057

Individuals With Disabilities Education Improvement Act of 2004, P.L. 108–446, 20 U.S.C. § 1400 *et seq.*

Iseman, J., & Naglieri, J. A. (2011). A cognitive strategy instruction to improve math calculation for children with ADHD: A randomized controlled study. *Journal of Learning Disabilities*, *44*, 184–195.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Kaufman, A. S. (1979). *Intelligent testing with the WISC–R.* New York, NY: Wiley.

Kaufman, A. S., & Kaufman, N. L. (Eds.). (2001). *Learning disabilities: Psychological assessment and evaluation.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511526794

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children* (2nd ed.). San Antonio, TX: Pearson.

Kavale, K. A., Kaufman, A. S., Naglieri, J. A., & Hale, J. B. (2005). Changing procedures for identifying learning disabilities: The danger of poorly supported ideas. *School Psychologist, 59*, 16–25.

Koehler-Hak, K. M., & Bardos, A. N. (2009). Dynamic Indicators of Basic Early Literacy Skills (DIBELS): General outcomes measurement for prevention and remediation of early reading problems. In J. A. Naglieri & S. Goldstein (Eds.), *A practitioner's guide*

to assessment of intelligence and achievement (pp. 389–416). New York, NY: Wiley.

LeBuffe, P. A., & Naglieri, J. A. (1999). *Devereux Early Childhood Assessment.* Lewisville, NC: Kaplan Press.

LeBuffe, P. A., & Naglieri, J. A. (2003). *Devereux Early Childhood Assessment—Clinical Form*. Lewisville, NC: Kaplan Press.

LeBuffe, P. A., Shapiro, V. B., & Naglieri, J. A. (2009). *Devereux Student Strengths Assessment*. Lewisville, NC: Kaplan Press.

Lichtenberger, E. O., Sotelo-Dynega, M., & Kaufman, A. S. (2009). The Kaufman Assessment Battery for Children—Second Edition. In J. A. Naglieri & S. Goldstein (Eds.), *Assessment of intelligence and achievement: A practitioner's guide* (pp. 61–94). New York, NY: Wiley.

Mackrain, M., LeBuffe, P., & Powell, G. (2007). *Devereux Early Childhood Assessment for Infants and Toddlers.* Lewisville, NC: Kaplan Press.

Mather, N., & Wendling, B. (2009). Woodcock–Johnson III Tests of Achievement. In J. A. Naglieri & S. Goldstein (Eds.), *Assessment of intelligence and achievement: A practitioner's guide* (pp. 503–536). New York, NY: Wiley.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8*, 290–302. doi:10.1177/073428299000800307

Meyer, M. S. (2000). The ability–achievement discrepancy: Does it contribute to an understanding of learning disabilities? *Educational Psychology Review, 12*, 315–337. doi:10.1023/A:1009070006373

Naglieri, J. A. (1999). *Essentials of CAS assessment.* New York, NY: Wiley.

Naglieri, J. A. (2000). Can profile analysis of ability test scores work? An illustration using the PASS theory and CAS with an unselected cohort. *School Psychology Quarterly, 15*, 419–433. doi:10.1037/h0088798

Naglieri, J. A. (2011). The discrepancy/consistency approach to SLD identification using the PASS theory. In D. P. Flanagan & V. C. Alfonso (Eds.), *Essentials of specific learning disability identification* (pp. 145–172). Hoboken, NJ: Wiley.

Naglieri, J. A., & Chambers, K. (2009). Psychometric issues and current scales for assessing autism spectrum disorders. In S. Goldstein, J. A. Naglieri, & S. Ozonoff (Eds.), *Assessment of autism spectrum disorders* (pp. 55–90). New York, NY: Springer.

Naglieri, J. A., & Das, J. P. (1997). *Cognitive Assessment System.* Itasca, IL: Riverside.

Naglieri, J. A., & Goldstein, S. (2009). *Assessment of intelligence and achievement: A practitioner's guide.* New York, NY: Wiley.

Naglieri, J. A., & Goldstein, S. (2011). Assessment of cognitive and neuropsychological processes. In S. Goldstein & J. A. Naglieri (Eds.), *Understanding and managing learning disabilities and ADHD in late adolescence and adulthood* (2nd ed., pp. 137–160). New York, NY: Wiley.

Naglieri, J. A., & Gottling, S. H. (1995). A cognitive education approach to math instruction for the learning disabled: An individual study. *Psychological Reports, 76*, 1343–1354. doi:10.2466/pr0.1995.76.3c.1343

Naglieri, J. A., & Gottling, S. H. (1997). Mathematics instruction and PASS cognitive processes: An intervention study. *Journal of Learning Disabilities, 30*, 513–520. doi:10.1177/002221949703000507

Naglieri, J. A., & Johnson, D. (2000). Effectiveness of a cognitive strategy intervention to improve math calculation based on the PASS theory. *Journal of Learning Disabilities, 33*, 591–597. doi:10.1177/002221940003300607

Naglieri, J. A., LeBuffe, P. A., & Shapiro, V. (2011). *Devereux Student Strengths Assessment—Mini.* Lewisville, NC: Kaplan Press.

Naglieri, J. A., & Otero, T. (2011). Cognitive Assessment System: Redefining intelligence from a neuropsychological perspective. In A. Davis (Ed.), *Handbook of pediatric neuropsychology* (pp. 320–333). New York, NY: Springer.

Naglieri, J. A., Otero, T., DeLauder, B., & Matto, H. (2007). Bilingual Hispanic children's performance on the English and Spanish versions of the Cognitive Assessment System. *School Psychology Quarterly, 22*, 432–448. doi:10.1037/1045-3830.22.3.432

Naglieri, J. A., & Pickering, E. (2010). *Helping children learn: Intervention handouts for use in school and at home* (2nd ed.). Baltimore, MD: Brookes.

National Association of School Psychologists. (2010). National Association of School Psychologists model for comprehensive and integrated school psychological services. *School Psychology Review, 39*, 320–333.

New Freedom Commission on Mental Health. (2003). *Achieving the promise: Transforming mental health care in America: Final report* (No. SMA03-3832). Rockville, MD: U.S. Department of Health and Human Services.

O'Donnell, L. (2009). The Wechsler Intelligence Scale for Children. In J. A. Naglieri & S. Goldstein (Eds.), *A practitioner's guide to assessment of intelligence and achievement* (4th ed., pp. 153–190). New York, NY: Wiley.

Payton, J., Weissberg, R. P., Durlak, J. A., Dymnicki, A. B., Taylor, R. D., Schellinger, K. B., & Pachan, M. (2008). *The positive impact of social and emotional learning for kindergarten to eighth grade students: Findings from three scientific reviews.* Chicago, IL: Collaborative for Academic, Social, and Emotional Learning.

Pearson. (2006). *Stanford Achievement Test—10th Edition*. San Antonio, TX: Author.

Prince-Embury, S. (2005). *Resiliency Scales for Children and Adolescents*. San Antonio, TX: Pearson Education.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*, 427–469. doi:10.1016/j.jsp.2009.07.001

Reynolds, C. R., & Kamphaus, R. (2004). *Behavior Assessment System for Children, Second Edition*. San Antonio, TX: Pearson.

Roid, G. (2003). *Stanford–Binet* (5th ed.). Itasca, IL: Riverside.

Roid, G., & Bos, J. (2009). Achievement assessment and progress monitoring with the Wide Range Achievement Test. In J. A. Naglieri & S. Goldstein (Eds.), *A practitioner's guide to assessment of intelligence and achievement* (4th ed., pp. 537–572). New York, NY: Wiley.

Roid, G. H., & Tippin, S. M. (2009). Assessment of intellectual strengths and weaknesses with the Stanford-Binet Intelligence Scales—Fifth Edition (SG5). In J. A. Naglieri & S. Goldstein (Eds.), *A practitioner's guide to assessment of intelligence and achievement* (pp. 127–152). New York, NY: Wiley.

Schofield, N. J., & Ashman, A. F. (1986). The relationship between Digit Span and cognitive processing across ability groups. *Intelligence, 10*, 59–73. doi:10.1016/0160-2896(86)90027-9

Shinn, M. R., Tindal, G. A., & Stein, S. (1988). Curriculum-based measurement and the identification of mildly handicapped students: A research review. *Professional School Psychology, 3*, 69–85. doi:10.1037/h0090531

Silverstein, A. B. (1993). Type I, Type II, and other types of errors in pattern analysis. *Psychological Assessment, 5*, 72–74. doi:10.1037/1040-3590.5.1.72

Stanovich, K. E. (1994). Are discrepancy-based definitions of dyslexia empirically defensible? In K. P. van den Bos, L. S. Siegel, D. J. Bakker, & D. L. Share (Eds.), *Current directions in dyslexia research* (pp. 15–30). Lisse, the Netherlands: Swets & Zeitlinger.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2005). *Wechsler Individual Achievement Test—Second Edition*. San Antonio, TX: Pearson.

World Health Organization. (2007). *International statistical classification of diseases and related health problems, 10th revision*. Geneva, Switzerland: Author.

Wendling, B. J., Mather, N., & Shrank, F. A. (2009). Woodcock-Johnson III Tests of Cognitive Abilities. In J. A. Naglieri & S. Goldstein (Eds.), *A practitioner's guide to assessment of intelligence and achievement* (pp. 191–229). New York, NY: Wiley.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). *Woodcock–Johnson III Tests of Cognitive Abilities*. Itasca, IL: Riverside.

CHAPTER 2

PRESCHOOL ASSESSMENT

Janet E. Panter and Bruce A. Bracken

Over the past 2 decades, educators and policymakers have given considerable attention to early intervention, school readiness, and a call for public prekindergarten programs. In light of these programs and concerns, the need for preschool assessment has grown considerably in the past 20 years. Psychologists assess children birth through age 5 for a variety of purposes: to identify developmental strengths and weaknesses, to determine eligibility for services (e.g., early intervention), to evaluate children's readiness for particular programs, to provide information for curriculum planning and instruction, and to evaluate program effectiveness. Under Part C of the Individuals With Disabilities Education Improvement Act (IDEIA) of 2004, children from birth through age 35 months are eligible for early intervention services if they have a diagnosed condition, a developmental delay, or are at risk of developing a delay if no services are provided. Children from age 3 through age 5 with developmental delays or other disabling conditions can receive services through public school special education programs. Identification for services might occur in several ways, such as screenings in pediatricians' offices, schools, or child care centers to determine whether children exhibit risk factors for delays or disabilities and thus need comprehensive assessment. In addition, if physicians or parents suspect a child has a disability or delay, they may refer the child directly to the appropriate agency for assessment and intervention as needed.

Given the wide range of assessment purposes, methods are similarly varied and cover a broad spectrum of approaches and instruments. Screening batteries, for example, may consist of relatively brief measures, which are individually or group administered to children or completed by parents as third-party rating scales. However, a comprehensive psychoeducational or developmental assessment typically involves individual testing, observation, and gathering background information from family members, teachers, and child care providers. This chapter addresses assessment purposes, issues, and the methods and instruments used on the basis of the purpose of the assessment.

PURPOSES

Purposes of preschool assessment typically include screening, diagnosis and special education eligibility, instructional planning, and program monitoring and evaluation (Bordignon & Lam, 2004; Brown, Scott-Little, Amwake, & Wynn, 2007; Epstein, Schweinhart, DeBruin-Parecki, & Robin, 2004; Kelley & Surbeck, 2007). Screening programs are often the first step in identifying young children with developmental delays or disabling conditions.

Children who perform below an established standard or who are the lowest performing in their group on the screening measure are referred for ongoing monitoring and intervention or for a comprehensive individualized assessment. Comprehensive assessments often lead to diagnoses as well as determination of eligibility for early intervention or special education services. However, when the intent of postscreening assessment is program

DOI: 10.1037/14049-002
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

planning and monitoring, results are used to identify appropriate remedial curricula, to group students for instruction, or to monitor instructional efficacy. Similarly, postscreening assessments may be conducted to evaluate program effectiveness (i.e., to hold educators and programs accountable for their performance).

Although experts have typically agreed that early childhood assessment is valuable, disagreement regarding specific methodology and instruments is considerable (e.g., see Bagnato & Neisworth, 1994; Bracken, 1994). Given the poor psychometric properties of many instruments (Bracken, 1987; Flanagan & Alfonso, 1995) and the challenges of assessing young children (Nagle, 2007), some professionals have argued against the use of traditional standardized instruments and methodology. Many individuals and groups have expressed particular concern with the use of high-stakes testing, that is, using results derived from preschool assessments for informing promotion or retention decisions or other outcomes that directly affect the young child assessed (Meisels, 1987, 1989, 1992; Shepard, 1997). Other experts have purported that assessment limitations and challenges are insufficient reasons to abandon traditional methods; rather, practitioners should engage in the thoughtful use of developmentally appropriate methods (traditional and alternative) to meet the varied purposes of the assessments they conduct (Bracken, 1994).

Appropriate use of assessment instruments and methods has been a concern of psychologists and other professionals for many years. The 1999 *Standards for Educational and Psychological Testing*, published by the joint task force of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, delineated "criteria for the evaluation of tests, testing practices, and the effects of test use" (p. 2) and provided a frame of reference for identifying and applying relevant criteria to test development, appropriate uses, and interpretation. The *Standards* address issues such as test reliability and validity, test construction, ensuring freedom from racial and cultural bias, professional training requirements, multimethod and multisource procedures, and attention to the outcomes of using tests. Each of these standards applies to tests or procedures used with young children and should guide practitioners' selection of instrumentation; however, the *Standards* did not specifically address the unique assessment challenges in working with very young, prekindergarten children.

The National Association for the Education of Young Children and the National Association of Early Childhood Specialists in State Departments of Education (2003) issued a joint position statement with a more direct focus on the preschool population. They recommended attention to technical adequacy (i.e., reliability and validity), linking results to educational decision making, including families in the assessment process, and identifying children's needs to provide intervention. Consistent with the *Standards* and with this position statement, the National Association of School Psychologists (2009) took the following position: It "supports comprehensive and authentic early childhood assessment processes and outcomes that are fair and useful to (a) detect the need for intervention, (b) enhance intervention delivery and individual child response to intervention, and (c) enhance program and system effectiveness" (p. 1).

In sum, early childhood assessments should be demonstrably valid for their intended purposes and sufficiently reliable to give practitioners confidence in the assessment outcomes. Moreover, early childhood assessments should take place within authentic contexts, be multidisciplinary in nature, and gather information from multiple sources. Kelley and Surbeck (2007) predicted just these features when they wrote, "The future trends for preschool assessment will undoubtedly focus on a multimethod, multi-disciplinary assessment process that includes significant family input" (p. 20). Before reaching this aspirational goal, professionals must address four critical issues, as discussed in the next section.

CRITICAL ISSUES IN PRESCHOOL ASSESSMENT

In the debate over appropriate assessment practices with preschool-age children, four issues consistently arise: developmental and behavioral influences,

instrumentation, working with children from diverse cultural and linguistic backgrounds, and addressing construct-irrelevant variance.

Developmental and Behavioral Influences

Assessing preschoolers requires considerable care and attention to multiple factors to ensure accurate interpretation of results. By virtue of their young age and rapid development, especially of cognitive and communication skills, preschoolers present psychologists with a number of challenges. Nagle (2007, p. 33) described this age group as a "unique population that is qualitatively different from their school-age counterparts." Preschool children experience rapid developmental changes that vary greatly both across and within domains (Bordignon & Lam, 2004; Epstein et al., 2004); consequently, their performance is often relatively unstable and subject to considerable change in a relatively brief period of time.

In addition to these rapid developmental spurts, preschool children's erratic behavior can adversely affect their performance on standardized measures and will sometimes result in an underestimation of their overall abilities or their functioning in specific domains. For instance, preschool children typically have limited attention spans and high activity levels; thus, assessment activities requiring sustained attention, limited movement, and considerable focus pose substantial challenges for preschoolers, independent of the cognitive demands of the tasks (Bracken, 2007b; Gredler, 1992; Nagle, 2007). Psychologists conducting assessments, therefore, must take care that a child's behavior is not the primary source of information for drawing inferences about cognitive abilities. In addition, the demands of the assessment setting might cause a child to exhibit behaviors atypical for him or her, such as hyperactivity or inactivity resulting from heightened anxiety or fear related to interacting with an unfamiliar examiner or novel assessment activities. To complicate the situation further, young children with exceptionalities bring additional challenges to the assessment process because of their often uneven development, limited language skills, lack of experience with standardized tasks, and adaptive and behavioral difficulties (Bracken, 1994, 2007b).

Instrumentation: Psychometric Properties

A second major consideration in preschool assessment is the technical adequacy of available instruments. In a 1987 review of 10 instruments commonly used to assess cognitive abilities, Bracken found significant limitations for most tests, especially for children younger than age 4. To evaluate the instruments, he established minimum criteria for the following psychometric properties: reliability, floors, item gradients, and validity information. These criteria were based on a review of the literature, the types of decisions to be made using the test results, and knowledge of and experience with preschool instrumentation. Bracken recommended standards for internal consistency coefficients of .90 or higher for total test scores used to make diagnostic or placement decisions and of at least .80 for subtests and scales. Reliability coefficients in the .80 to .89 range were deemed adequate for screening but not sufficiently strong for diagnosis or placement. Bracken also addressed assessment stability, examining each test manual to determine the extent to which the trait measured was stable over brief periods (or expected to remain stable). One problem Bracken encountered when evaluating test–retest reliability was the lack of representativeness of the samples used in these studies. In many instances, samples of older children were used to infer test–retest reliability, an unacceptable practice given the many differences between preschool children and their older peers.

Test floors (lowest achievable standard score) should permit scores of at least 2 standard deviations below the mean to "allow for the meaningful differentiation of low-functioning children from children who function in the average to low-average range of abilities" (Bracken, 1987, p. 317). *Item gradients* refer to the amount of change in standard scores that results from each 1-point raw score change. In other words, when a child answers one item correctly (earns 1 raw score point) on a test, how much does the standard score increase? Bracken (1987) recommended a gradient no steeper than one third of a standard deviation change in standard score based on a 1-point raw score change. Flanagan and Alfonso (1995) concurred with Bracken's emphasis on item gradients in early childhood

assessment, saying, "Although information on subtest item gradient is accorded little attention in the literature, it is an important test characteristic to evaluate because it allows one to determine the extent to which a test effectively differentiates among various ability levels" (p. 70).

Regarding validity, Bracken (1987) did not set minimum criteria; rather, he examined each measure for the presence or absence of validity evidence and recommended that test users evaluate the instrument's validity on the basis of their intended purposes and the decisions for which the test scores would be used. In general, his evaluation illustrated that most preschool instruments fail to meet these minimal psychometric standards, especially for children younger than age 4. He concluded,

> Through the examination of the information in this paper, it can be readily seen that preschool assessment below the age of 4 years seems to present the greatest psychometric problems. Selection of tests for use with low-functioning children below age 4 needs to be made with special care. As can be seen, many of these tests designed for preschool use are severely limited in floor, item gradient, and reliability, especially at the lower age levels. (p. 325)

In 1995, Flanagan and Alfonso conducted a similar review to determine whether commonly used cognitive measures had improved or continued to exhibit similar technical limitations. In their summary, they concluded that the standardization process and samples were exemplary for most instruments and that each test's internal consistency reliability was good. Problems were noted, though, with test–retest reliability procedures and samples. Flanagan and Alfonso concluded that the tests' most problematic areas were test floors and subtest item gradients. In addition, they concluded it was difficult to evaluate test validity, especially given the lack of an agreed-on definition of intelligence.

Flanagan and Alfonso (1995) concluded "that not much has changed with regard to the technical adequacy of intelligence tests for children at the lower end of the preschool age range" (p. 86). They identified two instruments—the Woodcock–Johnson Psycho-Educational Battery: Tests of Cognitive Ability—Revised (Woodcock & Johnson, 1989, 1990; Woodcock & Mather, 1989, 1990) and the Bayley Scales of Infant Development—Second Edition (Bayley, 1993)—that met the majority of the criteria for children younger than age 4. Although they noted improvements in some areas since Bracken's (1987) earlier review, serious limitations continue, and psychologists must exercise considerable care when choosing measures. Moreover, when interpreting test results, professionals should be especially attentive to tests' technical characteristics to avoid possible misinterpretations or overinterpretations of resulting test scores.

Regarding assessment in the social–emotional domain, Bracken, Keith, and Walker (1998) evaluated 13 instruments designed to assess preschool children's development and behavior, such as the Behavior Assessment System for Children (Reynolds & Kamphaus, 2004). As Bracken et al. pointed out, it is important to examine the technical properties of frequently used instruments to ensure their suitability for preschool populations. The rating scales studied varied in terms of reliability, with total test coefficients for internal consistency ranging from .73 to .98. More important, some tests did not provide global scores. In terms of the tests' ceilings and floors, Bracken et al. found that all scales except one provided scores 2 standard deviations below or above the mean (as appropriate to the directionality of the scale). They also examined each test's item gradient for sensitivity to small differences in behavior among examinees and reported that the tests generally had sufficient items to discern atypical from typical behavior. Another concern regarding third-party scales was interrater reliability, which generally fell below the desired level of .90. Bracken et al. reported that all of the instruments provided validity evidence, although the quality of that evidence was not assessed. On the basis of their evaluation of the 13 instruments, they concluded, "Regardless of type of assessment procedure used in the evaluation of preschool children, the instruments employed could and should be stronger psychometrically than they are currently" (p. 162). Bracken et al. also reported that the newer

instruments appear more technically sound than the older measures, indicating that test authors are on the right track regarding these psychometric properties. However, psychologists planning to use third-party rating scales should exercise care to ensure the technical adequacy of the instruments they use.

Given the distinctive behaviors of preschool-age children previously discussed (e.g., uneven development, high activity level) and the psychometric limitations of many instruments available for preschoolers, some experts have called for an end to traditional assessment methods. For instance, Bagnato and Neisworth (1994) recommended an end to the use of traditional intelligence tests, arguing for alternative, more subjectively scored assessment methods, such as observations, parent interviews, and play-based assessment. In his response to their article, Bracken (1994) pointed out that traditional and alternative methods are not mutually exclusive; rather, they are tools psychologists use to gather data as part of the assessment process. He also highlighted the important fact that all methods and tools used with those in early childhood are obligated to the same psychometric criteria; professionals must hold alternative methods to the same high standards set for traditional instruments.

Children From Culturally and Linguistically Different Backgrounds

Given the large number of children from culturally and linguistically different backgrounds, psychologists must be prepared to work with children with a variety of background experiences and cultural expectations, especially in school settings. Practitioners working with culturally diverse populations must first consider the meaning of culture and its influence on children's thinking and behavior. Unfortunately, professionals have not reached consensus regarding a definition of culture or the extent of its influence on test results (Frisby, 1998). Broadly defined, *culture* consists of "the customs, civilization, and achievements of a particular time or people" ("Culture," 1999). Some theorists have argued that culture drives behavior in a deterministic fashion, whereas others have seen its influence as more subtle and varied (Frisby, 1998). Psychologists, then, are faced with the complex task of understanding an individual's cultural heritage and placing each assessment within the context of that person's heritage. As Padilla (2001) suggested, "Assessment is made culturally sensitive through an incessant, basic, and active preoccupation with the culture of the group or individual being assessed" (p. 7). Thus, psychologists working with culturally different children must engage the child within the context of the child's culture to interpret assessment results appropriately and accurately. Volume 2, Chapter 12, this handbook describes a multicultural perspective of clinical assessment.

Terms such as *culture, race,* and *ethnicity* are frequently used interchangeably; however, important and relevant differences exist. Although individuals of a particular race or ethnicity may have a common culture, one should not presume they do simply because they have similar genetic backgrounds (i.e., race) or because they share a nationality or language (i.e., ethnicity; Frisby, 1998; Sattler, 2008). In fact, Santos de Barona and Barona (2007) strongly warned against assuming a child belongs to a particular racial or ethnic group on the basis of appearance alone. Instead, psychologists need to determine a family's self-described racial or ethnic identity and its level of *acculturation,* or the extent to which the individuals in question identify with or participate in a particular culture (see Ponterotto, Gretchen, & Chauhan, 2001, for a fuller discussion).

One of the main concerns when dealing with children from culturally different families or from diverse racial and ethnic groups is their primary language. In 2007, nearly 20% of children between ages 5 and 17 spoke a language other than English in the home (American Community Survey, 2008). As Nagle (2007) pointed out, one of the issues for preschoolers from linguistically and culturally diverse backgrounds is their limited shared experience with peers from other groups. One way such experiences are gained is through community-based preschool programs. Unfortunately, many non-English-speaking children do not have access to such programs. The National Household Survey Program (O'Donnell & Mulligan, 2008) reported that 36% of children whose parents both speak a language other than English are enrolled in center-based preschools or child care programs compared with 54% of

children with one English-speaking parent and 62% of students for whom English is the parents' primary language.

Culturally different children diverge from their mainstream peers on several dimensions, including domains such as social interaction style and verbal communication (Santos de Barona & Barona, 2007). One of the most distinct variations is in the child's style of interacting with adults. Behaviors such as making and sustaining eye contact, asking questions, initiating conversation, or incessant verbal engagement are viewed by some cultural groups as undesirable, especially for young children. These behavioral differences are the reason some professionals have called for an end to the use of standardized measures of cognitive ability and academic achievement with children with cultural differences. Santos de Barona and Barona (2007) have reminded users, however, that standardized instruments provide useful information, although results must be interpreted in light of a child's culture rather than assuming cultural equivalence. As with all preschool children, psychologists must carefully interpret test results in light of children's behavior and patterns rather than assuming that scores are unquestionably valid or representative of the child's true abilities.

Construct-Irrelevant Variance

A primary concern in the assessment of young children is the elimination of construct-irrelevant variance (Bracken, 2007b). To do so, psychologists must decide which constructs are targeted for assessment (i.e., which skills or abilities are to be measured) and then identify potential construct-irrelevant variables that threaten the validity of the assessment process. Bracken (2007b) identified several variables that could be either construct relevant or construct irrelevant depending on the referral questions and assessment goals; this list includes factors such as language proficiency, level of enculturation, and level of education and life experiences. He proposed,

> When a variable is identified as irrelevant to the assessed construct and yet negatively influences the child's test performance, that variable should be considered as a source of test bias and should be eliminated or moderated to as great an extent as possible. (p. 138)

Moreover, Bracken (2007b) identified several possible contributors to construct-irrelevant variance from four sources: examinee, examiner, environment, and instruments used. A complete discussion of each source is beyond the scope of this chapter. At a minimum, psychologists involved in preschool assessment must be attentive to the following potential threats to validity:

- child characteristics, such as health, motivation, fatigue, and anxiety;
- examiner characteristics, including approachability and affect, physical appearance, rapport, and psychometric skill;
- environmental factors, including furniture, distraction, and climate control; and
- psychometric characteristics of scores emerging from instruments, such as test floors, ceilings, item gradient, reliability, validity, norms tables, and age of norms.

ASSESSMENT METHODS

Under IDEIA, children from birth to age 3 receive early intervention services through Part C of the act. Consistent with the standards and position statements mentioned earlier, this legislation requires evaluations to be systematic, be multidisciplinary (i.e., collaborative), include the family as both participants and examinees, and be comprehensive. Although some disagreement exists among professionals regarding the relative importance of each characteristic, the consensus is that they all merit inclusion (Gredler, 1992, 1997, 2000; Nagle, 2007; Shepard, Taylor, & Kagan, 1996).

Reminding psychologists to be systematic in their approach seems initially unnecessary, because the work they do requires considerable attention to detail, organization, and careful planning. However, assessments can become so routine that examiners fail to plan carefully by attending to the characteristics of each child and ensuring appropriate instruments are selected and proper methods are followed (e.g., not administering a standard battery to every child).

In addition to approaching all assessments systematically, preschool assessments should be multidisciplinary. Nagle (2007) discussed terminology and definitions regarding the way in which professionals collaborate with each other, pointing out that *multidisciplinary* usually connotes a high level of professional independence with little contact or effort to integrate information. In other words, multidisciplinary assessment is limited as a means of collaboration and cross-fertilization; when professionals spend little time collaborating with each other, they are not likely to work together in the way intended by IDEIA. Psychologists and other team members should carefully consider the collaborative process and strive to understand and be informed by one another's perspectives (Nagle, 2007).

Given the problems with collaboration in the multidisciplinary approach, some professionals have argued for using an interdisciplinary style. Individuals on interdisciplinary teams also work independently but have an added emphasis on shared communication and consultation. As professionals engage in authentic collaboration, they work as a team to meet the needs of families and children in the way intended (Nagle, 2007). A third option for team members is the transdisciplinary model. Transdisciplinary assessment is a more interactive process than the traditional multidisciplinary approach and is designed to optimize collaboration by crossing discipline boundaries during the assessment, not just before or after. Such interaction is evident in Linder's (2008) Transdisciplinary Play-Based Assessment—2, which provides early childhood specialists across a variety of fields with an assessment structure and method whereby each professional collects data relevant to his or her specialty (see Athanasiou, 2007, for a full discussion of play-based assessment). Although a transdisciplinary approach has many benefits, one limitation is that the process is time consuming and requires considerable coordination among professionals (Nagle, 2007).

The collaborative process just described extends beyond professionals to the family. IDEIA gave considerable attention to the role of families in the assessment of young children (birth through age 3) and recognized the parents' unique role in knowing and understanding a child's functioning. Moreover, IDEIA recommends that examiners must pay careful attention to the functioning of a family. Parents' circumstances and functioning have direct and indirect effects on children; therefore, determining a family's functional and structural strengths and weaknesses is an essential component of comprehensive assessment.

As mentioned earlier, preschool assessments should be comprehensive in nature—including multiple sources, contexts and settings, and content areas or domains. Conducting a comprehensive evaluation ensures that psychologists identify a child's strengths and weaknesses and provides educators with all of the information they will need for educational planning. The five developmental domains that must be assessed to ensure a comprehensive evaluation are cognitive functioning, expressive and receptive language skills, adaptive and self-help behavior, gross and fine motor skills, and social–emotional functioning. More detailed measurement of each domain then follows. Specific methods and instrumentation capturing the range of assessment approaches—including large-scale readiness screening, individual, and comprehensive assessments—are discussed in light of possible assessment goals.

Readiness Screening

Screening of incoming students on school entry is a relatively commonplace phenomenon across the country. In a 1996 telephone survey of state-level early childhood specialists, Shepard et al. found that most states mandate some form of readiness screening, although they frequently allow local education agencies to choose their own screening methods and decide how the results will be used. Some states exclude the use of tests for placement or tracking young students' progress, and Shepard et al. noted positive trends toward appropriate use of readiness measures (e.g., not using screening results to recommend delayed entry into kindergarten).

LaParo and Pianta (2000) conducted a meta-analytic review of longitudinal studies using academic–cognitive and social–behavioral assessments administered in preschool or kindergarten to predict later performance on similar measures. They found small to moderate effect sizes for academic–cognitive measures and small effect sizes for scales

in the social–behavioral domain. They wrote, "Children's rank order changes over the preschool to second grade period, especially with regard to the social/behavioral domain. Instability or change may be the rule rather than the exception during this period" (p. 476). As discussed earlier, it is important to recognize that children's rapid developmental changes clearly affect the validity and usefulness of measurement outcomes, and care must be taken to ensure that assessment results are interpreted appropriately, within the limitations imposed by the instrument and the characteristics of preschool children.

When making a decision to implement a screening program, one of the most important questions is, "What will we do with the results?" If a child falls below a predetermined cut score or fails to exhibit competence in a specific skill set, what will be the appropriate next step? Current practices range from recommending delayed entry to school to adjusting classroom curriculum and instruction on the basis of children's current skills and needs. Most often, children classified as at risk on the basis of screening results receive a follow-up comprehensive assessment to determine whether they have developmental delays, meet criteria for special education services, or need other accommodations. Two relevant questions arise regarding the effectiveness of the screening batteries being used. First, what should be the target skills assessed? That is, do some domains predict future academic performance better than do other areas? Is information needed from several domains, and if so, which domains should be represented and which should be prioritized? Second, what degree of accuracy should be expected of screening measures or batteries? (See Panter, 2010, for a full discussion of these issues.)

Screening batteries have historically assessed content in one or more of the five domains identified by the National Education Goals Panel: physical well-being and motor development, social and emotional development, approaches to learning, language development, and cognition and general knowledge (Kagan, Moore, & Bredekamp, 1995). To determine the most accurate predictors of later achievement, Duncan et al. (2007) conducted a meta-analysis of six longitudinal datasets from the United States, Canada, and the United Kingdom. Children were screened between the ages of 4.5 and 6 in the following areas: reading achievement, language and verbal ability, math achievement, attention skills, attention problems, externalizing behavior problems, internalizing problems, and prosocial behaviors. Outcome measures in reading and mathematics were administered anywhere from Grade 3 to ages 13 to 14. No behavioral outcomes or other school-related variables were measured (e.g., attendance, graduation). Duncan et al. found that three of the screening variables predicted later performance: reading and language achievement, math achievement, and attention. Interestingly, basic math achievement equally predicted later performance in math and reading.

Similarly, a longitudinal study by the Santa Clara County Partnership for School Readiness and Applied Survey Research (2008) assessed children's performance from kindergarten entry through fifth grade. At the beginning of the kindergarten year, teachers completed the Kindergarten Observation Form, a rating scale that addresses self-care and motor skills, self-regulation, social expression, and kindergarten academics. On the basis of their performance on the Kindergarten Observation Form, children were placed into one of four categories: all stars (high in all four areas); focused on the facts (high kindergarten academics, low social expression); social stars (high social expression, low kindergarten academics), and needs prep (low in all four areas). Outcome measures were children's ratings as basic, proficient, or advanced on the California standardized achievement tests in English and language arts and mathematics. Students rated as proficient or advanced are deemed by the state to have met adequate yearly progress benchmarks. In third and fourth grades, the all stars were significantly more likely than children in the other three groups to earn proficient or advanced scores. Similarly, focused-on-the-facts students performed better than social-star or needs-prep students. In fifth grade, the all-star students' performance decreased as a group, and they were the equivalent of the focused-on-the-facts students; both groups outperformed social-star and needs-prep students. It is evident from the longitudinal studies reviewed here that early academic

knowledge—evident at the time of readiness screening—is essential to outcomes measured by standardized testing in higher grades.

A study by Panter (1998) produced similar results, showing that the tryout version of the Bracken Basic Concept Scale—Revised (Bracken 1998) was the best predictor of children's scores on the Metropolitan Readiness Test—6 (Nurss & McGauvran, 1995), teachers' ratings of children's readiness for first grade, and the Academic Skills subtest of the teacher version of the Social Skills Rating System (Gresham & Elliott, 1990). In Panter's study, 76 kindergartners were administered the tryout version of the Bracken Basic Concept Scale—Revised, the Geometric Design subtest of the Wechsler Preschool and Primary Scale of Intelligence—Revised (Wechsler, 1989), and the parent version of the Social Skills Rating System. Of the measures given, the Bracken Basic Concept Scale—Revised's School Readiness Composite best predicted retention and referral for services, correctly classifying 90% of the total sample. Moreover, the School Readiness Composite accounted for almost half of the variance in the Metropolitan Readiness Test—6 Pre-Reading scores ($r^2 = .45$). In a later study, Panter and Bracken (2009) found the Bracken School Readiness Assessment (Bracken, 2002) to be the best predictor of kindergarten students' retention in grade or referral for services. Overall, it correctly classified 91% of the 117 students in the study. The Bracken School Readiness Assessment also accounted for more than 60% of the variance in Metropolitan Readiness Test—6 Pre-Reading scores (corrected $r^2 = .66$). The Bracken School Readiness Assessment did not perform equally well for African American and Caucasian students. Although it accounted for a high percentage of the variability in the Caucasian students' Metropolitan Readiness Test—6 scores, it was not predictive of African American students' performance once the Brigance K and 1 Screen (Brigance, 1992) was taken into account.

To determine the effectiveness of screening measures for identifying children who are at risk for developmental or learning difficulties, Gredler (1992, 1997, 2000) outlined procedures for evaluating an instrument's predictive validity. Predictive validity is evaluated in terms of a measure's sensitivity and specificity. *Sensitivity* is the screening instrument's correct identification of low performers (i.e., students identified as at risk who perform poorly on the outcome measure, or true positives). *Specificity* refers to the number of children who meet the screening criterion and perform in the acceptable range on the outcome measure (i.e., true negatives). True, or accurate, positives and negatives are correct hits—the screening instrument is working properly to identify children who are at risk for school problems in the domain of interest. Of course, instances of false positives (i.e., students identified as at risk who perform in the acceptable range on the outcome) and false negatives (i.e., students identified as not at risk who do poorly on the outcome measure) also occur.

Gredler (1992) reviewed the performance of 12 screening instruments and found an average sensitivity index of .77, meaning that 77% of the students identified as at risk exhibited poor performance on the outcome criterion. Conversely, 23% of the students identified as at risk demonstrated acceptable skills on the outcome measure. The average specificity index of .81 indicated that 81% of students classified as not at risk performed within or above an acceptable standard on the outcome measure, whereas 29% performed below acceptable standards (i.e., failed to meet standards). Gredler (2000) conducted a similar analysis on the performance of six screening measures. Sensitivity indices ranged from .20 to .91 ($Mdn = .77$) with higher specificity indices (range = .66–.98). Overall, 57% of the children classified as at risk in these studies performed poorly, whereas 43% later performed at an acceptable level (i.e., classified as at risk but successful on the outcome measure). Of the children classified as not at risk, 90.8% did well on the outcome measure (i.e., were successful). So, just more than 9% of the children classified as not at risk were children who later failed and who had been misidentified as not at risk.

Debate exists regarding acceptable levels of sensitivity and specificity (Boan, Aydlett, & Multunas, 2007; Gredler, 1992, 2000; Panter, 2010). Some professionals have argued for highly sensitive instruments to ensure that students at risk for poor outcomes are identified early, allowing professionals to provide them with appropriate intervention services.

As Gredler (1992, 2000) pointed out, though, there are costs involved in misidentification of children as at risk (false positives). First, the parents of a child identified as at risk may develop concerns about their child's performance or abilities on the basis of these inaccurate screening results, concerns that may not be fully allayed by a comprehensive assessment indicating typical development and functioning. Similarly, teachers may prejudge children on the basis of negative screening results, regardless of later outcomes. Second, the child referred for further assessment or intervention may also recognize that he or she has failed to achieve at expected levels, resulting in negative self-judgments and expectations. Third, practical considerations related to the high cost of comprehensive assessments arise, especially if they are unnecessary.

Assessment of Family Functioning

There is wide agreement regarding the importance of the family to a child's healthy development and to ensuring academic and social success. Parental engagement, opportunities for cognitive stimulation, language and communication variables, and demographic factors, such as socioeconomic status, all play a role in children's early development. It is essential, then, that psychologists begin the evaluation process with the parents. Assessment of the home environment, including family dynamics, is an essential component of comprehensive preschool assessment and requires attention to "building rapport, acting in a culturally responsive manner, and attending to safety" (Nickerson, Duvall, & Gagnon, 2007, p. 156). See Table 2.1 for a partial list of measures of family functioning.

Building rapport with parents requires sensitivity and awareness of parental needs and concerns. Showing respect for parents and their concerns and issues establishes an atmosphere of trust and shared decision making. Too often, professionals approach parents with little respect for the parents' skills, culture, and opinions about their child, making it highly unlikely that a collaborative atmosphere will exist. Psychologists must exhibit even more

TABLE 2.1

Instruments and Methods Used in the Assessment of Family Functioning

Author and publication date	Instrument	Age range	Domains assessed	Methods
Abidin (1995)	Parenting Stress Index—3rd edition	1 month–12 years	Child characteristics, parent characteristics, and situations salient to parenting Designed for early identification of dysfunctional parenting as well as child or adult emotional and behavioral problems	Rating scale
Caldwell & Bradley (2001)	Home Observation for Measurement of the Environment Inventory	Birth–3 years 3–6 years 6–10 years	Child's learning environment, family relationships, and stressors	Direct observation and semistructured interviews
Fox (1994)	Parent Behavior Checklist	1–5 years	Beliefs about child-rearing (expectations, nurturing, discipline)	Rating scale
Moos & Moos (1994)	Family Environment Scale—3	All ages (parent form) 5–12 years (children's form)	Family cohesion, conflict, organization, and expressiveness	True–false questionnaire

sensitivity when working with children from culturally different backgrounds. As previously discussed, it is important that professionals become knowledgeable about children's culture and be prepared to address cultural differences (and biases) when assessing the family and the child (Frisby, 1998).

Home visits can be quite informative and often provide psychologists with considerable information about children and their families. Visiting the child's home allows professionals to get to know family members in their natural setting; provides information about a family's culture, socioeconomic status, and living arrangements; and may also be interpreted as a sign of respect for the family and their culture (Nickerson et al., 2007). As mentioned previously, safety is sometimes a concern for professionals making home visits, and psychologists are cautioned to deal with this important issue before making home visits.

To assess family functioning, Nickerson et al. (2007) recommended attention to the following four domains: demographics, family relationships, strengths, and stressors. Demographics include factors such as socioeconomic status, ethnicity, parental education, and family composition. Children whose families live in poverty or who are members of racial or ethnic minority groups are more likely to experience academic difficulties than their peers from circumstances that are more affluent or who have nonminority group membership (Garbarino & Ganzel, 2006; O'Donnell & Mulligan, 2008).

To assess family relationships, psychologists should consider the child's relationship with the primary caregiver, parenting style, sibling interactions, and involvement with extended family (Nickerson et al., 2007). In a chapter on parenting, Osofsky and Thompson (2006) discussed adaptive factors that produce healthy parenting styles: reciprocity (mutually satisfying relationships), emotional availability (accessible and responsive), mother's role (often discussed in the literature and so not addressed here at length), father's role (interactive and responsible for meeting child's needs), and social networks (supportive, stress-reducing relationships). They also discuss risk factors for maladaptive parenting, such as substance abuse, violence, teenage mothers, and parental psychopathology.

Assessment of Child Functioning

Comprehensive assessment requires attention to children's functioning in the following domains: cognitive, communication and language, physical, self-help and adaptive, and behavior and social–emotional. Some domains, such as cognitive or behavior and social–emotional, are typically assessed directly by psychologists, and other areas may be evaluated by professionals in other disciplines as part of the multidisciplinary team approach (e.g., speech–language pathologists). Please note that although domains are considered separately in this discussion, all ability domains are intertwined. Moreover, assessment in one area will be influenced by a child's functioning in other areas, such as a cognitive measure that requires a coordinated physical motor response. In some instances, the multiple domains measured by an instrument result in the construct-irrelevant influences previously discussed, and psychologists will need to choose measures and interpret scores in light of those issues.

Cognitive functioning is typically measured with traditional instruments such as the Wechsler Preschool and Primary Scale of Intelligence—IV (Wechsler, 2012) or the Stanford–Binet Intelligence Scale—5 (Roid, 2003). The Bayley Scales of Infant and Toddler Development—III (Bayley, 2006) are often used in the assessment of infants and toddlers and measure intelligence as well performance in other ability areas. Other cognitive measures are available as well (see Table 2.2); once again, psychologists are cautioned to examine a test carefully to ensure it meets the psychometric standards previously discussed, especially for children ages 2 and younger.

Communication and language abilities are closely linked to cognitive functioning, and many tests of intelligence require considerable knowledge of basic concepts (Bracken, 1986; Bracken & Panter, 2011; Flanagan, Alfonso, Kaminer, & Rader, 1995; Kaufman, 1978). In addition, many intelligence measures have a strong verbal loading, so children with limited language abilities may appear less cognitively able than their more verbally facile peers. In addition to the information provided by cognitive measures, psychologists can measure receptive and expressive language skills with tests such as the Bracken Basic Concept Scale—3 (Bracken, 2006).

TABLE 2.2

Measures of Cognitive Functioning

Author and publication date	Instrument	Age range	Scores provided	Comments
Bayley (2006)	Bayley Scales of Infant and Toddler Development—III	1–42 months	Cognitive, Language, Motor, Social–Emotional, Adaptive	Individually administered; 1–3 months (10 minutes); 4–8 months (15 minutes); 9+ months (20 minutes)
Bracken & McCallum (1998)	Universal Nonverbal Intelligence Test	5 years and older	Full Scale IQ Index Scores:, Reasoning Quotient, Memory Quotient, Symbolic Quotient, Non-Symbolic Quotient	Abbreviated battery (10–15 minutes); standard battery (30 minutes); extended battery (45 minutes)
Eliot (2007)	Differential Ability Scales—II	2 years, 6 months, and older	Composites: General Conceptual Ability, Special Nonverbal Composite Clusters: Verbal, Nonverbal Reasoning, Spatial	Individually administered; core battery (45–60 minutes); diagnostic subtests (30 minutes)
Roid (2003)	Stanford-Binet Intelligence Scales—5	2 years and older	IQs: Verbal, Nonverbal, and Full Scale Factors: Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual–Spatial Processing, Working Memory	Individually administered; 30–50 minutes for preschool children
Wechsler (2012)	Wechsler Preschool and Primary Scale of Intelligence—IV	2 years, 6 months–7 years, 7 months	Full Scale IQ Index Scores: Verbal Comprehension, Visual Spatial, Working Memory (all ages); Fluid Reasoning & Processing Speed (ages 4 through 7 years, 7 months)	Individually administered; 2 years, 6 months–3 years, 11 months—45 minutes; 4 years–7 years, 3 months—60 minutes
Woodcock, McGrew, & Mather (2001)	Woodcock–Johnson III Tests of Cognitive Abilities	2 years and older	General Intellectual Ability, Verbal Ability, Thinking Ability, Phonemic Awareness, Cognitive Efficiency	Individually administered; standard and supplemental batteries

Administration of the Bracken Basic Concept Scale—3, for instance, provides information about preschoolers' understanding of basic concepts, which are important for performance on intelligence tests (Bracken, 1986; Flanagan et al., 1995) and for success in school (Bracken, 2006; Bracken & Crawford, 2010).

Psychologists can also gather data through third-party rating scales designed to measure adaptive and self-help skills or social–emotional behavior (see Table 2.3 for instruments in these domains). This third-party information is valuable because it allows examiners to know how the preschooler behaves in his or her usual setting—at home, child care, or school—which might be quite different from behavior exhibited in the formal testing situation. As discussed earlier, gathering data from parents is especially important when working with preschoolers. In formal assessment settings, preschool children sometimes become overly shy or anxious and may

TABLE 2.3

Measures of Adaptive and Self-Help Skills and Social–Emotional Behavior

Author	Measure	Ages	Domains assessed	Method
Achenbach & Rescorla, 2000	Achenbach System of Empirically Based Assessment	1.5–5	Clinical scales, language development survey	Parent and teacher rating scales
Bracken & Keith (2004)	Clinical Assessment of Behavior	2–18 (parent) 5–18 (teacher)	Clinical and adaptive scales; includes validity indices	Parent and teacher rating scales
Harrison & Oakland (2003)	Adaptive Behavior Assessment System—II	Birth–89	Adaptive behavior in conceptual, social, and practical domains	Parent and teacher rating scales
Reynolds & Kamphaus, 2004	Behavior Assessment System for Children—2	2–21 (parent and teacher) 8–25 (self-report)	Clinical and adaptive scales, includes validity indices	Parent, teacher, and self-report rating scales
Sparrow, Cicchetti, & Balla (2005)	Vineland Adaptive Behavior Scales—II	Birth and older	Adaptive behavior in communication, motor skills, daily living skills, and socialization domains	Interview or survey form Includes parent and teacher forms

not exhibit their full repertoire of skills and knowledge. Parents can provide essential information regarding the full range of a child's developmental accomplishments, especially if the rating scale addresses several domains rather than being narrowly focused. For example, the Vineland Adaptive Behavior Scale—II (Sparrow, Cicchetti, & Balla, 2005) measures adaptive functioning in the areas of communication, motor skills, daily living skills, and socialization.

In addition to the use of standardized instruments, clinical observations are an essential part of the assessment process and provide important information about the preschool child; as Bracken (2007a) expressed,

> Observations should be employed to describe and explain children's test and nontest behaviors, attest to the validity or invalidity of test scores, at least partially explain children's variable test performance, lend support for diagnoses and remediation strategies made on the basis of standardized test results, and provide the examiner with information needed to develop hypotheses concerning a child's learning style and individual strengths and weaknesses. (pp. 95–96)

When assessing preschoolers, psychologists need to be attentive to an examinee's behavior within the following parameters:

- time—beginning, middle, and end of the assessment session;
- contexts and tasks—response to settings, task demands, and materials; and
- relationships—the way child relates to and separates from parents, level of comfort, and interaction with examiner.

By making note of the child's specific behaviors in these ways, the examiner has useful data for normative and ipsative comparisons (see Table 2.4 for a list of relevant behaviors; Bracken, 2007a). For instance, does the child exhibit typical behaviors during separation from parents? Did the child respond more positively to some tasks or materials than to others? Did the child's affect change when tasks became more challenging?

Using these clinical observations, the examiner can generate hypotheses about the child's functioning, globally and in each of the domains previously discussed. Hypothesis generation is an ongoing process of developing a theory to explain the child's behavior, confirming or disconfirming that explanation (with test results, parent and teacher reports, and other observations), and then modifying the

TABLE 2.4

Preschool Behaviors to Observe

Domain	What to observe
Appearance	Notable physical characteristics?
Height and weight	How does the child compare with peers?
Physical abnormalities	Are there unusual characteristics or indicators of problems with diet, abuse, lack of medical attention, improper sleep, and so forth?
Grooming and dress	Is the child receiving appropriate care and supervision? Note: Children who have been playing outside or in an active preschool may be disheveled or dirty. Seasonally inappropriate clothing (e.g., long sleeves in hot weather to cover bruises) might indicate abuse or neglect.
Gross and fine motor skills	How well does child walk, run, climb stairs, skip, hop, balance on one foot? Is motor development symmetrical? How well does child manipulate small objects, use a pencil, or color?
Speech	What is the child's vocabulary? Basic concept attainment? Are there speech disorders? Is the child verbally fluent?
Activity level	Is the child lethargic? Overly active?
Attention	Is the child appropriately attentive, especially on tasks requiring sustained attention? Did inattention influence child's performance?
Distractibility	Does the child respond to distractions in the environment in ways that disrupt his or her attention?
Impulsivity	Does the child act before hearing directions? Blurt out answers?
Affect	How does the child respond to various situations? How does the child deal with failure? With novel situations?
Anxiety	What causes the child to become anxious? How does the child display anxiety?
Comprehension and problem solving	How does the child approach problem solving? Are the child's efforts systematic or random?
Reactions to other people and situations	How does child interact with parents (together and individually)? Siblings? Teacher? Peers? Strangers?

From *Psychoeducational Assessment of Preschool Children* (4th ed., pp. 45–56) by B. A. Bracken & R. J. Nagle (Eds.), 2007, Mahwah, NJ: Erlbaum. Copyright 2007 by Lawrence Erlbaum Associates. Adapted with permission.

hypothesis or generating a new one. Informed clinical observations allow psychologists to develop a picture of a child that goes beyond test scores and that provides parents and teachers with a deeper and fuller understanding of that child's abilities, functioning, and behavior within and across various contexts.

SUMMARY

Preschool assessment presents a unique set of challenges because of the age and developmental characteristics of the children in this age group. Preschool children are especially challenging clients because of their rapid and often irregular rate of development, exuberant behavior, and lack of experience with formal schooling and testing. Additionally, psychologists must deal with issues related to the technical adequacy of preschool instruments, working with children from culturally diverse backgrounds, and eliminating, as best possible, construct-irrelevant variance from the assessment process. In spite of these challenges and potential pitfalls, preschool assessment is valuable and essential. It allows psychologists to conduct readiness screening, to engage in diagnosis and placement, to inform instruction and curriculum, and to evaluate programs, holding educators accountable.

This chapter addressed many of the issues related to improving the validity, reliability, and purposeful decision making of assessments with the youngest clients. Clearly, psychologists must be attentive to several factors, including the psychometric properties of the instruments they select, the appropriate use of such instruments and their respective scores within the context of comprehensive assessment,

skills and abilities to be assessed, and collaboration with educators and families in meaningful and practical ways. Although challenging, preschool assessment and screening can be beneficial for children, their families, and educators by providing meaningful information about the development and delay of children's skills and abilities before they become intractably problematic.

References

Abidin, R. R. (1995). *Parenting Stress Index* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for ASEBA preschool forms and profiles*. Burlington: Research Center for Children, Youth, and Families, University of Vermont.

American Community Survey. (2008). *Characteristics of people by language spoken at home: 2006–2008 American Community Survey 3-year estimates (Table S1603)*. Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_08_3YR_S1603&prodType=table

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Athanasiou, M. S. (2007). Play-based approaches to preschool assessment. In B. A. Bracken & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 219–238). Mahwah, NJ: Erlbaum.

Bagnato, S. J., & Neisworth, J. T. (1994). A national study of the social and treatment "invalidity" of intelligence testing for early intervention. *School Psychology Quarterly, 9*, 81–102. doi:10.1037/h0088852

Bayley, N. (1993). *Bayley Scales of Infant Development* (2nd ed.). San Antonio, TX: Psychological Corporation.

Bayley, N. (2006). *Bayley Scales of Infant Development* (3rd ed.). San Antonio, TX: Pearson Education.

Boan, C. H., Aydlett, L., & Multunas, N. (2007). Early childhood screening and readiness assessment. In B. A. Bracken & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 49–67). Mahwah, NJ: Erlbaum.

Bordignon, C. M., & Lam, T. C. M. (2004). The early assessment conundrum: Lessons from the past, implications for the future. *Psychology in the Schools, 41*, 737–749. doi:10.1002/pits.20019

Bracken, B. A. (1986). Incidence of basic concepts in the directions of five commonly used American tests of intelligence. *School Psychology International, 7*, 1–10.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 5*, 313–326. doi:10.1177/073428298700500402

Bracken, B. A. (1994). Advocation for effective preschool assessment practices: A comment on Bagnato and Neisworth. *School Psychology Quarterly, 9*, 103–108. doi:10.1037/h0088845

Bracken, B. A. (1998). *Bracken Basic Concept Scale—Revised*. San Antonio, TX: Psychological Corporation.

Bracken, B. A. (2002). *Bracken School Readiness Assessment*. San Antonio, TX: Psychological Corporation.

Bracken, B. A. (2006). *Bracken Basic Concept Scale—Third Edition: Expressive and Receptive*. San Antonio, TX: Harcourt Assessment.

Bracken, B. A. (2007a). Clinical observation of preschool assessment behavior. In B. A. Bracken & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 95–110). Mahwah, NJ: Erlbaum.

Bracken, B. A. (2007b). Creating the optimal preschool testing situation. In B. A. Bracken & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 137–153). Mahwah, NJ: Erlbaum.

Bracken, B. A., & Crawford, E. (2010). Basic concepts in early childhood educational standards: A 50-state review. *Early Childhood Education Journal, 37*, 421–430. doi:10.1007/s10643-009-0363-7

Bracken, B. A., & Keith, L. K. (2004). *Clinical Assessment of Behavior*. Lutz, FL: Psychological Assessment Resources.

Bracken, B. A., Keith, L. K., & Walker, K. C. (1998). Assessment of preschool behavior and social-emotional functioning: A review of thirteen third party instruments. *Journal of Psychoeducational Assessment, 16*, 153–169. doi:10.1177/073428299801600204

Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test*. Itasca, IL: Riverside.

Bracken, B. A., & Panter, J. E. (2011). Using the Bracken Basic Concept Scale and Bracken Concept Development Program in the assessment and remediation of young children's concept development. *Psychology in the Schools, 48*, 464–475.

Brigance, A. (1992). *Brigance K and 1 Screen* (3rd ed.). North Billerica, MA: Curriculum Associates.

Brown, G., Scott-Little, C., Amwake, L., & Wynn, L. (2007). *A review of methods and instruments used in state and local school readiness evaluations* (Issues & Answers Report REL 2007–No. 004). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs

Caldwell, B. M., & Bradley, R. H. (2001). *Home Inventory and administration manual* (3rd ed.). Little Rock: University of Arkansas for Medical Sciences and University of Arkansas at Little Rock.

Culture. (1999). In *The Oxford reference online*. Retrieved from http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t21.e7594

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Jape, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446. doi:10.1037/0012-1649.43.6.1428

Eliot, C. D. (2007). *Differential Ability Scales* (2nd ed.). San Antonio, TX: Pearson Education.

Epstein, A. S., Schweinhart, L. J., DeBruin-Parecki, A., & Robin, K. B. (2004, July). Preschool assessment: A guide to developing a balanced approach. *Preschool Policy Matters, 2*, 1–11. Retrieved from http://nieer.org/resources/policybriefs/7.pdf

Flanagan, D. P., & Alfonso, V. C. (1995). A critical review of the technical characteristics of new and recently revised intelligence tests for preschool children. *Journal of Psychoeducational Assessment, 13*, 66–90. doi:10.1177/073428299501300105

Flanagan, D. P., Alfonso, V. C., Kaminer, T., & Rader, D. E. (1995). Incidence of basic concepts in the directions of new and recently revised intelligence tests for preschoolers. *School Psychology International, 16*, 345–364. doi:10.1177/0143034395164003

Fox, R. A. (1994). *Parent Behavior Checklist.* Austin, TX: Pro-Ed.

Frisby, C. L. (1998). Culture and cultural differences. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Genier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 51–73). Washington, DC: American Psychological Association. doi:10.1037/10279-003

Garbarino, J., & Ganzel, B. (2006). The human ecology of early risk. In J. Shonkoff & S. Meisels (Eds.), *Handbook of early intervention* (2nd ed., pp. 76–93). New York, NY: Cambridge University Press.

Gredler, G. R. (1992). *School readiness: Assessment and educational issues.* Brandon, VT: Clinical Psychology Publishing.

Gredler, G. R. (1997). Issues in early childhood screening and assessment. *Psychology in the Schools, 34*, 99–106. doi:10.1002/(SICI)1520-6807(199704)34:2<99::AID-PITS3>3.0.CO;2-N

Gredler, G. R. (2000). Early childhood screening for developmental and educational problems. In B. A. Bracken (Ed.), *The psychoeducational assessment of preschool children* (3rd ed., pp. 399–411). Boston, MA: Allyn & Bacon.

Gresham, F. M., & Elliot, S. N. (1990). *Social Skills Rating System.* Circle Pines, MN: American Guidance Service.

Harrison, P., & Oakland, T. (2003). *Adaptive Behavior Assessment System* (2nd ed.). Minneapolis, MN: Pearson Assessment.

Individuals With Disabilities Education Improvement Act of 2004, Pub. L. 108–446, 20 U.S.C. § 1400 *et seq.*

Kagan, S. L., Moore, E., & Bredekamp, S. (1995). *Reconsidering children's early development and learning: Toward common views and vocabulary*. Washington, DC: National Education Goals Panel, Goal One Technical Planning Group.

Kaufman, A. S. (1978). The importance of basic concepts in the individual assessment of preschool children. *Journal of School Psychology, 16*, 207–211. doi:10.1016/0022-4405(78)90002-X

Kelley, M. F., & Surbeck, E. (2007). History of preschool assessment. In B. A. Bracken & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 3–28). Mahwah, NJ: Erlbaum.

La Paro, K. M., & Pianta, R. C. (2000). Predicting children's competence in the early school years: A meta-analytic review. *Review of Educational Research, 70*, 443–484. doi:10.3102/00346543070004443

Linder, T. (2008). *Transdisciplinary play-based assessment* (2nd ed.). Baltimore, MD: Paul H. Brookes.

Meisels, S. J. (1987). Uses and abuses of developmental screening and school readiness testing. *Young Children, 42*, 4–6.

Meisels, S. J. (1989). High-stakes testing in kindergarten. *Educational Leadership, 46*, 16–22.

Meisels, S. J. (1992). Doing harm by doing good: Iatrogenic effects of early childhood enrollment and promotion policies. *Early Childhood Research Quarterly, 7*, 155–174. doi:10.1016/0885-2006(92)90002-G

Moos, R. H., & Moos, B. S. (1994). *Family Environment Scale* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Nagle, R. J. (2007). Issues in preschool assessment. In B. A. Bracken & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 29–48). Mahwah, NJ: Erlbaum.

National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education. (2003). *Early childhood curriculum, assessment, and program evaluation: Building an effective, accountable system in programs for children birth through age 8.* Washington, DC: Author. Retrieved from http://www.naeyc.org/files/naeyc/file/positions/pscape.pdf

National Association of School Psychologists. (2009). *Early childhood assessment* (Position statement). Bethesda, MD: Author.

Nickerson, A. B., Duvall, C. C., & Gagnon, S. G. (2007). Assessment of home and family dynamics. In B. A. Bracken & R. J. Nagle (Eds.), *The psychoeducational assessment of preschool children* (4th ed., pp. 155–172). Mahwah, NJ: Erlbaum.

Nurss, J. R., & McGauvran, M. E. (1995). *Metropolitan Readiness Tests* (6th ed.). San Antonio, TX: Psychological Corporation.

O'Donnell, K., & Mulligan, G. (2008). *Parents' reports of the school readiness of young children from the national Household Education Surveys Program of 2007* (NCES 2008–051). Washington, DC: National Center for Education Statistics.

Osofsky, J. D., & Thompson, M. D. (2006). Adaptive and maladaptive parenting: Perspectives on risk and protective factors. In J. Shonkoff & S. Meisels (Eds.), *Handbook of early intervention* (2nd ed., pp. 54–75). New York, NY: Cambridge University Press.

Padilla, A. M. (2001). Issues in culturally appropriate assessment. In L. A. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (2nd ed., pp. 5–27). San Francisco, CA: Jossey-Bass.

Panter, J. E. (1998). *Assessing the school readiness of kindergarten children.* Unpublished doctoral dissertation, University of Memphis, Memphis, TN.

Panter, J. E. (2010). Kindergarten readiness. In S. B. Thompson (Ed.), *Kindergarten: Programs, functions, and outcomes* (pp. 51–92). Hauppauge, NY: Nova Science.

Panter, J., & Bracken, B. A. (2009). Validity of the Bracken School Readiness Assessment for predicting first grade readiness. *Psychology in the Schools, 46*, 397–409. doi:10.1002/pits.20385

Ponterotto, J. G., Gretchen, D., & Chauhan, R. V. (2001). Cultural identity and multicultural assessment: Quantitative and qualitative tools for the clinician. In L. A. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (2nd ed., pp. 67–99). San Francisco, CA: Jossey-Bass.

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children.* Circle Pines, MN: American Guidance Service.

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment System for Children* (2nd ed.). Circle Pines, MN: American Guidance Service.

Roid, G. (2003). *Stanford–Binet Intelligence Scales* (5th ed.). Itasca, IL: Riverside.

Santa Clara County Partnership for School Readiness and Applied Survey Research. (2008). *Does readiness matter? How kindergarten readiness translates into academic success.* Retrieved from http://www.appliedsurveyresearch.org/storage/database/early-childhood-development/school-readiness/sanmateosantaclara/DoesReadinessMatter_ALongitudinalAnalysisFINAL3.pdf

Santos de Barona, M., & Barona, A. (2007). Assessing multicultural preschool children. In B. A. Bracken & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 69–92). Mahwah, NJ: Erlbaum.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). San Diego, CA: Jerome M. Sattler.

Shepard, L. A. (1997). Children not ready to learn? The invalidity of school readiness testing. *Psychology in the Schools, 34*, 85–97. doi:10.1002/(SICI)1520-6807(199704)34:2<85::AID-PITS2>3.0.CO;2-R

Shepard, L. A., Taylor, G. A., & Kagan, S. L. (1996). *Trends in early childhood assessment policies and practices.* Washington, DC: National Education Goals Panel.

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales* (2nd ed.). San Antonio, TX: Pearson Education.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence—Revised.* Chicago, IL: Psychological Corporation.

Wechsler, D. (2012). *Wechsler Preschool and Primary Scale of Intelligence* (4th ed.). San Antonio, TX: Pearson Assessment.

Woodcock, R. W., & Johnson, M. E. (1989). *Woodcock–Johnson Psycho-Educational Battery—Revised.* Allen, TX: DLM Teaching Resources.

Woodcock, R. W., & Mather, N. (1989). *WJ–R Tests of Cognitive Ability—Standard and Supplemental Batteries: Examiner's manual.* Allen, TX: DLM Teaching Resources.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III Tests of Cognitive Abilities.* Itasca, IL: Riverside.

CHAPTER 3

ASSESSMENT OF INTELLECTUAL FUNCTIONING IN CHILDREN

John O. Willis, Ron Dumont, and Alan S. Kaufman

This chapter emphasizes the rationale, techniques, and special considerations for assessing the intellectual functioning of children in educational, clinical, and other settings. The term *children* refers to anyone younger than age 18, with an emphasis on school-age children. (Please see Chapter 2, this volume, for detailed information on testing younger children. Bracken & Nagle, 2007, have provided an extensive discussion of preschool assessment.) Additionally, given the differing opinions about what exactly the terms *cognitive ability*, *intellectual ability,* and *intellectual functioning* mean, we use the terms interchangeably throughout the chapter.

Children's intellectual functioning is assessed for multiple purposes (Kaufman, 2009; Sattler, 2008): to determine eligibility for special education services, to determine eligibility for gifted and talented programs, for intervention and placement decisions, to assist in the development of Individualized Education Programs, to identify at-risk preschoolers and design interventions for them (Bagnato, 2007; Lichtenberger, 2005), to evaluate intervention programs, and to measure and monitor progress (Bradley-Johnson, 2001; Brassard & Boehm, 2007; Epstein, 2004). To accomplish these purposes, assessment of intellectual functioning should be integrated with assessment of academic achievement (please see Chapters 5, 6, and 10, this volume, for more in-depth information).

Although one of the historical foundations of modern intellectual assessment is testing children for educational placement (Binet & Simon, 1916/1980; Terman, 1916), some of the instruments and practices in current use are based on tests for adults. As Kaufman (2009) noted, "The similarity of Wechsler's original set of subtests to the tasks used to evaluate recruits, soldiers, and officers during World War I is striking" (p. 31). Yet those same tasks formed the basis for Wechsler's children's scales, as Wechsler (1949) stated in the Wechsler Intelligence Scale for Children (WISC) manual:

> *The Wechsler Intelligence Scale for Children* has grown logically out of the *Wechsler–Bellevue Intelligence Scales* used with adolescents and adults. . . . In this brief manual the background of the new Scale can only be sketched. It is assumed that the reader is acquainted with the author's *The Measurement of Adult Intelligence*, third edition. (p. 1)

The younger the child, the less applicable adult instruments and procedures are likely to be, so poor matches often occur between test demands and young examinees' developmental skills. Simply dumbing down adult tests with easier items for children works about as well as shortening the legs and sleeves of adult clothing for children. Examiners must be careful that they are using developmentally appropriate tasks with children. Nonetheless, despite the adult roots of Wechsler's subtests, his current children's tests are quite child oriented. In our opinion, each successive edition of the WISC and Wechsler Preschool and Primary Scale of Intelligence (WPPSI) has become more child friendly.

DOI: 10.1037/14049-003
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

Children present special assessment challenges that are usually less common or less extreme among adults with no or mild disabilities. Some of these challenges are discussed later. Examiners who have not taught or worked extensively with children and whose children or grandchildren are well below or beyond school age may need to find opportunities to spend time with children in nontest settings to help build their understanding of childhood and adolescent functioning and current child and adolescent culture.

COGNITIVE ABILITIES

Jensen (1998, 2002), Gottfredson (2008), and many other authorities have asserted that general intelligence (*g*; Spearman, 1904) is the single most important predictor of important life outcomes such as academic success, job attainment, income level, and the likelihood of incarceration; assessment of *g* or overall mental ability remains an important aspect of intellectual assessment of children. As with earlier tests, some contemporary intellectual ability tests, such as the Reynolds Intellectual Assessment Scales (RIAS; Reynolds & Kamphaus, 2003), are explicitly designed to assess "general intelligence and its two primary components, fluid and crystallized intelligence" (Reynolds & Kamphaus, 2003, p. iv).

However, many contemporary researchers have placed less emphasis on *g*. Consequently, many newer intellectual instruments for children—for example, the WISC, fourth edition (WISC–IV; Wechsler, 2003); Differential Ability Scales, second edition (DAS–II; Elliott, 2007a); Kaufman Assessment Battery for Children, second edition (KABC–II; Kaufman & Kaufman, 2004a); and Stanford–Binet Intelligence Scales, fifth edition (SB5; Roid, 2003)—have been developed or restructured to yield not only a total score, a proxy for *g,* but also to reflect the change in the conceptualization of IQ from the unitary *g* to multiple indexes tapping various levels of mental processes (e.g., Hale & Fiorello, 2001, 2004). Such contemporary researchers, test authors, and clinicians as Flanagan and Kaufman (2009); Flanagan, Ortiz, and Alfonso (2007); Elliott (2007b); and Sattler (2008) have slightly different approaches to how they recommend interpretation of the subtests and the indexes, clusters, or composites of the intellectual ability scales. However, they are strong advocates of intellectual assessment and are in agreement that such an endeavor is a worthwhile investment of time because it sheds light on the individual's strengths and weaknesses, which can in turn be translated into useful and meaningful educational recommendations (Hale & Fiorello, 2004; Hale et al., 2010).

In the school setting, assessment of children is often a means to an end, a way to provide those who present with learning or other difficulties with intervention programs or special education services that successfully meet their special needs. Whereas *assessment* is the process of collecting, synthesizing, and interpreting information, *testing* is a formal and systematic practice for the collection of a sample of behavior and then using that information to make generalizations about performance or similar behaviors (Airasian, 2002). Faced with the task of testing children, as opposed to adults, examiners should be keenly aware of several special considerations and problems. These issues include selection of tests, special challenges, population diversity issues, concerns with young children, legal issues, theoretical considerations, disability and disadvantages, test sessions, accommodations and adaptations, examiner's limitations, and recommendations in evaluation reports.

SELECTION OF TESTS

Potential limitations to norm-referenced testing of children have been amply documented in the literature. For example, accurately determining children's performance levels using measurements at one point in time is difficult because they are growing and changing at a rapid pace (e.g., Glascoe, 2005; Valdivia, 1999). Generally, children referred for evaluation are likely to have characteristics such as short attention span, high distractibility and low frustration tolerance, a level of discomfort with unfamiliar adults, and inconsistent performance in unfamiliar settings, which may result in challenging assessments (e.g., Bracken & Walker, 1997; Cole & Cole, 1996; Valdivia, 1999). Moreover, young children tend to abstain from adult-directed activities, they may not fully engage in testing procedures that do

not incorporate play activities, and they may not respond to verbal items (Valdivia, 1999). Adolescents may also not be inclined to follow adult instructions and may remain disengaged from testing procedures. Conversely, young children's and adolescents' unique characteristics may also facilitate establishing rapport because children and adolescents are egocentric and sometimes sociable and they like the undivided attention provided during the testing activity (Bracken & Walker, 1997; Sattler, 2008). Tests' colorful and novel equipment may also attract and engage young children (Bracken & Walker, 1997), and the difference between routine classroom demands and intellectual assessment may appeal to older children and adolescents (Sattler, 2008).

Possible limitations are also associated with the standardized and inflexible administration procedures of norm-referenced testing (Bagnato & Neisworth, 1994; Sattler, 2008). Notably, these limitations may be even more pronounced when working with children with diverse impairments and may lower a test's validity (Brassard & Boehm, 2007; Lichtenberger, 2005). Other indicated limitations are related to tests' published psychometrics. Indeed, although the American Educational Research Association (AERA), American Psychological Association, and National Council on Measurement in Education (1999) recommended standards for norm-referenced tests, tests are not federally or professionally regulated (Glascoe, Martin, & Humphrey, 1990), and test manuals sometimes do not provide sufficient evidence of the psychometric integrity of their tests (Buros, 1938; Geisinger, Spies, Carlson, & Plake, 2007; Sattler, 2008; Snyder & Lawson, 1993).

SPECIAL CHALLENGES

Assessment of children's intellectual functioning presents, as noted earlier, several special challenges, some of which are listed here.

1. Children usually have shorter attention spans and greater distractibility than adults, a characteristic that would argue for shorter tests that could be completed within the child's span of good attention.
2. However, a child may be less reliable than an adult in responding to test items, which would argue for longer tests that might mitigate the effects of fluctuating attention.
3. Children are often inconsistent in their responses, which would also suggest using more test items to increase reliability, but again, longer tests would strain short attention spans.
4. Our clinical experience over the past several decades would encourage us to evaluate children in many short sessions, but that is not the way in which the tests were standardized, and the logistics would be extremely difficult.
5. Children, of course, are developing and learning at breath-taking speed compared with adults. A child might be able to pass items on Tuesday that had been impossible on Monday. Norms tables should probably be divided by single months, not spans of 3, 4, or 6 months.
6. Sampling error is a more serious problem with children than with adults. If, for example, one is trying to assess general information, it is fairly safe to select some questions, such as "Who was Thomas Jefferson?" that item tryouts have shown to be representative of most adults' broad spans of general knowledge. Evidence has shown that adults who can answer that question correctly tend to know a lot of information and that those who cannot do not. However, if a school has just made a big, week-long deal of celebrating Presidents' Day, children in that school will temporarily be able to pass that item, even if they are generally ignorant of U.S. history and other general-information topics. Conversely, if references to Jefferson have been downplayed in the children's new textbooks, generally well-informed children may fail that item.
7. Item gradients tend to be steeper for children than for adults. For example, many vocabulary subtests include both picture-naming and formally defining words. The steps between not being able to explain what words mean and being able to do so are really not clearly defined.
8. Children are not always particularly interested in doing their best on a test. Of course, adults may be depressed, may malinger, or may underperform for other reasons, but the assumption

that an examinee wants to answer questions correctly and solve puzzles swiftly may be especially questionable with children and adolescents.

9. Skills that are generally considered to be developmental in nature may be influenced by educational experiences. This issue also affects adults, but the wide variations among children's school and home environments may result in differences in, for example, oral language or visual–motor abilities that might not reflect genuine developmental or intellectual differences between children.
10. Small differences in test format and wording may be deal breakers for children. For example, if a test asks the child to select the biggest number rather than the greatest (at most a trivial difference for most adults), a child may be stymied, depending on the terminology used in the child's math class. The difference between a No. 2 pencil and a thick "primary" pencil might make a notable difference in a child's performance on a test of copying geometric designs or drawing them from memory.
11. Cognitive abilities that are usually of little interest when assessing adults can be very important when one assesses children. For one example, phonological awareness is widely considered an essential cognitive ability for early development of reading and writing skills (e.g., Torgesen, 2002) and is included in such cognitive ability measures as the Woodcock–Johnson III (WJ III) Tests of Cognitive Ability and WJ III Tests of Achievement (Woodcock, McGrew, & Mather, 2001a, 2001b) and the DAS–II (Elliott, 2007a). For adults, the ability to rhyme words or repeat a word with one sound omitted is usually not very important.
12. There is often insufficient "floor" for young children on cognitive tests (Dumont & Willis, n.d.; Goldman, 1989; Sattler, 2008). It is always prudent to see what scores would be achieved by a child who has no clue how to respond to certain types of items, that is, what subtest, cluster, and IQ scores would result from raw scores of zero or 1. For example, on the WJ III Normative Update, a child of age 2 who earned only 1 raw score point on each subtest would receive an Extended General Intellectual Ability standard score of 89 and a General Intellectual Ability—Early Development standard score of 73. By age 3, those scores would drop to 66 and 39, respectively, and by age 4, to 50 and 20, respectively. Raw scores of 0 on all subtests for a child age 6 would yield a WISC–IV Full Scale IQ (FSIQ) of 40 (the lowest available score at all ages), and raw scores of 1 would give a FSIQ of 44. With the WPPSI–III, at age 2.5, a raw score of 0 would give a FSIQ of 44, and a raw score of 1 would give a FSIQ of 58. On the DAS–II, at age 2.5, a raw score of 0 on all subtests would give a General Conceptual Ability (GCA) of 48 and a raw score of 1 would give a GCA of 59. On the RIAS, at age 3, raw scores of zero would yield a Composite Intelligence Index of 65, and a raw score of 1 would yield a Composite Intelligence Index of 76. On the KABC–II, raw scores of 0 at age 3 would yield Fluid–Crystallized Index and Mental Processing Index scores of 40, and a raw score of 1 would give a Fluid–Crystallized Index score of 41 and a Mental Processing Index score of 45.

POPULATION DIVERSITY ISSUES WITH CHILDREN

Using norm-referenced testing with children of diverse cultural backgrounds is especially challenging (Bracken & McCallum, 2001; Espinosa, 2005; Gutkin & Reynolds, 2009; Ortiz & Dynda, 2005; Ortiz, Ochoa, & Dynda, 2012; Sandoval, Frisby, Geisinger, Scheuneman, & Grenier, 1998; Sattler & Hoge, 2006, Chapter 4). Both Danesco (1997) and Harry (1992) contended, in fact, that disability is a social and cultural construct because children are compared with others of the same age. Such comparisons depend on the prevailing culture's definitions and expectations of disabilities and delay, which may be very different from those of the child's culture.

The assertion that the normative samples in norm-referenced tests truly represent children of different cultural backgrounds is questionable (e.g., Garcia Coll, 1990). Moreover, examining developmental domains separately to conclude that delay exists in specific areas conflicts with a more holistic approach that is used in various cultural groups

other than Western (Kagitcibasi, 1996). Another obstacle is the observation that culture influences the acquisition of developmental milestones (Garcia Coll, 1990). Cross-cultural studies have long shown, in fact, that children do not develop at the same pace across cultures (e.g., Gesell, 1925). Using tools that are insensitive to cultural diversity may thus result in a misdiagnosis of disability—either an unwarranted diagnosis of disability that will lead to erroneously identifying children as eligible for special education or a failure to diagnose that may deprive the child of much-needed services (McLean, 1998).

Limitations of assessment associated with cultural diversity are underscored by the increasing diversity of young children in the United States (Espinosa, 2005; Sattler & Hoge, 2006, Chapter 4). From 2000 to 2005, the Hispanic population in the United States increased by 6.9 million (U.S. Census Bureau, 2006), becoming the largest ethnic minority group in the country (Lichter, Quian, & Crowley, 2006). In comparison, the African American and Asian populations have increased by 1.9 million and the White population by 2.5 million (U.S. Census Bureau). In fact, according to the 24th and 26th Annual Reports to Congress on the Implementation of the Individuals With Disabilities Education Act of 1990 (IDEA; U.S. Department of Education, 2001, 2005), each ethnic and racial population category of children with disabilities between ages 3 and 5 served under IDEA has increased: Hispanics and African Americans by 10%, Whites by 5%, and Asians by less than 5%. This trend is expected to continue (Lichter et al., 2006), which presents a challenge because cultural and language barriers may interfere with the identification, assessment, and, in turn, the provision of special education services (Espinosa, 2005). Interestingly, although testing individuals with dissimilar cultural backgrounds has received increasing attention since the 1950s, concerns had already been raised in 1910 when diverse cultural groups immigrated to the United States (Anastasi & Urbina, 1997).

Indeed, annual reports to Congress on the implementation of IDEA have emphasized the disproportionate representation of ethnic and racial groups in special education; whereas African American children are often overrepresented, Asian children are underrepresented (Pavri, 2001). This disproportionality may be attributed to culturally biased assessment tools, inadequate use of translators during assessment, or professionals who are insensitive to the effect that cultural, bilingual, or ethnic background may have on children's performance when tested (McLean, 1998). A prominent additional factor that may confound minority children's performance is their socioeconomic status, a factor that is difficult to separate from racial and ethnic background (Brahan & Bauchner, 2005). Critics have, in fact, contended that standardized and norm-referenced testing is biased against children from socioeconomic statuses and cultures different from those of European American middle-class children (Cronshaw, Hamilton, Onyura, & Winston, 2006; Cummins, 1986). Brahan and Bauchner (2005) showed that children of ethnically and racially diverse backgrounds are usually from low socioeconomic statuses and, in addition, tend to have parents with limited education and job security. It is worth noting here, finally, that the health status of some minority infants places them at higher risk for developmental problems (Garcia Coll, 1990).

Examiners must select tests that will not penalize children from diverse backgrounds. It is appropriate and, in fact, necessary to carefully and thoroughly document a child's current functioning levels in oral and written English, in English-language academic achievement, and in adaptive behavior by norms for mainstream U.S. culture (please see Chapters 9 and 10, this volume, for a comprehensive discussion of these issues). However, it is absolutely inappropriate to confuse such measures of acquired skills (or of skills that have not been acquired) with measures of intellectual capacity.

In many cases, children for whom English is a second language will never have developed a high level of cognitive and academic language proficiency (Cummins, 1979) in their first language and will not yet have developed cognitive and academic language proficiency in English. They might have achieved sufficient basic interpersonal communicative skills to carry on a fluent conversation with the examiner, but not to demonstrate their verbal intellectual potential on a test of intellectual functioning. In that situation, the examiner cannot even seek and use

a verbal intelligence test standardized in the child's first language, assuming the test and the skills needed to administer and score it are even available.

Examiners are tempted to use translations of verbal tests (or even to make up their own translations). Geisinger (1994a) contended that translations are probably necessary but that complex issues of validity must be thoroughly addressed. Geisinger warned that translators must be sensitive to cultural as well as linguistic issues and that care must be taken to ensure that the translated test is still measuring the same construct. Ad hoc efforts by bilingual examiners clearly do not meet this standard. Even a conscientious and fluently bilingual examiner could not avoid the risk of translating, for instance, the vocabulary test word *cat* with an equivalent as difficult as *feline* or as easy as *kitty*.

Several tests now provide alternative instructions in languages other than English, most often Spanish. The DAS–II (Elliott, 2007a) even provides on a CD a demonstration of standardized test administration in American Sign Language. The manuals warn examiners not to use these translations unless they are truly fluent in the particular language. This rule is very important. Not only would badly pronounced instructions invalidate the test, but the examiner might receive responses he or she is unable to evaluate. Some tests also accept correct responses in languages other than English (e.g., WJ III), which seems to be a good idea on the whole but does introduce a source of interexaminer variability. A child who responds correctly in a language other than English, for example, naming a picture as *caballo, cheval,* or *Pferd,* will be marked as correct by an examiner who speaks Spanish, French, or German, but not by one who does not.

The examiner can, of course, use a nonverbal or nonvocal test of intellectual functioning such as the Leiter International Performance Scale—Revised (Leiter–R; Roid & Miller, 1997), Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998), KABC–II Nonverbal Scale (Kaufman & Kaufman, 2004a), DAS–II Special Nonverbal Composite (Elliott, 2007a), Comprehensive Test of Nonverbal Intelligence, second edition (CTONI–2; Hammill, Pearson, & Wiederholt, 2009), or Wechsler Nonverbal Scale of Ability (Wechsler & Naglieri, 2006). Bracken and Naglieri (2003) discussed this application of nonverbal tests and warned that "most 'nonverbal tests' in fact are best described as language-reduced instruments with verbal directions—sometimes with lengthy and complex verbal directions" (p. 246). Serious problems could result with such a nonverbal assessment if it was not supplemented with additional measures. First, the examiner has no way of knowing whether the child's verbal intelligence would be a strength, a weakness, or a disability, which is a huge omission in a comprehensive assessment of intellectual functioning that must be acknowledged in the report. Second, by assessing only abilities that can be tested nonvocally, the examiner risks mistaking a strength or weakness in visual abilities for the child's overall cognitive level. Again, this limitation must be spelled out clearly in the report. A definitive statement about the child's overall intellectual potential would have to be deferred until the child's English-language cognitive and academic language proficiency abilities were sufficient for a valid assessment.

Examiners must also be extremely sensitive to cultural differences that may lead to misinterpretation of a child's behavior and responses or may cause the child to misinterpret the examiner. Familiarity with the child's culture is ideal. Lacking such familiarity, the examiner must rely on common sense, sensitivity, and information sought from adult members of the child's culture. Again, please see Chapter 11, this volume; Gutkin and Reynolds (2009); and Sattler and Hoge (2006, Chapter 3), among other resources.

YOUNGER SCHOOL-AGE CHILDREN

Assessment of children younger than school age is covered in detail in Chapter 2, this volume (see also Bracken & Nagle, 2007). Assessment of young children is an especially challenging undertaking that requires great familiarity with children of that age range. Some older texts have provided especially helpful discussions of testing young children, for example, Haeussermann (1958), Kaufman and Kaufman (1977), and McCarthy (1972).

Despite considerable controversy, some agreement seems to exist in the literature that preschool

norm-referenced tests and norm-referenced tests for very young school-age children, especially those that yield an IQ and particularly with children younger than age 4, are generally less than adequate in terms of their psychometrics—although improvement in recent years has been considerable and ongoing (Bracken, 1994; Flanagan & Alfonso, 1995; Ford & Dahinten, 2005). Some consensus also exists that a variety of methodologies should be used in the assessment of young children (Bracken, 1994; Ford & Dahinten, 2005). Nevertheless, norm-referenced tests can be used cautiously with young children as long as professionals are aware of their limitations and consider those limitations when interpreting the results (Bracken, 1994; Flanagan & Alfonso, 1995).

Ultimately, however, norm-referenced testing has multiple practical purposes: to screen children suspected of being at risk, to make decisions in regard to eligibility for special education and related services, to plan and evaluate intervention programs, to assess and monitor progress (Bradley-Johnson, 2001; Brassard & Boehm, 2007), and to analyze developmental trajectories (Batshaw, Pellegrino, & Roizen, 2007). The advantages of using norm-referenced tests are well documented in the literature. Norm-referenced tests provide normative data that easily allow for the comparison between young children and a normative sample; standard scores thus differentiate between young children of the same age who can and cannot perform certain skills (Bracken, 1994; Flanagan & Alfonso, 1995; Sattler, 2008). In essence, norm-referenced tests generate valid and reliable quantitative data that can be used to allow young children to gain access to special education by determining eligibility (McLean, Bailey, & Wolery, 2004; Sattler, 2008), to afford diagnosis, and to predict children's future performance (Bagnato & Neisworth, 1994; Flanagan & Alfonso, 1995; McLean et al., 2004), ultimately leading to an accurate depiction of children's difficulties (Meyer et al., 2001).

Norm-referenced tests are usually administered by qualified professionals with specialized education and training who adhere strictly to standardized administration procedures; in addition, testing is usually conducted in defined and controlled settings, and the administration format is structured and follows direct testing procedures that attempt to elicit specific responses and behaviors from children (Sattler, 2008). As Sattler (2008) stated, (a) standardized administration procedures are designed to reduce the effects of professionals' personal biases and other possible extraneous variables that may affect young children's performance; (b) they are an economical and efficient means of quickly sampling children's behavior and functioning to identify those who have the greatest need of resources; (c) they are particularly valuable in evaluating behavioral deficits and strengths; (d) they conveniently and efficiently provide a baseline against which to measure young children's changes in performance and progress during interventions; and (e) they help examiners evaluate developmental changes and effects of interventions, which also allow for increased accountability. Building on a developmental perspective, norm-referenced developmental tests determine intraindividual and interindividual differences, measuring variations among young children in relation to reference groups while taking into consideration factors such as demographics and socioeconomic status (Sattler, 2008). This determination is beneficial because it allows examiners to establish realistic intervention goals, select specific areas in which interventions are needed, compare generated data to information provided by other sources such as parents' and teachers' reports, and identify typical and atypical key behaviors (Sattler, 2008). Pertinent to research, finally, normative data generated by norm-referenced tests assist researchers to determine group differences by comparing group samples across studies (Sattler, 2008).

LEGAL ISSUES

Please see Chapters 12 and 25, this volume; Volume 1, Chapter 38, this handbook; and Volume 2, Chapter 6, this handbook for detailed discussion of legal issues in psychological assessment. The significance of assessment as a means to early identification and subsequent intervention is also evident in the United States in the passage of multiple legislative acts safeguarding children's rights. The Individuals With Disabilities Education Improvement Act of 2004 (IDEIA) defined special education as specially designed

instruction to meet the unique needs of children with disabilities at no cost to parents. To be eligible for such modified instruction, children must demonstrate cognitive, physical, or behavioral impairments that interfere with their ability to learn the general education curriculum. IDEIA thus mandated that states evaluate all children suspected to have disabilities, including young children not yet enrolled in schools, to determine whether they are eligible to receive early intervention or special education services. *Evaluation* refers to procedures used to determine eligibility for special education services (IDEIA). Although the Education for All Handicapped Children Act of 1975 delineated the requirements of assessments (e.g., they should be psychometrically adequate and racially and culturally just), the Education for All Handicapped Children Amendments of 1986 extended these requirements to preschoolers with suspected disabilities. In particular, the amendments mandated states to conduct multidisciplinary assessments of preschool children to determine eligibility for special education. The Individuals With Disabilities Education Act Amendments of 1997, moreover, required states to assess and report the performance and progress of young children with disabilities.

States are required to conduct multidisciplinary comprehensive assessments when working with young children (IDEIA, 2004). Comprehensive assessments have traditionally used multiple sources of data, including scores generated by norm-referenced developmental measures (Bagnato, 2007; Bradley-Johnson, 2001; Brassard & Boehm, 2007; Rydz, Shevell, Majnemer, & Oskoui, 2005). AERA et al. (1999) indicated that norm-referenced developmental testing has the aim of providing differentiated profiles of young children by way of quantitative measurement. A general framework for using norm-referenced developmental testing includes determining a distinct developmental status, contrasting typical and atypical development, and assessing potential for further development (Petermann & Macha, 2008)

The advantages of using norm-referenced tests are well documented in the literature and were discussed in the preceding section. Norm-referenced tests provide valid and reliable quantitative data that can be used to determine eligibility for special education and related services as required by law (McLean et al., 2004; Sattler, 2008), to afford diagnosis, and to predict children's future performance (Bagnato & Neisworth, 1994; Flanagan & Alfonso, 1995; McLean et al., 2004). Standard scores thus differentiate between young children of the same age who can and cannot perform certain skills (Flanagan & Alfonso, 1995; Sattler, 2008).

There is, of course, a difference between the constraints and special considerations imposed by IDEIA on identification of a disability when one is evaluating preschool and school children and the guidelines for a diagnosis by, for example, the criteria of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text revision; American Psychiatric Association, 2000). Although federal regulations provide general guidelines in regard to identification and classification, states use differing policies and practices in definitions, classification criteria, assessment processes, and other considerations (Bergeron, Floyd, & Shands, 2008; Reschly & Hosp, 2004).

Although the advent of the Education for All Handicapped Children Amendments of 1986 created an increased use of preschool testing instruments (Brassard & Boehm, 2007), it did not mandate standardized acceptable technical criteria. Moreover, the *Standards for Educational and Psychological Testing* (AERA et al., 1999) provide only a frame of reference rather than specific technical criteria allowing practitioners to decide whether published criteria in the manuals are adequate. The lack of specific numerical criteria for, say, test–retest reliability is not altogether a bad thing because examiners must select the best test available for a particular purpose. The most reliable test for routine intellectual assessment of children with mild disabilities might have much higher reliability coefficients than the best available instrument for assessing the intellectual functioning of a child with severe cerebral palsy, including inaccurate pointing and unintelligible speech.

SELECTING A THEORETICAL BASIS FOR ASSESSMENT OF INTELLECTUAL FUNCTIONING

Selecting tests for assessment of children requires careful consideration. The chosen instruments must

be reliable and must be normed on an appropriate sample of children (usually a nationwide sample with random selection of examinees within stratification variables). Volume 1, Chapters 2 and 11, this handbook, discuss these issues in detail. Usually, the norm sample is intended to be a representative microcosm of the population of all children of selected ages within a nation or other specified region. The scores of the children in the norming sample provide a trustworthy yardstick for evaluating the scores earned by an examinee, even though the norming sample may have few, if any, children who share important characteristics with the examinee, such as cerebral palsy, deafness, or only recent exposure to the English language. On first consideration, it is tempting to contemplate norming a test on a special sample of children with, for example, a specific disability to make the norms more fair or more relevant for a particular examinee. However, assembling a nationwide sample of children that matched on all important variables with a national population of children with a particular diagnosis would be almost impossible. In fact, even defining unequivocal criteria for membership in a particular disability group would be difficult. For example, would mild ataxia qualify as cerebral palsy? How the examiner would interpret scores based on special norming samples is also not clear.

Most recently published test manuals have instead provided at least a little information about test scores of children in various special groups, usually compared with matched samples of children without disabilities. The samples used in these studies are usually small, so examiners must wait for larger studies to appear in the literature, but the data do serve to demonstrate the test's validity for differentiating various groups of children and to give at least some indication of anticipated score levels or patterns for children with certain disabilities or other characteristics.

The issue of a test's validity for differentiating groups of children raises the broader issue of test validity in general. This essential concern is discussed at length in Volume 1, Chapter 4, this handbook (see also, e.g., Watkins, 2009). Validity, of course, does not exist in a vacuum. A test may be valid for one purpose, but not for another, and it may be valid for a particular purpose with one group, but not with another. Validity is also relative, not absolute, so examiners must select tests that are the most valid for the intended purpose for the particular examinee. In some instances, that validity may be very strong, but in others, the best might not be very good. (Please see Volume 1, Chapters 4 and 13–15, this handbook, for extensive information on test validity.)

For children with specific disabilities (e.g., deafness or cerebral palsy), particular circumstances (e.g., English language learners), and various behavioral characteristics (e.g., shyness or short attention span), the format in which the items are presented can be extremely important. For example, a child with slow motor speed and poor visual–motor coordination would probably score lower on an intellectual assessment that included timed paper-and-pencil tests than on one that did not. If the examiner's goal is to assess the child's intellectual functioning, the motor speed and coordination issues would constitute construct-irrelevant variance (Messick, 1989). However, if the examiner believes that motor speed and visual–motor coordination are essential components of overall intellectual functioning, then that examiner would want to use a test that did include timed paper-and-pencil tasks.

Ay, there's the rub. The validity of a measure of intellectual functioning as such for a particular child depends on the examiner's or the reviewer's definition of intelligence or intellectual functioning. If, for example, one adheres to the Cattell–Horn–Carroll (CHC) model of intelligence (Carroll, 1997/2005; Flanagan et al., 2009; Horn & Blankson, 2005; McGrew & Flanagan, 1998; Schneider & McGrew, 2012; Woodcock, 1990), then one would want to include in an assessment of intellectual functioning all of the Stratum II broad abilities: fluid reasoning, crystallized ability, spatial thinking, long-term storage and retrieval, short-term memory, auditory processing, and processing speed (and perhaps correct decision speed and even reading and writing and mathematics achievement). One would need to either select an instrument designed to assess this wide array of abilities, such as the WJ III (Woodcock, McGrew, Schrank, & Mather, 2001/2007) or assemble a battery of tests and subtests to

measure all of the desired abilities (e.g., Flanagan et al., 2007).

However, if one believes that intellectual functioning is a matter of *g,* including only fluid and crystallized abilities (see, e.g., Cattell & Horn, 1978), then a test such as the RIAS (Reynolds & Kamphaus, 2003, 2005) is ideally suited to one's needs.

Other tests are based on different conceptualizations of intellectual functioning. For example, the DAS–II (Elliott, 2007a) includes only verbal ability, nonverbal (fluid) reasoning, and spatial ability in its total GCA score and assesses other abilities with diagnostic subtests and clusters that do not influence that GCA score (Dumont, Willis, & Elliott, 2008; Elliott, 2007b). The SB5 (Roid, 2003; see also Roid & Barrum, 2004; Roid & Pomplum, 2005, 2012) divides intellectual functioning into verbal and nonverbal domains, within which it assesses fluid reasoning, knowledge, quantitative reasoning, visual–spatial ability, and working memory.

The WISC–IV (Wechsler, 2003; see also Flanagan & Kaufman, 2009; Prifitera, Saklofske, & Weiss, 2005, 2008; Wahlstrom, Breaux, Zhu, & Weiss, 2012; Weiss, Saklofske, Prifitera, & Holdnack, 2006; Zhu & Weiss, 2005), WPPSI—Third Edition (WPPSI–III; Wechsler, 2002b; see also Lichtenberger & Kaufman, 2003; Wahlstrom, Breaux, Zhu, & Weiss, 2012), and Wechsler Adult Intelligence Scale (Wechsler, 2008; see also Drozdick, Wahlstrom, Zhu, & Weiss, 2012; Lichtenberger & Kaufman, 2009; Sattler & Ryan, 2009) are based on Wechsler's (1944) famous definition of intelligence: "the capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment" (p. 3). However, over their many versions and editions, the Wechsler scales have gradually given greater emphasis to separate index scores, which have expanded from Verbal and Performance IQ scales to Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed index scores, which—although developed from an array of research studies in cognitive neuroscience—more or less correspond to the CHC broad abilities of crystallized ability, fluid reasoning and visual–spatial thinking, short-term memory, and processing speed (Wechsler, 2002b, 2003, 2008).

Although CHC theory is the basis, or part of the basis, for many instruments or is at least acknowledged in the test manuals, entirely different theories form the basis for many tests. For example, the Cognitive Assessment System (Naglieri & Das, 1997a; see also Naglieri, 1999; Naglieri & Das, 1997b, 2005; Naglieri & Otero, 2012a) is based on the PASS model (Das, Kirby, & Jarman, 1975; Naglieri & Das, 2005; Naglieri, Das, & Goldstein, 2012) with Planning, Attention, Simultaneous, and Successive cognitive processes. Their work expanded and operationalized for intelligence testing the work of Luria (e.g., 1966, 1973, 1980). The KABC–II (Kaufman & Kaufman, 2004a; see also Kaufman, Lichtenberger, Fletcher-Janzen, & Kaufman, 2005; Singer, Lichtenberger, Kaufman, Kaufman, & Kaufman, 2012) uniquely offers two interpretative options, based on the same set of subtests. In the Luria system, the scales are called Sequential Processing, Simultaneous Processing, Learning Ability, and Planning Ability, and they yield a total score, the Mental Processing Index. In the CHC system, the same scales are called Short-Term Memory, Visual Processing, Long-Term Storage and Retrieval, and Fluid Reasoning, and a fifth scale (Knowledge or Crystallized Ability) is added to complete the Fluid–Crystallized Index (Kaufman & Kaufman, 2004a, p. 2).

If an examiner wants to assess intellectual functioning, it is very important that the test or battery selected represent the examiner's conceptualization of intelligence and that the examiner explains that conceptualization in the evaluation report (see Volume 2, Chapter 3, this handbook for extensive discussions of explaining assessment findings; see also Lichtenberger, Mather, Kaufman, & Kaufman, 2004; Sattler & Hoge, 2006, Chapter 25; and Chapter 23, this volume). Both the examiner and the reader of the report must be clear about how intellectual functioning is defined, and the test used must reflect that definition.

Some examiners use the McGrew, Flanagan, and Ortiz (Flanagan & McGrew, 1997; Flanagan, McGrew, & Ortiz, 2000; Flanagan et al., 2007) cross-battery approach to supplement preferred tests of intellectual functioning with other measures to assess the full range of CHC abilities. This approach

does not yield a total score based on all of the CHC broad abilities and has a strong following among trainers and practitioners, although it is not universally accepted (see, e.g., Ortiz & Flanagan, 2002a, 2002b; Watkins, Glutting, & Youngstrom, 2002; Watkins, Youngstrom, & Glutting, 2002).

A related issue is so-called *profile analysis,* a term usually used to describe and condemn one of two practices: generating profile templates of subtest scores on a test associated with certain disabilities or diagnoses or focusing attention on ipsative scores rather than purely normative ones (see, e.g., McDermott, Fantuzzo, & Glutting, 1990; McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992; Watkins, Glutting, & Youngstrom, 2002, 2005; Watkins & Kush, 1994).

We cannot find much to recommend (nor many instances in actual practice of) using profile templates, often based on very small differences between subtest scores, to diagnose particular disabilities. There are obvious problems with purely ipsative analyses of test scores in which each subtest score is marked by its deviation from the individual's mean score rather than by its deviation from the population mean. Such an analysis removes overall intellectual functioning from the scores and might even, if taken to extremes, fail to distinguish a child functioning in the range of intellectual giftedness from one functioning at the level of intellectual disability. Critics (e.g., Canivez & Watkins, 1998, 1999, 2001) have also questioned whether scores other than total test scores are sufficiently reliable to be used for analysis.

There are, however, in addition to the CHC cross-battery method mentioned earlier, approaches to the investigation of a pattern of, or relationships among, subtest scores that consider both the normative standing of the scores and their strengths and weaknesses. Most of these approaches are carried out top–down, beginning with the total, composite, or full-scale score before working down through component composites to individual subtests. Most of these approaches require statistically significant differences between scores, and many also include base rates or frequencies of differences in their analyses. Methods of analyzing test scores have been published at least since the Wechsler scales first provided subtest scores capable of being analyzed (e.g., Rapaport, Gill, & Schafer, 1945), but the most widely cited and used current methods date back to Sattler (1974), Kaufman (1979), and other authors in the 1970s who recommended methods of analyzing test scores that began with the most reliable scores and ended with the least reliable scores, that favored groups of subtest scores over scores from single subtests, and that required statistical tests of significance and frequency (base rate) for decisions. Many recent test manuals have included discussions of such approaches to analysis of test scores for the particular test.

Whether individual subtest scores or even clusters of two or more subtest scores add any reliable and useful information to that provided by the total score on a test of intellectual functioning has been debated. Because the subtest scores are all included in the total score, different methods of statistical analysis can give very different results about the contributions of subtest scores to scores for groups (clusters, scales, factors) of subtest scores (see, e.g., Fiorello, Hale, McGrath, Ryan, & Quinn, 2001; Hale, Fiorello, Kavanagh, Hoeppner, & Gaitherer, 2001; Watkins & Glutting, 2000; Watkins, Glutting, & Lei, 2007).

On the basis of our study of the question and our clinical experience, we do strongly recommend the cautious and statistically based analysis of a child's strengths and weaknesses on a test of intellectual functioning. We believe that valuable information can be gained from those data in addition to the total score on the test. However, others (e.g., Watkins, 2003) have disagreed, and examiners would do well to thoroughly study this issue.

DISABILITY, DISADVANTAGES, AND INTELLECTUAL FUNCTIONING

Despite the importance of selecting a test that faithfully reflects the examiner's conceptualization of intellectual functioning, instances will occur in which an examiner may believe that his or her definition of intellectual functioning, and a test selected to reflect that definition, may be unfair to a child because of a disability, disadvantage, or cultural difference. For example, no matter how much one

values crystallized verbal ability as a core component of intelligence, it would not be reasonable to use English-language questions and answers to assess the intellectual functioning of a child with little or no exposure to spoken English or a deaf child who has little oral communication ability. Similarly, a matrix reasoning–type test, no matter how valid a measure of *g* or of fluid reasoning for most children, would not measure intellectual functioning in a blind child, and timed block-design tests do not assess purely intellectual functioning in children with moderate to severe motor disabilities.

Consequently, the examiner must sometimes amend the selection of a test to accommodate disabilities, disadvantages, or differences that would invalidate the test as a measure of intellectual ability. Such decisions should also be explained in the report. Some tests offer options for such accommodations. For example, the WISC–IV (Wechsler, 2003) is designed to yield a FSIQ based on the Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed Indexes. However, Prifitera et al. (2005) provided tables for computing a General Ability Index score based on only the Verbal Comprehension and Perceptual Reasoning Indexes, and Saklofske et al. (2006) provided tables of Cognitive Processing Index scores based on the Working Memory and Processing Speed Indexes, both alone as well as with a table of base rates for differences between the General Ability Index and Cognitive Processing Index scores. These additional scores, based on the original WISC–IV norming sample, allow the examiner to consider intellectual functioning without including working memory and processing speed.

In some instances, examiners may want to use entirely nonverbal or at least nonoral measures of intellectual functioning. Such tests do omit an important area of intellectual ability, as discussed previously, but may afford a more accurate or more appropriate measure of intellectual functioning for a child who is deaf, a child with a diagnosed language disability, or a child with limited exposure to English. Chapters 4 and 11, this volume, provide detailed information on these issues. The Leiter–R, Universal Nonverbal Intelligence Test, and Wechsler Nonverbal Scale of Ability are strong examples of comprehensive nonverbal (or nonoral) tests that use more than one item format and assess more than one cognitive ability. See Braden and Athanasiou (2005), Brunnert, Naglieri, and Hardy-Braz (2008), McCallum (2003), and McCallum, Bracken, and Wasserman (2001) for additional discussion of nonverbal assessment.

The issue of nonverbal (or nonoral) intelligence tests omitting an important aspect of intellectual functioning raises a broader issue. If the purpose of intellectual assessment is simply to predict school achievement or other life outcomes, such as professional success or income, a test that is unfair to a child because it penalizes a disability or disadvantage may actually perform its intended function better than an assessment of intellectual functioning that bypasses those areas of impairment or disadvantage. For example, the same disability that depresses the total score on an assessment of intellectual functioning is also likely to interfere with school achievement and earning potential (especially to the extent that special education is unsuccessful for that child). From that standpoint, some might argue that tests should be selected, administered, and interpreted without regard to the child's disabilities, disadvantages, or differences.

Although we understand this argument, we do not agree with either the premise or the conclusion. First, we believe that the purpose of assessment of intellectual functioning is more than simply predicting academic or vocational success. We believe that the primary purpose is to understand and explain the child's intellectual functioning to allow teachers, therapists, parents, and others to assist the child as needed. We believe that the purpose of special education is to help the child surpass the pessimistic predictions based on the child's disability. Therefore, we want to measure intellectual functioning separately from the effects of disabilities, disadvantages, and differences. This is not to say that those factors are not important or that they do not affect academic and vocational progress. However, we do believe that they are separate from intelligence and that intelligence should be measured separately from those factors. Although many authorities would not agree with us, we urge examiners to try

to assess intelligence or thinking ability independently of the effects of disabilities, disadvantages, and differences.

EVALUATION SESSION

Assessment of children presents special considerations for the evaluation session. Some of these considerations are caused by the special characteristics of children, others by the fact that children usually attend schools.

Testing Room

If you see children in your own office, you are responsible for providing a quiet, distraction-free environment with comfortable, stable furniture appropriate for the child's height and leg length. Some children require special lighting, special seating cushions, slant boards for writing, and other accommodations, which you should have available. Often, however, evaluations of children are carried out in schools with limited space available for such activities. It is important to insist on adequate testing conditions—even threatening to describe the testing conditions in clear detail in the evaluation report (Joel P. Austin, personal communication, October 10, 1980). It is not prudent to assume that adequate testing conditions will routinely be available when you arrive at a school. Allow time before the scheduled evaluation for procuring a suitable room and furniture. It is important that the child's feet not dangle in the air and that the examiner not sit in a higher chair, towering over a shy child. (One of us carries folding, child-size chairs and a table in his car.) We have also learned that it is important to find out in advance about scheduled fire drills and other potential interruptions, which always seem to occur during timed tests.

Sadly, for self-protection, it is also prudent to balance the child's need for privacy and lack of distractions with your need to be protected from false allegations of misconduct with the child. We never lock the door to a test room, and Guy M. McBride (personal communication, July 11, 2010) has recommended ostentatiously rattling the doorknob to demonstrate to the child that it is not locked. We try never to use a test room without a window in the door. Seating the child with his or her back to the window diminishes distractions and puts the child closer to the door than the examiner in case the child wants to leave at any point.

It is best practice to find out from parents, teachers, and therapists what physical accommodations may be needed for the test session. Children with severe and multiple disabilities often have paraprofessionals who work with them individually throughout the day in most or all settings. Schools often fail to put evaluators in contact with such paraprofessionals, so it is up to you to initiate communication with those individuals, who often have the most extensive and most practical information about working with the child.

Sessions

Although each child is an individual, you need to be alert to certain frequently encountered issues. Children are often unclear, totally ignorant, or misinformed about the purposes and possible results of testing. It is wise to discuss these issues even if you think the child has been adequately prepared. It often helps to begin by asking the child what he or she thinks the testing is all about. Children often assume the testing will determine special education placement, retention in grade, or accelerated promotion to the next grade. It is important not to lie to the child, although you may temporize by explaining that the issue in question is a team decision, not a direct result of the test scores. Test manuals usually offer some advice about testing children. Textbooks on assessment, such as Gutkin and Reynolds (2009) and Sattler (2008, Chapters 1 and 6), have discussed these issues at length.

Test manuals usually provide a standardized introduction to the test, which must be used as presented. You may, however, need to provide additional information before reading the standardized introduction and beginning the test. You can explain that your job is helping teachers teach more effectively (and even mention any college or graduate teaching of teachers you may have done). Sometimes it is helpful to explain that teachers want to teach things the child does not know and skip things the child already knows, so both correct and incorrect answers are useful. It is extremely important

not to let the child figure out that a certain number of incorrect responses terminates a subtest. We have actually had previously tested children ask, "How many more do I have to miss so we can stop?"

The child's attention span needs to be monitored closely. Many children continue to demonstrate peak performance for surprising lengths of time with child-friendly tests and frequent shifts from one task and one type of task to another. Most children's tests are now designed with frequent changes of activity. However, when a child becomes bored, tired, or inattentive, it is essential to terminate testing and resume at another time. This practice can be terribly inconvenient and expensive (for you, if you bill by the case, and for the purchaser, if you bill by the hour), but the alternative is invalid results.

Verbatim recording of the child's responses (both correct and incorrect) is essential. You (and perhaps others) will need to recheck your scoring; psychologists are notoriously poor clerks and commonly make errors when scoring test protocols (e.g., Alfonso, Johnson, Patinella, & Rader, 1998; Klassen & Kishor, 1996; Watkins, 2009). Also, the content of specific responses may turn out to be important in light of later information. Recording correct responses also prevents the child from learning that your writing always indicates an incorrect answer. We find it helpful to mark the starting time for each subtest on the record form near the subtest title. Questions sometimes arise later about which subtests were administered before the child's medication usually wears off or which was the last subtest before recess or the first after lunch. The start times also allow you to compute the total time for a subtest in case that becomes an issue of interest. We also find it helpful to unobtrusively make one pencil mark every second while awaiting the child's response to a question on an untimed test. As you gain more information and as new questions arise, the child's usual response pace and even latencies in responding to particular questions may become important. It is, of course, valuable to record comments unrelated to the test and the child's behaviors. However, it is unwise to note a behavior that occurs only once or twice without also noting that it did not occur again. Single instances of a behavior that might be diagnostically important if it was frequent could be very common. A single, unelaborated recording of such a behavior might take on undue importance when you, or someone else, later reviews the test record. Again, Chapters 12 and 25, this volume, and Volume 1, Chapter 38, and Volume 2, Chapter 6, this handbook, discuss legal issues more extensively. We should note, however, that several circumstances may result in the test record forms becoming part of a public record. Never write anything on a record form that you are not willing to share with attorneys and others.

The pace of testing must be adapted to the child's needs. Young children are especially intolerant of pauses between items or subtests while the examiner scribbles notes. Audio recording of test sessions can sometimes be helpful, although there is always the risk of equipment malfunction. If you are likely to lose the child's attention or cooperation during a pause, the audio recording even allows you to preserve comments such as "You stacked seven cubes before the tower fell over" for future reference. Some children respond badly to a sense of being hurried, and you must slow the pace of testing to accommodate the child's needs without letting your impatience show.

It often helps to balance the child's mood with an opposite response on your own part. If a child's exuberance is on the verge of spinning out of control, you need to become slower and calmer. The examiner should not join a glum, pessimistic child in a slough of despond.

Testing in schools also raises the issue of the classes the child misses for testing. Arrangements need to be made to ensure the child is not penalized (e.g., teacher hostility, a lowered grade, or additional homework) for missing class; that the child does not miss an important exam or essential class session; and that the child is not excessively upset by missing a favorite class. One of us (John O. Willis) tested a child who seemed to have serious depression and mild mental retardation. When she began to cry, questioning revealed that the school secretary had taken the child out of gym class for testing and that the child's team had been winning a volleyball game. When the child was rushed back to the gym, she rejoined her team and helped them

snatch victory from the jaws of defeat. Subsequently, she turned out to be a cheerful youngster with high average intellectual ability. Some teachers and administrators will disapprove of your concern that the child not miss favorite classes. Nonetheless, it can be important.

Young children tend to respond well to toylike objects, blocks, and other materials they can manipulate (*manipulatives*) as well as colorful pictures. Tension can result between the child's expectations of a game and the examiner's compulsion to complete the standardized assessment. It is prudent to ask parents and teachers not to tell children they are going to play games with you. Many test manuals wisely warn examiners not to let children play with test materials but to instead bring age-appropriate toys in case play is necessary.

Some children like to have a list of tasks that will be completed. For younger children, you can use pictures or draw squares on paper to represent the tasks and then give the child a sticker to place on each picture or in each square as the task is completed.

Children usually want to know the overall schedule for testing and to be reassured that they will not miss recess, lunch, favorite activities, special events, or the bus home. If there is any possibility of additional sessions, the child should know that in advance. It is never prudent to promise that a test session is going to be the last. New issues may arise.

With older children and adolescents, it is important to review ground rules at the outset of testing. It is important that the child understand the limits of confidentiality and what will be done with the results of the assessment. Older children and adolescents often want to know what's in it for them before committing to the assessment. They may also need to understand the possible consequences of blowing off the tests and earning low scores. Children will often ask questions about the purpose of the test and other technical information. Some of those questions can be answered at the end of the assessment, but discussing subtests during the assessment might change the difficulty or the nature of the subtest.

It is important to explain in advance that you are usually not allowed to tell whether answers are correct. You should not review with the child items after the test because the child might take the same or a similar test again. You may wish to arrange to discuss the general results after the testing, but do not commit to any particular form of debriefing, such as test scores. If you introduced the assessment by discussing the intention to identify individual strengths and weaknesses, you may be able simply to discuss in a positive light which tasks the student did best on and which were most difficult without reference to norms. A more detailed discussion of results requires considerable planning and coordination with parents, teachers, and therapists.

Accommodations and Adaptations

If the child has a significant hearing or vision impairment, a school-based or itinerant teacher of children who are deaf and hard of hearing or of children with visual impairments (titles vary from state to state) should have essential information about the child's capacities and needs. Consultation before the assessment will allow the examiner to select an assessment of intellectual functioning that can be used with the fewest and smallest possible accommodations and to plan and prepare for any accommodations that will still be needed. A functional vision assessment before your assessment is essential. It is better—within the limits of using tests that are reliable and that are proven valid for the intended purpose—to adopt an appropriate test than to try to adapt a less appropriate one. Please see, for example, Chapters 20 and 21 in Sattler and Hoge (2006) and the discussions of testing children with sensory and motor disabilities in various test manuals for tests. The DAS–II (Elliott, 2007a), for example, includes a CD with a video of signed standard sentences to show examiners fluent in American Sign Language the standardized presentation of the DAS–II instructions. It is easy to rule out some tests or subtests at first glance. For example, picture vocabulary tests cannot be adapted validly for use with students who are totally blind, and oral vocabulary tests lose validity when translated into American Sign Language (or any other language, for that matter). It is helpful to select a test whose manual includes an extensive discussion of your examinee's particular disability and, as is increasingly true, includes in the validity section of the manual data

on test scores obtained by children with your examinee's disability. For just one example, the SB5 includes an 11-page appendix (Roid, 2003, pp. 311–322), "Use of the Stanford–Binet Intelligence Scales, Fifth Edition, With Deaf and Hard of Hearing Individuals: General Considerations and Tailored Administration."

The issue of test accommodations is difficult (Geisinger, 1994b, mostly addressing large-scale group tests). Psychologists want to make a test as fair as possible for the child simply as a matter of fairness as well as legal protections. They also want to ensure that they are truly measuring the construct they intend to measure (Messick, 1989). Psychologists also do not want to alter the difficulty level of a test item. There is a distinction, albeit one that is difficult to define precisely in actual practice, between making an accommodation that may threaten the validity of the test norms and making an adaptation that changes the administration of the test without apparently altering the constructs that are measured or changing the difficulty of any item. (This use of the terms *accommodation* and *adaptation* is not universal. Be sure to understand how the author is using these terms whenever you encounter them.) To take just one example, consider altering a vocabulary test to a multiple-choice format for a child who cannot speak intelligibly. The WISC–IV (Wechsler, 2003), for example, includes a Vocabulary subtest that, for most items, requires the child to state a definition of a word spoken and presented in print by the examiner. It would clearly be an impermissible accommodation for the examiner to make up multiple-choice definitions and ask the child simply to select one for each word. The format would have been totally changed, and the effect on the difficulty of each test item would not be known. Wechsler et al. (2004), however, have standardized two such versions: one with four pictorial choices and one with four possible verbal definitions for each word. These WISC–IV Integrated subtests use the same words as the WISC–IV. These versions are standardized and are normed on a smaller but nationally stratified sample: 730 children (as opposed to the 2,200 in the WISC–IV norming sample). Electing to use the WISC–IV Integrated multiple-choice subtests instead of the WISC–IV subtest would be a case of adopting an appropriate test format (albeit one with a smaller norming sample) rather than adapting a test with an inappropriate format for your examinee. You might also elect to assess oral, receptive vocabulary with an entire test designed with a multiple-choice format, such as the Peabody Picture Vocabulary Test, fourth edition (Dunn & Dunn, 2007).

Another approach would be to use part of an existing test. For example, the Kaufman Brief Intelligence Test—Second Edition (Kaufman & Kaufman, 2004b) includes a single-subtest Nonverbal Scale that is multiple choice and a Verbal Scale with two subtests, one of which is also multiple choice. Two of the three subtests could be used, but the Verbal Scale and the IQ Composite scores would be lost.

Now, suppose that a child also cannot point very accurately. You have found a multiple-choice vocabulary test that does not require an invalidating accommodation of the basic test format for the child, but the choices printed in the test booklet are much too close together to allow this child to reliability indicate her or his choices. You could then arrange various adaptations, such as sacrificing an expensive test book by cutting apart the choices and spreading them out; hanging pieces of paper with numbers or letters for the choices on the wall (which would require the cognitive ability to relate those marked papers to the choices in the booklet); arranging the child's computer keyboard or electronic communication device to permit only the four, five, or six choices used by the test (again requiring the child to be able to associate the two sets of choices); or using an eye-gaze communication system (e.g., writing the four, five, or six choices on the edges of a sheet of transparent plastic, holding it between yourself and the child, and asking the child to stare at the intended choice). (When using such systems with a child who cannot speak intelligibly, it is prudent to include a fail-safe option that allows the child reject your confirmation of what you thought the child indicated.)

These are examples of permissible adaptations that would not alter the fundamental nature of the task and that would not make any test item easier than it was for the children in the standardization sample. If anything, the task is more difficult when

using such methods. It would be essential to describe your adaptations in the report and to explain why you believe that they did not alter the tests or invalidate the test scores.

Geisinger (1994b) noted that "many individuals with disabilities develop compensatory skills to help them offset weaknesses associated with their particular disabilities" (p. 136). In suggesting that the makers and users of large-scale group tests assess those skills that are systematically developed by "individuals with specific classes of disabilities" Geisinger suggested that "when an individual assessment is to be made, few would argue that we should attempt to assess those compensatory skills and abilities so we can obtain a more complete understanding of the individual's level of functioning" (p. 136). In fact, our sad experience is that many evaluators do not do that. Reports of assessments of individual intellectual functioning would be much more useful if evaluators did explore the child's compensatory skills and did report those developed skills and new compensations that worked effectively in the assessment. Similarly, if the examiner succeeds in helping a child successfully focus and sustain attention or modulate disruptive behavior, parents and teachers might appreciate a description of the examiner's successful techniques.

Examiners' Limitations

We discussed earlier the need to be familiar with the behavior, interests, and culture of children similar to the one you are testing. This knowledge is important both for establishing rapport with the child or adolescent and for recognizing that seemingly bizarre behavior may be well within the current norm for children of that age. You also need to recognize your physical limitations. As we age, we are becoming aware that the faint, high-pitched voices of young children are becoming more difficult to hear and that sitting in very low chairs or on the floor is more challenging than it used to be. Some examiners simply do not like uncooperative small children or hostile adolescents. There may be times when an examiner simply has to request that a child be tested by a colleague if the alternative is an assessment of questionable validity.

RECOMMENDATIONS IN REPORTS ON ASSESSMENTS OF INTELLECTUAL FUNCTIONING

In most instances outside public school settings, psychologists routinely offer specific recommendations for diagnosis and treatment in their reports (see Chapter 23, this volume). The issue is more complicated when an examiner makes a report within the public school special education system. Because of different interpretations of legal requirements and liabilities, states, local education agencies, and building-level administrators often impose different rules on examiners working for them. Some demand specific recommendations as part of the evaluation process (even when the examiner has done only the assessment of intellectual functioning and has not seen the concurrent assessments of educational achievement, speech and language functioning, social and emotional adjustment, and other essential areas). Many administrators do not permit any recommendations at all (for fear that the local educational agency might be required to follow any and all recommendations suggested by the various examiners for a student). Yet others permit or demand certain categories of recommendations but forbid others (perhaps identification of specific disabilities, placement, or specific instructional programs by name). An examiner working in a public school system needs to become familiar with the local rules. Assuming that recommendations are permitted, there are at least three good ways of offering recommendations. These are discussed next.

Comprehensive Recommendations in the Report

One approach is simply to include specific recommendations in the report. The scope of those recommendations must be limited by the examiner's knowledge. First, the examiner's knowledge of the child is limited by the background information that was made available and by the scope of the evaluation. A comprehensive evaluation, including intellectual functioning, academic achievement, and social–emotional assessment, will obviously provide more information than a 90-minute assessment of intellectual functioning. It would, for example, be

presumptuous and unwise to offer recommendations for specific reading instruction programs on the basis of only an IQ test. Second, the examiner's ability to offer specific recommendations is limited by the examiner's knowledge of specific interventions. We have seen, for instance, far too many evaluations in which examiners with limited expertise in reading instruction have confidently recommended a specific remedial program by name apparently because it is the only remedial reading program of which the evaluator is aware. In making educational recommendations, as in all professional activities, psychologists must work within the limits of their specific professional expertise.

Collaborative Recommendations

If they are invited to the evaluation team meeting to review the child's assessment, many examiners prefer to work collaboratively in that meeting with parents, teachers, specialists, and administrators to cooperatively develop a comprehensive and detailed set of recommendations based on all of the information that has become available. This approach is likely to produce more effective, more comprehensive, and better integrated recommendations for the benefit of the child. The recommendations will be based on more information, and team members, including parents, may be able to work together to produce better recommendations. This approach also gives more ownership to the teachers, therapists, parents, and administrators who will have to carry out the recommendations or provide administrative support for them. However, this approach does have risks. First, it is absolutely essential that the recommendations be recorded in specific detail, that they be preserved where they will remain accessible, and that they be incorporated completely in the meeting minutes, the written prior notice, or both as well as in the child's regular education action plan, formal plan of special accommodations under §504 of the Rehabilitation Act of 1973, or the Individualized Education Program. Otherwise, all the effort will be for naught. The other risk is that an examiner's recommendation (that would have been permanently recorded in the evaluation report in the first approach described) might be rejected, vetoed, or watered down in the meeting process. A powerful administrator who would have been stuck with a clear recommendation in the evaluation report might be able to weaken or eliminate the recommendation through the team meeting process.

Two Reports

Finally, some evaluators write a preliminary report for the meeting and a final report based on the additional information revealed at and the recommendations developed in the meeting. This method tends to yield a better report, and recommendations can be explicitly attributed to the people who offered them, who may be the same people who will be implementing them—an incentive for treatment integrity. The disadvantages of this approach are that the second report may not be perceived as fully independent, may not be distributed and preserved, and, absent a deadline, may not be written quickly or at all.

Additional Issues With Recommendations in Reports

A recommendation is usually more clear and more persuasive if it includes the rationale for the recommendation. An evaluator's simply stating that some accommodation or method of special instruction should be used is not very persuasive. If the examiner explains the reason for the recommendation, connecting it to a reported test observation, a test score, or other information in the report, the reader is more likely to understand precisely what is being recommended and may be more likely to accept the recommendation.

Reasons Given for Not Permitting Recommendations

Some school administrators hold to the mistaken belief that recommendations should not be permitted in reports or even in meetings because the school district would be held liable for following all recommendations made by all evaluators. Faced with this argument, examiners can first ask the administrator to show where such a rule appears in the federal regulations for special education. To the very best of our knowledge, this rule does not appear in any federal special education regulations, and we have yet to see such a rule in any state

regulations. Second, such a rule would be impossibly illogical. Suppose one evaluator recommends that the child be taken out of regular class for 90 minutes a day of intensive, individual, Orton-Gillingham reading instruction; another evaluator on the team insists that the child must be sent to a private, special education boarding school; and a third evaluator is adamant that the child must have full inclusion with whole language reading instruction, never setting foot outside the regular classroom and never working in a group of fewer than five children. Obviously, the school cannot possibly follow the three sets of mutually exclusive recommendations. Therefore, it is equally as obvious that the school cannot be required to follow any of them simply because someone wrote them.

A SAMPLING OF A FEW SPECIFIC INSTRUMENTS

Any list of particular tests will be obsolete by the time it is typed, let alone by the time it is published. As the field of intellectual assessment continues to grow and change (please see, e.g., Chapter 28, this volume, and Daniel, 1997), new editions and entirely new tests continue to appear. Far too many tests are on the market today to review or even mention them in a single chapter. The following sections include some of the more commonly used instruments as of August 2012, but space limitations require many arbitrary choices. The omission of a test should not be taken as a condemnation of it (it might merely reflect our ignorance), nor should the inclusion of one imply our whole-hearted approval of all aspects of the instrument.

Cognitive Assessment System

The Cognitive Assessment System (Naglieri & Das, 1997; see also Naglieri, 1999; Naglieri & Das, 1997b, 2005; Naglieri & Otero, 2012a) is an individually administered test of cognitive ability for children ages 5 years through 17 years, 11 months. The test includes 12 subtests and can be administered in two forms. The standard battery consists of all 12 subtests, and the basic battery is made up of eight subtests. Administration times are 60 minutes and 45 minutes, respectively.

The Cognitive Assessment System, based on the PASS model (Das et al., 1975; Naglieri & Das, 2005; Naglieri, Das, & Goldstein, 2012), which derives in part from Luria's theories (e.g., Luria, 1966, 1973, 1980), is represented by four scales representing planning, attention, simultaneous, and successive cognitive processes. *Planning* is the ability to conceptualize and then apply the proper strategies to successfully complete a novel task. The individual must be able to determine, select, and then use a strategy to efficiently solve a problem. Attention is a cognitive process by which an individual focuses on one cognitive process while excluding extraneous competing stimuli. Simultaneous processing is the integration of stimuli into a coherent whole. Successive processing involves organizing various things into a specific sequential order. A Full Scale score can also be obtained from the data.

Standard scores are provided for all subtests, with a mean of 10 and a standard deviation of 3. The four scales, along with the Full Scale score, are reported as standard scores with a mean of 100 and a standard deviation of 15.

Examiners who find Luria's (1966, 1973, 1980) theories especially helpful and examiners seeking a test structure and content different from Wechsler, Binet, and CHC formulations may find the Cognitive Assessment System useful, especially when the PASS theory is most likely to answer the specific referral questions for a particular student.

Comprehensive Test of Nonverbal Intelligence, Second Edition

The CTONI–2 (Hammill et al., 2009) measures nonverbal reasoning abilities of individuals ages 6 through 89. Because the CTONI–2 contains no oral responses, reading, writing, or object manipulation, it is presented as being particularly appropriate for students who are bilingual, speak a language other than English, are socially or economically disadvantaged, are deaf, or who have a language disorder, motor impairment, or neurological impairment.

The CTONI–2 measures analogical reasoning, categorical classification, and sequential reasoning in two different contexts: pictures of familiar objects (people, toys, and animals) and geometric designs (unfamiliar sketches, patterns, and drawings). There

are six subtests in total. Three subtests use pictured objects, and three use geometric designs. Examinees indicate their answers by pointing to alternative choices. The CTONI–2 provides three composite IQs: Nonverbal Intelligence Quotient, Pictorial Nonverbal Intelligence Quotient, and Geometric Nonverbal Intelligence Quotient. A computer-administered version of the previous edition of this nonverbal intelligence test is also available.

The CTONI–2 offers the advantage of completely nonoral administration. The lack of verbal subtests is obviously helpful for testing children with limited oral language or English language abilities, but it also limits the scope of the assessment. The consistent multiple-choice format can be an asset for children who have difficulty learning or shifting between tasks, but again it limits the scope of the assessment. See also the Test of Nonverbal Intelligence (4th ed.; Brown, Sherbenou, & Johnsen, 2010), which is a recently normed, nonoral multiple-choice test and offers two parallel forms.

Differential Ability Scales, Second Edition

The DAS–II (Elliott, 2007a, 2007b; see also Dumont et al., 2008; Elliott, 2005, 2012; Sattler, 2008) is an individually administered measure of cognitive ability (which Elliott has refused to call *intelligence*) designed to measure specific abilities and assist in determining strengths and weaknesses for children and adolescents ages 2 years, 6 months, through 17 years, 11 months. The DAS–II is composed of 10 core cognitive subtests and 10 diagnostic subtests. The core subtests are used to calculate a high-level composite score called the *GCA score* ("the general ability of an individual to perform complex mental processing that involves conceptualization and transformation of information"; Elliott, 2007b, p. 17) and three lower level composite scores: Verbal Ability, Nonverbal (Fluid) Reasoning Ability, and Spatial Ability cluster scores. With lower *g* loadings, the diagnostic subtests are used predominantly to assess strengths and weaknesses and do not contribute to the composite scores. They yield three cluster scores: Processing Speed, Working Memory, and School Readiness. A Special Nonverbal Composite, based on only the Nonverbal (Fluid) Reasoning and Spatial Ability clusters, is also available for ages 3 years, 6 months, to 17 years, 11 months.

The DAS–II uses standard scores ($M = 100$, $SD = 15$) for the composite scores and *T* scores ($M = 50$, $SD = 10$) for the 20 individual subtests. The GCA is highly *g* saturated, the time of administration is relatively short, the test is adaptive in nature, and children as young as ages 2 and 3 can be assessed. Overlapping age ranges for the Lower Early Years (with fewer subtests and clusters), Upper Early Years, and School-Age batteries and out-of-level norms allow considerable flexibility in standardized testing of low- and high-scoring children. Examiners who agree with Elliott (2007b) that verbal, fluid reasoning, and spatial abilities are core intellectual abilities and that other important cognitive functions with lower *g* loadings should be measured but not included in the total score will find the DAS–II especially appealing. Obviously, examiners who believe that other cognitive abilities must be included in a total score will be inclined to use other instruments.

Kaufman Assessment Battery for Children, Second Edition

The KABC–II (Kaufman & Kaufman, 2004a; see also Kaufman et al., 2005; Singer et al., 2012) contains a total of 18 subtests grouped into core or supplementary tests. It has two interpretative models: the CHC (Carroll, 1997/2005; Horn & Blankson, 2005) and the Luria (1966, 1973, 1980). The core subtests have individual scaled scores and are used to compute either the CHC Fluid–Crystallized Index or the Luria Mental Processing Index, and the supplementary subtests provide expanded coverage of the abilities measured by the core KABC–II subtests and allow for the computation of a Nonverbal Index. At all ages except 3 years, 0 months, to 3 years, 11 months, the subtests not only combine to produce the Global Index scores (Fluid–Crystallized Index or Mental Processing Index) but also yield as many as four (Luria model) or five (CHC model) indexes. These index scores represent Sequential Processing–Short-Term Memory, Simultaneous Processing–Visual Processing, Learning Ability–Long-Term Storage and Retrieval, Planning Ability–Fluid Reasoning, and Crystallized Ability. This last (Crystallized Ability) is represented only in the CHC model. Different subtests are used to compute the

scales at different ages (3, 4–6, 7–18). The KABC–II uses standard scores ($M = 100$, $SD = 15$) for the five scales and the three Global Indexes and scaled scores ($M = 10$, $SD = 3$) for the 18 individual subtests.

Some examiners will welcome the option of choosing either the Luria or the CHC interpretation and the availability of a Nonverbal Index. Kaufman and Kaufman (2004a) stated that "measures of [Crystallized Ability] should be excluded from any score that purports to measure a person's intelligence or overall cognitive ability whenever the measure of [Crystallized Ability] is not likely to reflect that person's level of ability" (p. 4) and that "an examiner with a firm commitment to the Luria processing approach [would believe] that acquired knowledge should be excluded from any global cognitive score" (p. 5).

Kaufman Brief Intelligence Test, Second Edition

The Kaufman Brief Intelligence Test—Second Edition (Kaufman & Kaufman, 2004b; see also Homack & Reynolds, 2007) is an individually administered test of verbal and nonverbal ability for people ages 4 through 90. The Kaufman Brief Intelligence Test—Second Edition takes approximately 20 minutes to administer and consists of two scales, Verbal and Nonverbal. The Verbal scale is composed of two parts, Verbal Knowledge and Riddles, and the Nonverbal scale contains the subtest Matrices. For Verbal Knowledge, the individual is asked to point to one of six pictures to match a vocabulary word spoken by the examiner or to answer a question of general knowledge. Riddles requires the examinee to answer oral questions that require both knowledge and logical reasoning. Matrices is a nonverbal test in which the individual looks at a sequence or pattern and then selects the one of five or six alternative pictures or abstract designs that best completes the logical pattern.

All subtests are administered using an easel. The items are in color and are designed to appeal to children. The Kaufman Brief Intelligence Test—Second Edition provides standard scores ($M = 100$, $SD = 15$) for both the subtests and the resulting IQ composite. Tables are provided for statistical significance and base rates for differences between the Verbal and Nonverbal scores.

The Kaufman Brief Intelligence Test—Second Edition is, as stated, a brief intelligence test. Brief tests are especially valuable when an assessment is directed at broader purposes and the examiner wants to check intellectual ability as one part of the assessment.

Leiter International Performance Scale—Revised

The Leiter–R (Roid & Miller, 1997; see also Braden & Athanasiou, 2005; McCallum et al., 2001) is an individually administered nonverbal test designed to assess intellectual ability, memory, and attention functions in children and adolescents ages 2 years to 20 years, 11 months. The Leiter–R consists of two groupings of subtests: the Visualization and Reasoning Battery, consisting of 10 subtests (four Reasoning and six Visualization–Spatial), and the Attention and Memory Battery, also consisting of 10 subtests (eight Memory and two Attention). It also includes four social–emotional rating scales (Examiner, Parent, Self, and Teacher) that provide information from behavioral observations of the examinee. The majority of Leiter–R items require the examinee to move response cards into slots on the easel tray. Some items require arranging foam rubber shapes or pointing to responses on the easel pictures.

Raw scores on the subtests and rating scales are converted to scaled scores ($M = 10$, $SD = 3$), and Brief and Full Scale IQs are calculated from sums of subtest scaled scores and converted to IQ standard scores ($M = 100$, $SD = 15$). Composite scores can also be obtained for Fluid Reasoning, Fundamental Visualization, Spatial Visualization, Attention, and Memory.

The Leiter–R requires no spoken language by either the examiner or the examinee. Instructions are given by pantomime and facial expression. The Leiter–R shares the strengths and limitations of other nonoral tests. It has a wider variety of tasks than other nonoral tests, which makes it more cumbersome, but it also provides a rich measure of abilities assessed in different ways. The conormed (with a smaller norming sample) Attention and Memory battery provides considerable additional information.

Reynolds Intellectual Assessment Scales

The RIAS (Reynolds & Kamphaus, 2003; see also Reynolds & Kamphaus, 2005; Reynolds, Kamphaus, & Raines, 2012) is an individually administered test of intelligence assessing two primary components of intelligence, verbal (crystallized) and nonverbal (fluid). Verbal intelligence is assessed with two tasks (Guess What and Verbal Reasoning) involving verbal problem solving and verbal reasoning. Nonverbal intelligence is assessed by visual fluid reasoning and spatial ability tasks, Odd-Item Out and What's Missing. These two scales combine to produce a Composite Intelligence Index. In contrast to many existing measures of intelligence, the RIAS eliminates dependence on motor coordination, visual–motor speed, and reading skills.

A Composite Memory Index can be derived from two supplementary subtests: Verbal Memory and Nonverbal Memory. These short-term memory assessments require approximately 10 minutes of additional testing time. Reynolds and Kamphaus (2003) reluctantly provided norms for a combined score including both the Composite Intelligence Index and Composite Memory Index scores, but discourage examiners from reporting that combined score. The three nonverbal subtests have time limits for each item and provide for a second chance for half credit if the examinee selects the wrong choice on the first try.

The subtests are reported as *T* scores, and the indexes are reported as standard scores ($M = 100$, $SD = 15$). Although the RIAS is brief, it is not offered as a brief test. Reynolds and Kamphaus (2003) listed as the first of their eight "goals for the development of the RIAS. . . . Provide a reliable and valid measure of *g* and its two primary components, verbal and nonverbal intelligence, with close correspondence to crystallized and fluid intelligence" (p. 1). The 259-page manual includes a brief form, the Reynolds Intellectual Screening Test (Reynolds & Kamphaus, 2003; see also Homack & Reynolds, 2007), consisting of only the Guess What and Odd-Item-Out subtests. Examiners who want to efficiently assess intelligence through only crystallized verbal and nonverbal (fluid reasoning and visual) measures would find the RIAS valuable for this purpose.

Stanford–Binet Intelligence Scale, Fifth Edition

The SB5 (Roid, 2003; see also Roid & Barrum, 2004; Roid & Pomplum, 2005, 2012) is an individually administered test of cognitive abilities for ages 2 to 85. The FSIQ is derived from the administration of 10 subtests (five verbal and five nonverbal). Subtests are designed to measure five factors: Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual–Spatial Processing, and Working Memory. SB5 subtests are composed of "testlets"—brief minitests at each level (1–6) of difficulty. The SB5 also provides examiners the option of calculating change-sensitive scores—a method of criterion-referenced rather than normative-referenced scoring, which avoids truncation at high and low ends, as well as an extended IQ—a special-case application for evaluating subjects with extremely high (or low) IQs.

The SB5 FSIQ and five factor scores have a mean of 100 and a standard deviation of 15. Individual subtests use scaled scores with a mean of 10 and a standard deviation of 3.

The SB5 uses routing tests (one verbal and one nonverbal) to determine the starting-level testlets for the other four verbal and other four nonverbal subtests. Basal scores and ceilings are determined independently for each subtest, so examiners may administer as few as one or as many as four verbal or nonverbal testlets at each level. This procedure is a little different from that used by other tests of intellectual functioning. The nonverbal subtests do involve some oral language; they are not purely nonoral as are the Leiter–R, Universal Nonverbal Intelligence Test, CTONI–2, or Test of Nonverbal Intelligence (4th ed.), for example. Examiners who want to include crystallized ability, working memory, visual–spatial ability, and separate measures of fluid reasoning and quantitative reasoning (rather than subsuming quantitative reasoning under fluid reasoning) would find the SB5 appropriately structured for their needs.

Universal Nonverbal Intelligence Test

The Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998; see also McCallum & Bracken, 2005, 2012; McCallum et al., 2001) is an individually administered instrument designed for use with

children and adolescents from age 5 years, 0 months, through 17 years, 11 months. It is intended to provide a fair assessment of intelligence for those who have speech, language, or hearing impairments; have different cultural or language backgrounds; or are unable to communicate verbally, in addition to individuals with mental retardation, autism, giftedness, and learning disabilities.

The Universal Nonverbal Intelligence Test measures intelligence through six culture-reduced subtests that combine to form two Primary Scales (Reasoning and Memory), two Secondary Scales (Symbolic and Nonsymbolic), and a Full Scale. Each of the six subtests is administered using eight reasonably universal hand and body gestures, demonstrations, scored items that do not permit examiner feedback, sample items, corrective responses, and transitional checkpoint items to explain the tasks to the examinee. The entire process is nonverbal but does require motor skills for manipulatives, paper and pencil, and pointing.

Three administrations are available for use depending on the reason for referral. These are an Abbreviated Battery containing two subtests (10–15 minutes), the Standard Battery containing four subtests (30 minutes), and the Extended Battery containing six subtests (45 minutes).

Wechsler Abbreviated Scale of Intelligence

The Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999; see also Homack & Reynolds, 2007) is an individually administered test that can be administered in approximately 30 minutes to individuals between the ages of 6 and 89.

The Wechsler Abbreviated Scale of Intelligence consists of four subtests: Vocabulary, Similarities, Block Design, and Matrix Reasoning. All items are new and parallel to their full Wechsler counterparts. The four subtests yield a FSIQ and can also be divided into Verbal IQ and Performance IQ. The Verbal IQ is based on the Vocabulary and Similarities subtests of the Wechsler Abbreviated Scale of Intelligence. The Performance IQ is based on Matrix Reasoning, which measures nonverbal fluid ability, and Block Design, which measures visual–spatial thinking. An estimate of general intellectual ability can also be obtained from just a two-subtest administration that includes Vocabulary and Matrix Reasoning and provides only the FSIQ.

Wechsler Intelligence Scale for Children—Fourth Edition

The WISC–IV (Wechsler, 2003; see also Flanagan & Kaufman, 2009; Prifitera et al., 2005, 2008; Weiss et al., 2006) is an individually administered clinical instrument for assessing the cognitive ability of children of ages 6 years, 0 months, through 16 years, 11 months. It includes 10 core subtests with five supplemental subtests, which can be used for additional information or, under rigidly specified conditions, to substitute for core subtests. Administration takes approximately 65 to 80 minutes for most children. The test provides a composite score (FSIQ) that represents general intellectual ability as well as four factor index scores (Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed). Each of the IQs and factor indexes is reported as a standard score with a mean of 100 and a standard score of 15. The subtests on the WISC–IV provide scaled scores with a mean of 10 and a standard deviation of 3.

Three subtests compose the Verbal Comprehension Index: Similarities, Vocabulary, and Comprehension. In addition, two supplementary verbal subtests, Information and Word Reasoning, are also available, and one may be substituted for any one of the other Verbal Comprehension subtests if needed. These subtests assess verbal reasoning, comprehension, and conceptualization.

Three subtests compose the Perceptual Reasoning Index: Block Design, Picture Concepts, and Matrix Reasoning. Picture Completion is a supplementary subtest that can be used as a substitute if necessary. These subtests measure perceptual reasoning and organization.

The Working Memory Index has two subtests: Digit Span and Letter–Number Sequencing. One supplementary subtest, Arithmetic, can be used to replace either of the Working Memory subtests. These subtests measure attention, concentration, and working memory.

The Processing Speed Index also has two subtests: Coding and Symbol Search. Cancellation is a

supplementary subtest and can be used as a substitute for either subtest of the Processing Speed Index. These subtests measure the speed of mental and graphomotor processing.

The WISC–IV is part of the long tradition of Wechsler scales, which allows examiners to draw on a wealth of research and interpretive opinions, including their own experience with various Wechsler scales. The theoretical model underlying the current development of this version is complex and, although there is some overlap with CHC theory, which is discussed in the manual, CHC theory is not the basis for the WISC–IV.

Wechsler Intelligence Scale for Children—Fourth Edition Integrated

The WISC–IV Integrated (Wechsler et al., 2004) adds an array of supplemental subtests, additional procedures, and standardized observations to the WISC–IV. These subtests and procedures have been developed from Kaplan's (1988) process approach to assessment. Norms are based on a smaller sample than that of the WISC–IV, which necessitates some caution in use of the scores, but tables are provided for determining significance and base rates of differences within the WISC–IV Integrated and between WISC–IV and WISC–IV Integrated scores. Examiners may use as few or as many WISC–IV Integrated procedures as they wish. Some examples of additional measures on the WISC–IV Integrated include Elithorn Mazes, multiple-choice versions of the Verbal Comprehension and Block Design subtests, and visual analogs to Digit Span: Visual Digit Span and Spatial Span, which requires imitating or reversing the examiner's sequence of tapping cubes scattered on a board. Examiners who want to explore and analyze a child's performance on the WISC–IV in depth and to follow up on questions raised by the WISC–IV scores will find the WISC–IV Integrated subtests useful.

Wechsler Preschool and Primary Scale of Intelligence—Third Edition

The WPPSI–III (Wechsler, 2002b; see also Lichtenberger & Kaufman, 2003; Wahlstrom et al., 2012) is an individually administered instrument that assesses cognitive functioning and global intelligence for early childhood (ages 2 years, 6 months–7 years, 11 months). The instrument can provide information pertaining to a child's cognitive strengths and weaknesses related to language, visual–perceptual skills, visual–motor integration, and reasoning.

Because the WPPSI–III covers a broad age range in which rapid advances in development are typical for youngsters, it is divided into two separate batteries (the first for ages 2 years, 6 months–3 years, 11 months, and the second for ages 4 years–7 years, 3 months). The test consists of 14 subtests (not all used at any particular age) that combine into four or five composites: Verbal IQ, Performance IQ, Processing Speed Quotient (for upper ages only), General Language Composite (actually oral vocabulary), and FSIQ.

The FSIQ is a general measure of global intelligence reflecting performance across various subtests within the Verbal IQ and Performance IQ domains. In general, the Verbal IQ contains subtests that measure general fund of information, verbal comprehension, receptive and expressive language, attention span, and degree of abstract thinking. The Performance IQ consists of subtests that collectively assess visual–motor integration, perceptual–organizational skills, concept formation, speed of mental processing, nonverbal problem solving, and graphomotor ability.

Scores provided include scaled scores for subtests, standard scores for composite scores, percentiles, and qualitative descriptors. The composite standard scores have a mean of 100 and a standard deviation of 15. Scaled scores have a mean of 10 and a standard deviation of 3.

The WPPSI–III provides moderate continuity of format and organization with the WISC–IV and Wechsler Adult Intelligence Scales—IV. It seems to us to be much more child friendly and developmentally appropriate than previous editions, but perhaps less gamelike than some other tests for young children.

Wechsler Nonverbal Scale of Ability

The Wechsler Nonverbal Scale of Ability (Wechsler & Naglieri, 2006; see also Brunnert et al., 2008; Naglieri & Otero, 2012b) is a cognitive ability test with nonoral administration and materials for ages 4 years, 0 months, through 21 years, 11 months. It

has six subtests, with four (Matrices, Coding, Object Assembly, and Recognition) used for the Full Scale at ages 4 years, 0 months, through 7 years, 11 months, and four (Matrices, Coding, Spatial Span, and Picture Arrangement) used at ages 8 years, 0 months, through 21 years, 11 months. There are also norms for a two-subtest battery at each age range. Subtests use *T* scores ($M = 50$, $SD = 10$), and the IQs are standard scores ($M = 100$, $SD = 15$). Administration is normally accomplished with pictorial instructions and standardized gestures. Examiners are also permitted to use standardized verbal instructions provided in six languages, and they may use a qualified interpreter to translate the instructions into other languages in advance. Examiners may also provide additional help as dictated by their professional judgment.

The Wechsler Nonverbal Scale of Ability is an efficient nonverbal test with clear instructions. The four-subtest format, using mostly subtests similar to those on other Wechsler scales, is quicker than the Leiter–R but provides fewer subtests and abilities to analyze. The flexibility of standardized pictorial, pantomime, and verbal instructions, with a provision for additional help as dictated by the examiner's judgment, makes the Wechsler Nonverbal Scale of Ability especially useful.

Woodcock–Johnson III Tests of Cognitive Abilities and Diagnostic Supplement

The WJ III Tests of Cognitive Abilities (Woodcock et al., 2001b, 2007; see also Schrank & Flanagan, 2003; Schrank, Miller, Wendling, & Woodcock, 2010; Schrank & Wendling, 2012) is an individually administered test of abilities appropriate for ages 2 to 90. Unlike many individual ability tests, the WJ III Tests of Cognitive Abilities are explicitly designed to assess a person's abilities on many specific CHC Fluid Reasoning–Crystallized Ability "cognitive factors," not just a total score or a few factors. The General Intellectual Ability (GIA) score of the WJ III is based on a weighted combination of tests that best represent a common ability underlying all intellectual performance. Examiners can obtain a GIA (standard) score by administering the first 7 tests in the Tests of Cognitive Abilities or a GIA (Extended) score by administering 14 cognitive tests. Additional tests from both the Tests of Cognitive Ability and the Diagnostic Supplement (Woodcock, McGrew, & Mather, 2001c) can be used to further explore CHC abilities and other factors. Each of the cognitive tests represents a different broad CHC factor. A three-test Brief Intellectual Ability score is available and takes about 10 to 15 minutes to administer and is useful for screenings and reevaluations. Examiners are permitted to select the tests they need to assess abilities in which they are interested for a particular student. The WJ III Tests of Cognitive Abilities provide interpretive information from 20 tests to measure cognitive performance. The Diagnostic Supplement adds 11 more tests, and some of the 21 tests from the WJ III Tests of Achievement (Woodcock et al., 2001a) also tap cognitive as well as achievement abilities. Several of the tests appropriate for younger children are in the Diagnostic Supplement.

The WJ III provides raw scores that are converted, using age- or grade-based norms, to standard scores, percentile ranks, age and grade equivalents, relative proficiency index scores and W scores (Jaffe, 2009), instructional ranges, and cognitive–academic language proficiency levels. All score transformation is performed through the use of the computer program (WJ III Compuscore). The program also can generate several "discrepancy" analyses: intra-ability discrepancies (intracognitive, intra-achievement, and intraindividual) and ability achievement discrepancies (predicted achievement vs. achievement, GIA vs. achievement, and oral language ability vs. achievement).

Examiners who wish to assess the full range of CHC abilities will find the WJ III precisely designed for this purpose. Examiners wishing to conduct a CHC cross-battery assessment (Flanagan et al., 2007) can use the WJ III tests to supplement their chosen core measure of intellectual functioning. The test manuals (Woodcock et al., 2001a, 2001b, 2001c) explicitly permit examiners to select as few or as many tests as they need for an assessment.

References

Airasian, P. (2002). *Assessment in the classroom.* New York, NY: McGraw-Hill.

Alfonso, V. C., Johnson, A., Patinella, L., & Rader, D. E. (1998). Common WISC–III examiner errors: Evidence from graduate students in training. *Psychology in the Schools, 35*, 119–125. doi:10.1002/(SICI)1520-6807(199804)35:2<119::AID-PITS3>3.0.CO;2-K

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall International.

Bagnato, S. J. (2007). *Authentic assessment for early childhood intervention: Best practices.* New York, NY: Guilford Press.

Bagnato, S. J., & Neisworth, J. T. (1994). A national study of the social and treatment "invalidity" of intelligence testing for early intervention. *School Psychology Quarterly, 9,* 81–102. doi:10.1037/h0088852

Batshaw, M. L., Pellegrino, L., & Roizen, N. J. (2007). *Children with disabilities* (6th ed.). Baltimore, MD: Paul H. Brookes.

Bergeron, R., Floyd, R. G., & Shands, E. I. (2008). State eligibility guidelines for mental retardation: An update and consideration of part scores and unreliability of IQs. *Education and Training in Developmental Disabilities, 43*, 123–131.

Binet, A., & Simon, T. (1980). *The development of intelligence in children: With marginal notes by Lewis M. Terman and preface by Lloyd M. Dunn* (E. S. Kite, Trans.). Nashville, TN: Williams Printing. (Original work published 1916)

Bracken, B. A. (1994). Advocating for effective preschool assessment practices: A comment on Bagnato & Nisworth. *School Psychology Quarterly, 9*, 103–108. doi:10.1037/h0088845

Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test*. Austin, TX: Pro-Ed.

Bracken, B. A., & McCallum, R. S. (2001). Assessing intelligence in a nation that speaks more than 200 languages. In L. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (2nd ed., pp. 405–431). San Francisco, CA: Jossey-Bass.

Bracken, B. A., & Nagle, R. J. (2007). *Psychoeducational assessment of preschool children.* Mahwah, NJ: Erlbaum.

Bracken, B. A., & Naglieri, J. A. (2003). Assessing diverse populations with nonverbal tests of general intelligence. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (2nd ed., pp. 243–274). New York, NY: Guilford Press.

Bracken, B. A., & Walker, K. C. (1997). The utility of intelligence tests for preschool children. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 484–502). New York, NY: Guilford Press.

Braden, J. P., & Athanasiou, M. S. (2005). A comparative review of nonverbal measures of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 557–577). New York, NY: Guilford Press.

Bradley-Johnson, S. (2001). Cognitive assessment for the youngest children: A critical review of tests. *Journal of Psychoeducational Assessment, 19*, 19–44. doi:10.1177/073428290101900102

Brahan, D., & Bauchner, H. (2005). Changes in reporting of race/ethnicity, socioeconomic status. *Pediatrics, 115*, e163–e166. Retrieved from http://pediatrics.aappublications.org/cgi/content/full/115/2/e163

Brassard, M. R., & Boehm, A. E. (2007). *Preschool assessment.* New York, NY: Guilford Press.

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (2010). *Test of Nonverbal Intelligence* (4th ed.). Austin, TX: Pro-Ed.

Brunnert, K. A., Naglieri, J. A., & Hardy-Braz, S. T. (2008). *Essentials of WNV assessment.* New York, NY: Wiley.

Buros, O. K. (Ed.). (1938). *The 1938 mental measurements yearbook*. Lincoln: University of Nebraska Press.

Canivez, G. L., & Watkins, M. W. (1998). Long term stability of the Wechsler Intelligence Scale for Children—Third Edition. *Psychological Assessment, 10*, 285–291. doi:10.1037/1040-3590.10.3.285

Canivez, G. L., & Watkins, M. W. (1999). Long term stability of the Wechsler Intelligence Scale for Children—Third Edition among demographic subgroups: Gender, race, and age. *Journal of Psychoeducational Assessment, 17*, 300–313. doi:10.1177/073428299901700401

Canivez, G. L., & Watkins, M. W. (2001). Long term stability of the Wechsler Intelligence Scale for Children—Third Edition among students with disabilities. *School Psychology Review, 30,* 438–453.

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 41–68). New York, NY: Guilford Press. (Original work published 1997)

Cattell, R. B., & Horn, J. L. (1978). A check on the theory of fluid and crystallized intelligence

with description of new subtest designs. *Journal of Educational Measurement, 15*, 139–164. doi:10.1111/j.1745-3984.1978.tb00065.x

Cole, M., & Cole, S. R. (1996). *The development of children* (3rd ed.). New York, NY: Freeman.

Cronshaw, S. F., Hamilton, L. K., Onyura, B. R., & Winston, A. S. (2006). Case for non-biased intelligence testing against Black Africans has not been made: A comment on Rushton, Skuy, and Bons (2004). *International Journal of Selection and Assessment, 14*, 278–287. doi:10.1111/j.1468-2389.2006.00346.x

Cummins, J. (1979). *Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters* (Working Papers on Bilingualism No. 19). Toronto, Ontario, Canada: Bilingual Education Project.

Cummins, J. (1986). Psychological assessment of minority students: Out of context, out of focus, out of control? *Journal of Reading, Writing, & Learning Disabilities International, 2*, 9–19.

Daniel, M. H. (1997). Intelligence testing: Status and trends. *American Psychologist, 52,* 1038–1045. doi:10.1037/0003-066X.52.10.1038

Das, J. P., Kirby, J., & Jarman, R. F. (1975). Simultaneous and successive synthesis: An alternative model for cognitive abilities. *Psychological Bulletin, 82,* 87–103. doi:10.1037/h0076163

Drozdick, L. W., Wahlstrom, D., Zhu, J., & Weiss, L. G. (2012). The Wechsler Adult Intelligence Scale—Fourth Edition and the Wechsler Memory Scale—Fourth Edition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 197–223). New York, NY: Guilford Press.

Dumont, R., & Willis, J. O. (n.d.). *The evaluation of Sam McGee.* Retrieved from http://alpha.fdu.edu/psychology/McGee.htm

Dumont, R., Willis, J. O., & Elliott, C. D. (2008). *Essentials of DAS–II assessment.* New York, NY: Wiley.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). Minneapolis, MN: NCS Pearson.

Education for All Handicapped Children Act of 1975, Public Law No. 94-142, 20 U.S.C. § 1400 *et seq.*

Education for All Handicapped Children Amendments of 1986, Pub. L. 99–457, 20 U.S.C. § 1401, Part H, Section 677.

Elliott, C. D. (2005). The Differential Ability Scales. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 402–424). New York, NY: Guilford Press.

Elliott, C. D. (2007a). *Differential Ability Scales* (2nd ed.). San Antonio, TX: Psychological Corporation.

Elliott, C. D. (2007b). *Differential Ability Scales 2nd edition introductory and technical handbook.* San Antonio, TX: Psychological Corporation.

Elliott, C. D. (2012). The Differential Ability Scales—Second Edition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 336–356). New York, NY: Guilford Press.

Epstein, A. (2004). Preschool assessment: What's right for four-year-olds? *High/Scope Resource, 23*(1).

Espinosa, L. M. (2005). Curriculum and assessment considerations for young children from culturally, linguistically, and economically diverse backgrounds. *Psychology in the Schools, 42,* 837–853. doi:10.1002/pits.20115

Fiorello, C. A., Hale, J. B., McGrath, M., Ryan, K., & Quinn, S. (2001). IQ interpretation for children with flat and variable test profiles. *Learning and Individual Differences, 13*, 115–125. doi:10.1016/S1041-6080(02)00075-4

Flanagan, D. P., & Alfonso, V. C. (1995). A critical review of the technical characteristics of new and recently revised intelligence tests for preschool children. *Journal of Psychoeducational Assessment, 13*, 66–90. doi:10.1177/073428299501300105

Flanagan, D. P., & Kaufman, A. S. (2009). *Essentials of WISC–IV assessment* (2nd ed.). New York, NY: Wiley.

Flanagan, D. P., & McGrew, K. S. (1997). A cross-battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and cognitive science. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 314–325). New York, NY: Guilford Press.

Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). *The Wechsler Intelligence Scales and Gf–Gc theory: A contemporary approach to interpretation.* Boston, MA: Allyn & Bacon.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). *Essentials of cross-battery assessment* (2nd ed.). New York, NY: Wiley.

Ford, L., & Dahinten, V. S. (2005). Use of intelligence tests in the assessment of preschoolers. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 487–503). New York, NY: Guilford Press.

Garcia Coll, C. T. (1990). Developmental outcome of minority infants: A process-oriented look into our beginnings. *Child Development, 61,* 270–289. doi:10.2307/1131094

Geisinger, K. F. (1994a). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment

instruments. *Psychological Assessment, 6,* 304–312. doi:10.1037/1040-3590.6.4.304

Geisinger, K. F. (1994b). Psychometric issues in testing students with disabilities. *Applied Measurement in Education, 7,* 121–140. doi:10.1207/s15324818ame0702_2

Geisinger, K. F., Spies, R. A., Carlson, J. F., & Plake, B. S. (Eds.). (2007). *The seventeenth mental measurements yearbook*. Lincoln: University of Nebraska Press.

Gesell, A. (1925). *The mental growth of the pre-school child.* New York, NY: Macmillan.

Glascoe, F. P. (2005). Screening for developmental and behavioral problems. *Mental Retardation and Developmental Disabilities Research Reviews, 11,* 173–179. doi:10.1002/mrdd.20068

Glascoe, F. P., Martin, E. D., & Humphrey, S. A. (1990). Comparative review of developmental screening tests. *Pediatrics, 86*, 547–554.

Goldman, J. J. (1989). On the robustness of psychological test instrumentation: Psychological evaluation of the dead. In G. G. Ellenbogen (Ed.), *The primal whimper: More readings from the Journal of Polymorphous Perversity* (pp. 57–68). New York, NY: Ballantine.

Gottfredson, L. S. (2008). Of what value is intelligence? In A. Prifitera, D. Saklofske, & L. G. Weiss (Eds.), *WISC–IV applications for clinical assessment and intervention* (2nd ed., pp. 545–563). Burlington, MA: Elsevier.

Gutkin, T. B., & Reynolds, C. R. (Eds.). (2009). *The handbook of school psychology* (4th ed.). New York, NY: Wiley.

Haeussermann, E. (1958). *Developmental potential of preschool children.* New York, NY: Grune & Stratton.

Hale, J. B., Alfonso, V., Berninger, V., Bracken, B., Christo, C., Clark, E., Goldstein, S. (2010). Critical issues in response-to-intervention, comprehensive evaluation, and specific learning disabilities identification and intervention: An expert white paper consensus. *Learning Disability Quarterly, 33*, 223–236.

Hale, J. B., & Fiorello, C. A. (2001). Beyond the academic rhetoric of "g": Intelligence testing guidelines for practitioners. *School Psychologist, 55*, 113–117, 131–135, 138–139.

Hale, J. B., & Fiorello, C. A. (2004). *School neuropsychology: A practitioner's handbook.* New York, NY: Guilford Press.

Hale, J. B., Fiorello, C. A., Kavanagh, J. A., Hoeppner, J. B., & Gaitherer, R. A. (2001). WISC–III predictors of academic achievement for children with learning disabilities: Are global and factor scores comparable? *School Psychology Quarterly, 16*, 31–55. doi:10.1521/scpq.16.1.31.19158

Hammill, D. D., Pearson, N., & Wiederholt, J. L. (2009). *Comprehensive Test of Nonverbal Intelligence* (2nd ed.). Austin, TX: Pro-Ed.

Harry, B. (1992). An ethnographic study of cross-cultural communication with Puerto Rican-American families in the special education system. *American Educational Research Journal, 29*, 471–494.

Homack, S. R., & Reynolds, C. R. (2007). *Essentials of assessment with brief intelligence tests.* New York, NY: Wiley.

Horn, J. L., & Blankson, N. (2005). Foundation for better understanding of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 41–68). New York, NY: Guilford Press.

Individuals With Disabilities Education Act of 1990, Pub. L. 101–476, 20 U.S.C., Ch. 33. Retrieved from http://law.justia.com/us/codes/title20/20usc1400.html

Individuals With Disabilities Education Act Amendments of 1997, Pub. L. No. 105–117, 20 U.S.C. § 1400 *et seq.* Retrieved from http://www.ed.gov/policy/speced/guid/idea/omip.html

Individuals With Disabilities Education Improvement Act of 2004, Pub. L. No. 118–2647. Retrieved from http://idea.ed.gov/download/statute.html

Jaffe, L. E. (2009). *Development, interpretation, and application of the W score and the relative proficiency index* (Woodcock–Johnson III Assessment Service Bulletin No. 11). Rolling Meadows, IL: Riverside. Retrieved from http://www.riverpub.com/products/wjIIIComplete/pdf/WJ3_ASB_11.pdf

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Jensen, A. R. (2002). Psychometric g: Definition and substantiation. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (pp. 39–53). Mahwah, NJ: Erlbaum.

Kagitcibasi, C. (1996). *Family and development across cultures: A view from the other side.* Mahwah, NJ: Erlbaum.

Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. Boll & B. K. Bryant (Eds.), *Clinical neuropsychology and brain function: Research, measurement, and practice* (pp. 127–167). Washington, DC: American Psychological Association. doi:10.1037/10063-004

Kaufman, A. S. (1979). *Intelligent testing with the WISC–R.* New York, NY: Wiley Interscience.

Kaufman, A. S. (2009). *IQ testing 101.* New York, NY: Springer.

Kaufman, A. S., & Kaufman, N. L. (1977). *Clinical evaluation of young children with the McCarthy Scales.* New York, NY: Grune & Stratton.

Kaufman, A. S., & Kaufman, N. L. (2004a). *Kaufman Assessment Battery for Children* (2nd ed.). Circle Pines, MN: AGS.

Kaufman, A. S., & Kaufman, N. L. (2004b). *Kaufman Brief Intelligence Test* (2nd ed.). Circle Pines, MN: AGS.

Kaufman, A. S., Lichtenberger, E. O., Fletcher-Janzen, E., & Kaufman, N. L. (2005). *Essentials of KABC-II assessment*. New York, NY: Wiley.

Klassen, R. M., & Kishor, N. (1996). A comparative analysis of practitioners' errors on WISC–R and WISC–III. *Canadian Journal of School Psychology, 12*, 35–43. doi:10.1177/082957359601200106

Lichtenberger, E. O. (2005). General measures of cognition for the preschool child. *Mental Retardation and Developmental Disabilities, 11*, 197–208.

Lichtenberger, E. O., & Kaufman, A. S. (2003). *Essentials of WPPSI–III assessment*. New York, NY: Wiley.

Lichtenberger, E. O., & Kaufman, A. S. (2009). *Essentials of WAIS–IV assessment*. New York, NY: Wiley.

Lichtenberger, E. O., Mather, N., Kaufman, N. L., & Kaufman, A. S. (2004). *Essentials of assessment report writing*. New York, NY: Wiley.

Lichter, D. T., Quian, Z., & Crowley, M. L. (2006). Race and poverty: Divergent fortunes of America's children? *Focus, 24*(3), 8–16. Retrieved from http://www.irp.wisc.edu/publications/focus/pdfs/foc243b.pdf

Luria, A. R. (1966). *Human brain and psychological processes*. New York, NY: Harper & Row.

Luria, A. R. (1973). *The working brain: An introduction to neuropsychology*. New York, NY: Basic Books.

Luria, A. R. (1980). *Higher cortical functions in man* (2nd ed.). New York, NY: Basic Books. doi:10.1007/978-1-4615-8579-4

McCallum, R. S. (Ed.). (2003). *Handbook of nonverbal assessment*. New York, NY: Kluwer Academic/Plenum Press. doi:10.1007/978-1-4615-0153-4

McCallum, R. S., & Bracken, B. A. (2005). The Universal Nonverbal Intelligence Test. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 425–440). New York, NY: Guilford Press.

McCallum, R. S., & Bracken, B. A. (2012). The Universal Nonverbal Intelligence Test. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 357–375). New York, NY: Guilford Press.

McCallum, R. S., Bracken, B. A., & Wasserman, J. (2001). *Essentials of nonverbal assessment*. New York, NY: Wiley.

McCarthy, D. (1972). *The McCarthy Scales of Children's Abilities*. San Antonio, TX: Psychological Corporation.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8*, 290–302. doi:10.1177/073428299000800307

McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *Journal of Special Education, 25*, 504–526. doi:10.1177/002246699202500407

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc Cross-Battery Assessment*. Boston, MA: Allyn & Bacon.

McLean, M. (1998). Assessing young children for whom English is a second language. *Young Exceptional Children, 1*(3), 20–25. doi:10.1177/109625069800100304

McLean, M., Bailey, D. B., & Wolery, M. (2004). *Assessing infants and preschoolers with special needs* (3rd ed.). Upper Saddle River, NJ: Pearson.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128–165. doi:10.1037/0003-066X.56.2.128

Naglieri, J. A. (1999). *Essentials of CAS assessment*. New York, NY: Wiley.

Naglieri, J. A., & Das, J. P. (1997a). *Cognitive Assessment System*. Itasca, IL: Riverside.

Naglieri, J. A., & Das, J. P. (1997b). *Cognitive Assessment System interpretive handbook*. Itasca, IL: Riverside.

Naglieri, J. A., & Das, J. P. (2005). Planning, attention, simultaneous, successive (PASS) theory. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 120–135). New York, NY: Guilford Press.

Naglieri, J. A.. Das, J. P., & Goldstein, S. (2012). Planning, attention, simultaneous, successive: A cognitive-processing-based theory of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 178–194). New York, NY: Guilford Press.

Naglieri, J. A., & Otero, T. M. (2012a). The Cognitive Assessment System: From theory to practice. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 376–399). New York, NY: Guilford Press.

Naglieri, J. A., & Otero, T. M. (2012b). The Wechsler Nonverbal Scale of Ability: Assessment of diverse populations. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 436–455). New York, NY: Guilford Press.

Ortiz, S. O., & Dynda, A. M. (2005). Use of intelligence tests with culturally and linguistically diverse

populations. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 545–556). New York, NY: Guilford Press.

Ortiz, S. O., & Flanagan, D. P. (2002a). Cross-battery assessment revisited: Some cautions concerning "Some Cautions" (Part I). *Communiqué, 30*(7), 32–34.

Ortiz, S. O., & Flanagan, D. P. (2002b). Cross-battery assessment revisited: Some cautions concerning "Some Cautions" (Part II). *Communiqué*, *30*(8), 36–38.

Ortiz, S. O., Ochoa, S. H., & Dynda, A. M. (2012). Testing with culturally and linguistically diverse populations: Moving beyond the verbal-performance dichotomy into evidence-based practice. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 526–552). New York, NY: Guilford Press.

Pavri, S. (2001). Developmental delay or cultural difference? Developing effective child find practices in identifying children from culturally and linguistically diverse backgrounds. *Young Exceptional Children, 4*(4), 2–9. doi:10.1177/109625060100400401

Petermann, F., & Macha, T. (2008). Developmental assessment: A general framework. *Zeitschrift für Psychologie/Journal of Psychology, 216*, 127–134.

Prifitera, A., Saklofske, D. H., & Weiss, L. G. (Eds.). (2005). *WISC–IV: Clinical use and interpretation: Scientist–practitioner perspectives.* Burlington, MA: Elsevier.

Prifitera, A., Saklofske, & Weiss, L. G. (Eds.). (2008). *WISC–IV applications for clinical assessment and intervention* (2nd ed.). Burlington, MA: Elsevier.

Rapaport, D., Gill, M., & Schafer, R. (1945). *Diagnostic psychological testing* (Vol. I). Chicago, IL: Year Book. doi:10.1037/10834-000

Rehabilitation Act of 1973, Pub. L. 93-112, 29 U.S.C. 701 *et seq.*

Reschly, D. J., & Hosp, J. L. (2004). State SLD identification policies and practices. *Learning Disability Quarterly, 27*, 197–213. doi:10.2307/1593673 Retrieved from http://idea.ed.gov/download/statute.html

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales.* Lutz, FL: Psychological Assessment Resources.

Reynolds, C. R., & Kamphaus, R. W. (2005). Introduction to the Reynolds Intellectual Assessment Scales and Reynolds Intellectual Screening Test. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 461–483). New York, NY: Guilford Press.

Reynolds, C. R., Kamphaus, R. W, & Raines, T. C. (2012). The Reynolds Intellectual Assessment Scales and the Reynolds Intellectual Screening Test. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 400–421). New York, NY: Guilford Press.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales* (5th ed.). Itasca, IL: Riverside Publishing.

Roid, G. H., & Barram, R. A. (2004). *Essentials of Stanford–Binet Intelligence Scales (SB5) assessment.* New York, NY: Wiley.

Roid, G. H., & Miller, L. J. (1997). *Leiter International Performance Scale—Revised.* Wood Dale, IL: Stoelting.

Roid, G. H., & Pomplum, M. (2005). Interpreting the Stanford-Binet Intelligence Scales. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 325–343). New York, NY: Guilford Press.

Roid, G. H., & Pomplum, M. (2012). The Stanford-Binet Intelligence Scales, Fifth Edition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 249–268). New York, NY: Guilford Press.

Rydz, D., Shevell, M. I., Majnemer, A., & Oskoui, M. (2005). Developmental screening. *Journal of Child Neurology, 20,* 4–21. doi:10.1177/08830738050200010201

Sandoval, J., Frisby, C. L., Geisinger, K. F., Scheuneman, J. D., & Grenier, J. R. (Eds.). (1998). *Test interpretation and diversity: Achieving equity in assessment.* Washington, DC: American Psychological Association. doi:10.1037/10279-000

Sattler, J. M. (1974). *Assessment of children's intelligence.* Philadelphia, PA: Saunders.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). San Diego, CA: Author.

Sattler, J. M., & Hoge, R. (2006). *Assessment of children: Behavioral, social, and clinical foundations* (5th ed.). San Diego, CA: Jerome M. Sattler.

Sattler, J. M., & Ryan, J. J. (2009). *Assessment with the WAIS–IV.* San Diego, CA: Jerome M. Sattler.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99–144). New York, NY: Guilford Press.

Schrank, F. A., & Flanagan, D. P. (2003). *WJ III clinical use and interpretation: Scientist-practitioner perspectives.* Burlington, MA: Elsevier.

Schrank, F. A., Miller, D. C., Wendling, B. J., & Woodcock, R. W. (2010). *Essentials of WJ III cognitive abilities assessment* (2nd ed.). New York, NY: Wiley.

Schrank, F. A., & Wendling B. J. (2012). The Woodcock–Johnson III normative update. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 297–335). New York, NY: Guilford Press.

Singer, J. K., Lichtenberger, E. O., Kaufman, J. C., Kaufman, A. S., & Kaufman, N. (2012). The Kaufman Assessment Battery for Children—Second Edition and the Kaufman Test of Educational Achievement—Second Edition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 269–296). New York, NY: Guilford Press.

Snyder, P., & Lawson, S. (1993). Evaluating the psychometric integrity of instruments used in early intervention research: The Battelle Developmental Inventory. *Topics in Early Childhood Special Education, 13,* 216–232. doi:10.1177/027112149301300209

Spearman, C. E. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201–293. doi:10.2307/1412107

Terman, L. M. (1916). *The measurement of intelligence.* Boston, MA: Houghton-Mifflin. doi:10.1037/10014-000

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology, 40,* 7–26. doi:10.1016/S0022-4405(01)00092-9

U.S. Census Bureau. (2006). *Annual estimates of the population by sex, race and Hispanic or Latino origin for the United States: April 1, 2000 to July 1, 2005* (NC-EST2005-03). Retrieved from http://www.census.gov/popest/data/historical/2000s/vintage_2005/index.html

U.S. Department of Education. (2001). *Twenty-fourth annual report to Congress on the implementation of IDEA.* Retrieved from http://www2.ed.gov/about/reports/annual/osep/2002/index.html

U.S. Department of Education. (2005). *Twenty-sixth annual report to Congress on the implementation of IDEA.* Retrieved from http://www2.ed.gov/about/reports/annual/osep/2004/index.html

Valdivia, R. (1999). *The implications of culture on developmental delay* (ED438663). Retrieved from http://www.ericdigests.org/2000-4/delay.htm

Wahlstrom, D., Breaux, K. C., Zhu, J., & Weiss, L. G. (2012). The Wechsler Preschool and Primary Scale of Intelligence—Third Edition, the Wechsler Intelligence Scale for Children—Fourth Edition, and the Wechsler Individual Achievement Test—Third Edition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 224–248). New York, NY: Guilford Press.

Watkins, M. W. (2003). IQ subtest analysis: Clinical acumen or clinical illusion? *Scientific Review of Mental Health Practice, 2*, 118–141.

Watkins, M. W. (2009). Errors in diagnostic decision making and clinical judgment. In T. B. Gutkin & C. R. Reynolds (Eds.), *Handbook of school psychology* (4th ed., pp. 210–229). New York, NY: Wiley.

Watkins, M. W., & Glutting, J. J. (2000). Incremental validity of WISC–III profile elevation, scatter, and shape information for predicting reading and math achievement. *Psychological Assessment, 12*, 402–408. doi:10.1037/1040-3590.12.4.402

Watkins, M. W., Glutting, J. J., & Lei, P.-W. (2007). Validity of the full scale IQ when there is significant variability among WISC–III and WISC–IV factor scores. *Applied Neuropsychology, 14*, 13–20. doi:10.1080/09084280701280353

Watkins, M. W., Glutting, J., & Youngstrom, E. (2002). Cross-battery cognitive assessment: Still concerned. *Communiqué, 31*(2), 42–44.

Watkins, M. W., Glutting, J., & Youngstrom, E. (2005). Issues in subtest profile analysis. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 251–268). New York, NY: Guilford Press.

Watkins, M. W., & Kush, J. C. (1994). WISC–R subtest analysis: The right way, the wrong way, or no way? *School Psychology Review, 23*, 640–651.

Watkins, M. W., Youngstrom, E. A., & Glutting, J. J. (2002). Some cautions regarding cross-battery assessment. *Communiqué, 30*(5), 16–20.

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Baltimore, MD: Williams & Wilkins. doi:10.1037/11329-000

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children.* New York, NY: Psychological Corporation.

Wechsler, D. (2002a). *Wechsler Abbreviated Scale of Intelligence.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2002b). *Wechsler Preschool and Primary Scale of Intelligence* (3rd ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale* (4th ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D., & Naglieri, J. A. (2006). *Wechsler Nonverbal Scale of Ability.* San Antonio, TX: Psychological Corporation.

Weiss, L. G., Saklofske, D. H., Prifitera, A., & Holdnack, J. A. (2006). *WISC–IV: Advanced clinical interpretation.* Burlington, MA: Elsevier.

Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment, 8,* 231–258. doi:10.1177/073428299000800303

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001a). *Woodcock–Johnson III Tests of Achievement*. Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001b). *Woodcock–Johnson III Tests of Cognitive Ability*. Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001c). *Woodcock–Johnson III diagnostic supplement*. Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., Schrank, F. A., & Mather, N. (2007). *Woodcock–Johnson III normative update*. Rolling Meadows, IL: Riverside. (Original work published 2001)

Zhu, J., & Weiss, L. (2005). The Wechsler scales. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 297–324). New York, NY: Guilford Press.

CHAPTER 4

ASSESSING INTELLIGENCE NONVERBALLY

R. Steve McCallum

During the latter half of the 20th century, and particularly during the late 1960s and 1970s, objections were raised (Hoffman, 1962; Jackson, 1975; Williams, 1971) regarding the use of existing and well-established intelligence tests to assess minority children (e.g., the Wechsler Intelligence Test for Children; Wechsler, 1949). Most of the extant tests were highly language loaded. As such, they were characterized as incapable of assessing intellectual (sub)constructs and predicting performance comparably across diverse racial and ethnic groups, of having inappropriate standardization samples for minority examinees, of containing examiner and language bias against minority examinees, and of producing inequitable social consequences for certain groups. Many of these objections were stated as facts based on rational (and perhaps emotional) grounds rather than on empirical evidence.

In the intervening years, many of the objections to intelligence testing have been addressed empirically (see Reynolds & Lowe, 2009). In addition to the increased empirical activity in response to the charges leveled against intelligence tests, a number of additional specific and related outcomes have resulted from the scrutiny focused on ability tests. For example, more recent test consumers and experts have begun to exercise more precision in the language they used to describe test characteristics and results (e.g., see Jensen's 1980 discussion of the term *discrimination*), to consider important distinctions between terms such as *cultural bias* versus *culture fair* versus *cultural loading* and to use definitions of terms that could actually be operationalized. Reynolds (1982) considered several conceptualizations of test bias and concluded that the most defensible definition of bias is one that can be couched in scientific terms: whether there is systematic error in the measurement of a psychological attribute as a function of membership in one or another subgroup (e.g., gender, cultural, racial). Perhaps the most salient consequence of the criticism of language-loaded intelligence tests has been the development of several psychometrically strong nonverbal intelligence–cognitive tests for use with an increasingly diverse U.S. population. The demonstrable need for sound nonverbal assessment instruments was underscored by Braden and Athanasiou (2005), who noted,

> Because verbal (or to be more precise, language-loaded) tests of intelligence presume that examinees have met a threshold of exposure to a standard form of the dominant language, the use of language-loaded tests has been suspect when examinees do not meet these assumptions (e.g., they speak a foreign language, have deafness, hearing impairments, or come from families using nonstandard forms of spoken language) (Lopez, 1997). Likewise, the presumption that individuals have the ability to understand language, and to use it to reason and respond, may be inappropriate for examinees who have experienced traumatic brain injury, stroke, or

DOI: 10.1037/14049-004
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

> degenerative neurological conditions. . . . Therefore, nonverbal (or, to be more precise, language reduced) measures . . . are essential tools. (p. 557)

Braden and Athanasiou (2005) concluded that nonverbal tests are essential complements to verbally loaded measures and that nonverbal tests of intelligence are superior to alternative methods of assessing intelligence in many situations. However, as is the case for all intelligence tests, careful and cautious use of nonverbal tests is warranted. The information contained in this chapter can be used to help examiners conduct careful and sensitive evaluations, consistent with Braden and Athanasiou's recommendations. In this chapter, the characteristics of the eight most prominent nonverbal tests are reviewed to help examiners make sound choices regarding selection of nonverbal tests on the basis of referral concerns. In addition, other more specific and related goals are addressed, including a brief history of nonverbal intelligence testing; a rationale for using these instruments; controversies surrounding the use of nonverbal tests; a detailed description of the administrative, statistical, and fairness characteristics of recently developed tests; the sociopolitical context for nonverbal cognitive assessment; criteria experts have used to characterize the linguistic and cultural demands of various nonverbal tests (e.g., Flanagan, Ortiz, & Alfonso, 2009); and some of the promising future-oriented research with nonverbal tests.

BRIEF HISTORY OF NONVERBAL INTELLIGENCE TESTING

Historically, tests of cognition or intelligence have relied on the exchange of linguistic communications from the examiner, examinee, or both; sometimes expressive and sometimes receptive, and often both. When it became obvious to practitioners that not all examinees had sufficient language facility to comprehend or respond to examiners' questions, because of physiological limitations (e.g., impaired hearing or language processing), psychological characteristics, (e.g., elective mutism), or lack of cultural or environmental exposure to the primary language, test authors began to develop creative nonverbal strategies to assess general cognitive abilities without the use of language.

This development of nonverbal measures began in earnest during the later part of the 19th century. Among the first documented efforts to use nonverbal strategies occurred when Jean Itard attempted to assess Victor, the "Wild Boy of Aveyron," a feral youth discovered wandering the countryside in France in the 1800s (Carrey, 1995). Early in the next century, Seguin (1907) began working on novel strategies to nonverbally assess the cognitive abilities of clients who were unable to effectively use their native language to respond to verbal questions. Seguin's novel "form board" test required placement of common geometric shapes into same-shape holes cut out of a wooden board. Although this test dates back more than a century, variations of this task are used currently for young examinees and those with limited use of the dominant language.

Nonverbal assessment became particularly relevant on a large scale in the United States during World War I, when the army developed the nonverbal Group Examination Beta version of the Army Mental Tests as a supplement to the verbally laden Army Alpha. Both forms of the test were used to classify potential soldiers on the basis of their mental ability, aiding in the identification of recruits with significant cognitive limitations as well as those with special talents who might be good candidates to become officers. Nonverbal testing continued to grow after the war throughout the private sector. In 1924, Arthur developed the first version of the Arthur Point Scale of Performance Tests, which combined several existing nonverbal tasks into a single battery; he revised this test in 1943. In his first battery, Arthur included such psychometric tasks as the Knox Cube Test (Knox, 1914), the Sequin Form Board, and the Porteus Maze Test (Porteus, 1915). In addition to their use in assessing recruits, some of these instruments had been used on Ellis Island to assess and classify immigrants (e.g., the Knox Cube Test).

The government-sponsored tradition of using nonverbal assessment techniques to assess cognitive functioning of individuals who could not be assessed optimally using language-loaded instruments continued in the private sector with the

development of such seminal measures as the Leiter International Performance Scale (Leiter, 1948), the Draw-a-Person test (Goodenough, 1926), and much later, the Columbia Mental Maturity Scale (Burgemeister, Blum, & Lorge, 1972). Although some of these tests continued to be used into the late 1970s, most had been criticized and largely abandoned by that time because of their poorly developed administration procedures, limited psychometric properties, or outdated norms. Consequently, many psychologists began to rely on the Performance scale of one or another of the Wechsler tests (e.g., Wechsler Intelligence Scale for Children; Wechsler, 1949) for an approximated nonverbal assessment because these scales required only minimal or no expressive language of the examinee. More important, the Wechsler Performance Scales did require receptive language from the examinee and both expressive and receptive language from the examiner.

Dissatisfaction with the Wechsler Performance subtests as nonverbal measures led to the development of several new-generation nonverbal tests during the 1990s, and today several psychometrically strong measures are available. Although many of the current measures assess a unidimensional construct, some of the current tests are multidimensional (i.e., characterized by diverse tasks and interactive materials). Current multidimensional tests include the Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998), the Leiter International Performance Scale—Revised (Leiter–R; Roid & Miller, 1997), and the Wechsler Nonverbal Intelligence Scale (WNV; Wechsler & Naglieri, 2006). In addition, several commonly used unidimensional tests are available, including the Test of Nonverbal Intelligence—4 (TONI–4; L. Brown, Sherbenou, & Johnsen, 2010), the Comprehensive Test of Nonverbal Intelligence—2 (CTONI–2; Hammill, Pearson, & Wiederholt, 2009), and the Naglieri Nonverbal Ability Test Individual Administration (NNAT–I; Naglieri, 2003).

Today, nonverbal tests vary considerably in their content, breadth, and administration methodology. Tests that are purported to be nonverbal tests vary on a variety of dimensions, including the extent of language required of the examinee or examiner, the number of different operationalizations (or tasks) used to define intelligence, the quality of their standardization sample, group versus individual administration, and so on. In the next section, some of the reasons for and benefits associated with use of these nonverbal tests are reviewed.

RATIONALE FOR USING NONVERBAL COGNITIVE TESTS

Traditionally, language-based interactions between an examiner and examinee have been used to gauge an examinee's intellect and personality by requiring examinees to respond verbally to examiners' questions (e.g., "How are words related?" "What do words mean?" "What is the appropriate thing to do in a particular situation?"). However, as detailed previously, the intellect of some examinees cannot be accessed optimally, if at all, via verbal interactions. In fact, for some individuals language is an obstacle rather than a conduit for effective communication, and nonverbal tests can avoid problematic communication within the assessment environment and thereby provide a fairer evaluation of a client's abilities.

Nonverbal assessment strategies and techniques have been developed, in part, to assess those individuals who are unable to use language because of physiological deficits (e.g., hearing loss) and, in part, because of the limitations associated with the need to assess an increasingly diverse U.S. population. This diversity is largely a result of immigration into the United States, and many of these immigrants and their children are not proficient in English. In fact, while this chapter was being written an article appeared in the popular press (El Nasser & Overberg, 2010) noting that the Diversity Index in this country is at an all time high. The Diversity Index is the probability that two people chosen at random from the U.S. population would be of a different race or ethnicity, on a scale ranging from 0 to 100. Currently, the Diversity Index stands at 52, as opposed to 1998, when it was 34. Much of the diversity results from an influx of young Hispanic immigrants whose birthrates are higher than those of non-Hispanic Whites (El Nasser & Overberg, 2010). Many Hispanic youths do not have English as a first language and English is not spoken in the home;

consequently, they are required to become facile in two languages, and may consequently be limited in both languages. Other authors have reported significant cultural diversity within the U.S. schools. For example, McCallum and Bracken (2001) reported the number of foreign languages used by students in various school systems in the United States. The Chicago system reported use of more than 200 languages by its student body; more than 67 languages were used in Tempe, Arizona, and more than 45 languages were spoken in Cobb County, Georgia, public schools.

Similarly, a large number of students are deaf or hard of hearing and have other language-related limitations. According to figures from the National Institutes of Health (2010a), 28,600,000 Americans are deaf or have other significant hearing impairments. In 2003, around 7,500,000 Americans were reported to have speech impairments that limited their ability to communicate effectively (National Institutes of Health, 2010b). Other students have neurological and psychiatric conditions that inhibit effective verbal communications (e.g., autism spectrum disorders, traumatic brain injury, and selective mutism).

Because at-risk populations have increased in size proportionately with an increasing general population, professionals who are responsible for creating standards governing the testing industry have begun to create documents with guidelines or standards for responsible and ethical service delivery to diverse populations, such as *Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations* (American Psychological Association, 1991) and *Standards for Educational and Psychological Testing,* developed by a joint commission of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education in 1999. In addition, recent legislative mandates included in the Individuals With Disabilities Education Improvement Act of 2004 and related guidelines have described recommendations for delivering fair and equitable assessment of at-risk diverse populations. When children within these populations experience academic difficulty, parents and teachers often want to know the origin of the difficulty and are motivated to rule in or rule out within-child limitations, such as intellectual disability or a learning disability. Nonverbal assessment can be helpful when ruling in or ruling out these other possible diagnoses.

CONTROVERSIES AND PROBLEMS REGARDING THE USE OF NONVERBAL TESTS

As mentioned previously, a number of controversies are related to intellectual assessment, especially when using language-loaded intelligence tests, but also when using nonverbal tests. At a very basic level, experts have agreed that intelligence itself is a hypothetical construct, and as with other constructs it cannot be observed directly. Consequently, intelligence must be inferred from behaviors, verbal or nonverbal, that are then operationalized to reflect levels of intelligence. This indirect assessment strategy is limited, in part because experts do not agree on which behaviors should be used as indicators that typify the construct or even on how those behaviors should be measured (Reynolds & Lowe, 2009). In some instances, answers to these questions will depend on the specific assessment context and the characteristics of the examinee. For example, some behavioral indicators work well for some examinees but are not appropriate for particular examinees. Examinees vary considerably in sophistication and facility with the primary language and core culture assumed within the assessment context as well as developmental level and ability to satisfy the task demands of the assessment.

Practitioners who evaluate examinees with language-related limitations face a problem unique to the assessment of this target population. Specifically, the term *nonverbal* is not (and has not) been used consistently within the profession. On one hand, Braden and Athanasiou (2005) suggested using the term to refer to tests that allow language-reduced test administration. On the other hand, some have used the term to characterize a more restricted use of language during administration of nonverbal tests. For example, within the context of administration of one nonverbal test, the UNIT, Bracken and McCallum (1998) characterized nonverbal assessment as that which requires no verbal

exchange between the examiner and the examinee to satisfy the standardized administration of the test, although examiners are encouraged to use a common language to establish rapport and exchange other relevant information (e.g., developmental, medical, educational history). Others have used the term *nonverbal* in a very liberal fashion. For example, as mentioned earlier, during the 1960s and into the 1970s, examiners for years used the Performance subtests from the various Wechsler scales to obtain nonverbal operationalizations of intelligence because these subtests required minimal spoken language from examinees, even though the administration directions required lengthy verbal directions. Because test authors, publishers, and practitioners have not, and probably will not, reach a consensus any time soon on what constitutes a nonverbal assessment, perhaps the best solution is to carefully define the meaning of *nonverbal* within context. That is, examiners who use the term *nonverbal* must be responsible for carefully defining to clients and to consumers of the test results how the term is used in a particular assessment context.

Just as there is controversy about the most appropriate use of the term *nonverbal,* there is confusion about the how best to define the construct assessed by nonverbal tests. My colleagues and I have argued that true to Jensen's (1984) concept of "positive manifold," nonverbal tests of intelligence assess intelligence, period. We have contended that nonverbal tests do not assess nonverbal intelligence, but intelligence (McCallum, Bracken, & Wasserman, 2001). As Braden and Athanasiou (2005) noted, the argument against characterizing intelligence as either verbal or nonverbal rests on the assumption that underlying cognitive processes are consistently independent of verbal mediation. For example, a "verbal" task requiring examinees to discern who is biggest among three individuals when their relative sizes have been described verbally (e.g., Bob > Larry; Bob < Sam) may be solved by applying internal representation using visual–spatial–abstract processing. Although some examinees may mediate the task using words that define juxtapositions (e.g., *larger than, smaller than, middle sized*), the underlying internal processes required to solve the problem probably do not change as a function of whether the problem is presented verbally or visually. That is, an examinee will engage this problem by using either a visual–spatial format or a verbal mediation strategy as a function of preference for and history of success with one or the other.

On the basis of factor-analytic studies reported by Horn and Noll (1997) and Carroll (1993), Braden and Athanasiou (2005) concluded that the argument in support of independence of cognition and language has gained credibility. They reached their conclusions on the basis of group data from nonexceptional populations; however, the situation may be less clear cut when exceptional populations are studied (e.g., individuals with autism or neurological impairments). Some researchers, for example, have found evidence for the existence of nonverbal processing deficits or nonverbal learning disabilities (see Petti, Voelker, Shore, & Hayman-Abello, 2003; Rourke, 1991; Rourke & Conway, 1997) and have argued that nonverbal cognitive processes differ from verbal cognitive processes.

Interpretation of findings from various groups has suggested that the performance of some examinees will be limited as a function of construct-irrelevant variance when verbally laden tasks are used to assess intelligence (e.g., diminished language facility, not diminished intelligence). To confound the picture further, the origin of the linguistic limitation may be the result of biology (central nervous system capacity) or the environment (e.g., limited exposure). Moreover, in either case the limiting factor may be in the visual or verbal modality used to access the test demands, not the underlying problem-solving processes.

Most typically developing individuals, those equipped with well-functioning physiology and reared in the mainstream culture by parents who provided exposure to mainstream cultural and educational opportunities can be evaluated fairly with either verbal or nonverbal measures. For these examinees, comparison of performance on verbally versus nonverbally laden tasks will likely yield similar results and the verbal–nonverbal distinction is probably irrelevant. For others, however, those who may be biologically (e.g., hearing loss) or environmentally (e.g., culturally different) disadvantaged,

the type of instrument used will be critical. For these individuals, the test of choice will most likely be a nonverbal one to ensure (the) fair(est) assessment (Braden & Athanasiou, 2005).

Fair assessment is also related to the breadth of test coverage. That is, a test that provides assessment of more facets of intelligence is likely to be considered fair(er) and representative of one's overall intellectual ability than one that assesses fewer intellectual elements, assuming that the task demands do not contribute to test-irrelevant variance. Given that nonverbal tests limit administration strategies by avoiding significant verbal interactions between the examinee and the examiner, one might justifiably ask whether nonverbal tests adequately represent the broad construct of intelligence. This question is especially poignant because significant underrepresentation of the construct may lead to a lack of confidence in the test, because it (a) may lack face validity, that is, the test does not appear to measure the various facets of intelligence; (b) is less capable of operationalizing general intelligence, *g*; and (c) fails to reflect important language-loaded subconstructs associated with intelligence.

The response to the face validity criticism relies on the subjective judgment of the user and is not subject to data analyses. The response to the second criticism is found in the research literature (i.e., the extent to which nonverbal tests are found to load on the *g* factor.) A number of well-regarded theories or models of intelligence assume that intelligence is hierarchically arranged, with *g* at the apex and superordinate to a number of subconstructs (e.g., Carroll, 1993; McGrew & Flanagan, 1998). In fact, Wechsler's (1939) famous definition of intelligence uses the term *global ability to*, and many test authors have used a hierarchical model as a starting point (e.g., Differential Ability Scales—Second Edition [Roach & Elliott, 2006]; Woodcock–Johnson III [Woodcock, McGrew, & Mather, 2001]), including the authors of some nonverbal tests (e.g., Bracken & McCallum, 1998).

As previously noted, within hierarchical models, *g* is typically at the apex, with increasingly narrow subconstructs below, arranged in successive tiers of more specific abilities, with the narrowest constructs at the bottom, combining further to form the more superordinate constructs. Many experts consider *g* to be the single best, and most psychometrically sound, estimate of intelligence (e.g., Carroll, 1993; Jensen, 1984) in part because it's operationalization relies on all of the more narrow subconstructs that contribute (to its overall score).

The narrower subconstructs that collectively contribute to general intelligence include hypothesized constructs such as crystallized abilities, fluid abilities, auditory processing, visualization, long-term retrieval, short-term memory at the second level of organization, memory span, working memory, mental comparison speed, and general sequential reasoning at a third tier (Flanagan et al., 2007); more important, the number of more specific subconstructs increases from the apex to the lowermost level. Flanagan et al. (2007) have identified 10 second-tier subconstructs and approximately 80 at third-tier subconstructs.

Although this three-tiered model has strong empirical (primarily factor-analytic) support, it is not accepted by all experts. In particular, Gardner (1999) objected to the notion of a hierarchical arrangement and described various multiple intelligences (e.g., verbal, musical, spatial, logical–mathematical, bodily–kinesthetic, interpersonal, intrapersonal, and natural) as being more or less independent and relevant for problem solving depending on the task demands. Nonetheless, because of the strong psychometric evidence in support of *g*, many test experts use it as a starting point to develop models of intelligence, including authors of nonverbal tests (e.g., Bracken & McCallum 1998; Wechsler & Naglieri, 2006). Much of the evidence in support of *g* within nonverbal assessment is based on factor-analytic data and the strong relationships between nonverbal test scores and other measures that are assumed to assess *g* well (e.g., verbally laden intelligence tests, real-world products that are assumed to rely on *g* such as school and vocational success).

Because nonverbal tests minimize verbal content to render them more appropriate for examinees presumed to be language limited because of hearing deficits, lack of familiarity with English, and so on, they are not capable of assessing all language-based indicators of intelligence (expressive vocabulary).

As discussed previously, for some examinees these language-loaded test scores are confounded by irrelevant variance, that is, they are influenced by language or culture factors that are not necessarily related to general intelligence. For those examinees who are not limited by language or culture, however, an examiner may be interested in obtaining estimates of intelligence that reflect verbal processing and production; in such cases, nonverbal tests can be supplemented by tests that require varying degrees of language (e.g., receptive, expressive). Even though nonverbal tests are capable of predicting verbally loaded academic content because both constructs are *g* saturated, some examiners may be interested in obtaining estimates of cognitive abilities that are more language laden (e.g., perhaps for assessment of all known functions associated with central nervous system functioning and trauma).

Despite the limited ability of nonverbal measures to access verbal abilities, the literature contains some evidence that purposefully designed nonverbal tests are sensitive to internal verbal mediation. For example, Bracken and McCallum (1998) included in the UNIT subtests that they referred to as *symbolic,* as opposed to other subtests categorized as *nonsymbolic*. These symbolic subtests contain items that are assumed to be more successfully completed by internal verbal mediation using whatever idiosyncratic language system the examinee possesses. The evidence in support of this dichotomy is not conclusive, but the correlation coefficients reported in the UNIT manual between the symbolic subtests and (symbolically laden) academic achievement scores are higher than those between the nonsymbolic subtests and academic achievement for approximately 60% of the comparisons reported. Although some experts have called for more research to determine the validity of this model (Kamphaus, 2001), the data are promising and have been supported by relevant findings from other researchers (Borchese & Gronau, 2005).

In spite of the controversies and perceived limitations associated with the use of nonverbal tests, many experts have concluded that use of nonverbal tests is superior in some situations to the use of verbal tests (Kaufman & Kaufman, 1983; Ortiz & Ochoa, 2005; Sattler, 2001). When the cognitive ability of an examinee appears limited because of poor language skills, lack of exposure to a particular culture, physical trauma, emotional distress, and so forth, language-loaded test scores may be depressed, leading to underestimates of cognitive ability.

HOW NONVERBAL TESTS CAN HELP PROVIDE FAIR(ER) ASSESSMENT

Experts in the field have contributed significantly to the creation of fair(er) test practices. For example, Ortiz's (2002) overarching model for nondiscriminatory testing provides guidelines for practitioners and recommends that examiners (a) develop culturally and linguistically based hypotheses; (b) assess language history, development, and proficiency; (c) assess effects of cultural and linguistic differences; (d) assess environmental and community factors; (e) evaluate, revise, and retest hypotheses; (f) determine appropriate languages of assessment; (g) reduce bias in traditional practices; (h) use authentic and alternative assessment practices; (i) apply cultural–linguistic context to all data; and (j) link assessment to intervention. This model provides a broad-based perspective for creating a fair(er) assessment paradigm. Although testing is only a small part of this overall model, use of nonverbal tests fit nicely into this perspective, subsumed under the recommendation to reduce bias in traditional (testing) practices. It should be recognized that whenever a standardized test is used, even a nonverbal one, cultural loading is never fully eliminated.

Users of standardized tests often erroneously begin the assessment process by accepting the *assumption of comparability*, described by Salvia and Ysseldyke (2004) as the belief that the acculturation of examinees is similar to those on whom the test was standardized. Obviously, that assumption is not always justified. According to Flanagan et al. (2007), a test may produce a biasing effect when it is administered to individuals whose cultural backgrounds, experiences, and exposures are dissimilar from those in the standardization sample, and the effect likely yields lower scores because the tests primarily sample cultural content related to the mainstream milieu and not the full range of experiences of the examinee (also see Valdés & Figueroa, 1994).

Consequently, examiners must take into account the examinee's level of acculturation and the extent to which test performance is culture specific.

To aid in this process of determining acculturation, examiners may explore some general guidelines provided by Cummins (1979). He introduced practitioners to the need to consider two phrases to help distinguish between the very different time periods typically required by immigrant children to acquire conversational fluency in their second language (about 2 years) versus the time required to catch up to native speakers in the academic aspects of the second language (about 5 years). The skill level acquired within the initial 2-year period is referred to as *basic interpersonal communicative skills* and the skill level referenced by the initial 5-year period is called *cognitive academic language proficiency*.

An examinee's cognitive academic language proficiency may be estimated by the procedures described by Woodcock–Johnson III Tests of Achievement and Cognitive Batteries (Woodcock et al., 2001). To match an examinee's level of proficiency and the language or cultural demands of particular tests, examiners can consult Flanagan et al. (2007). They provide criteria characterizing the cultural (and language) loading within their constructed Culture–Language Test Classification (C-LTC) and the Culture–Language Interpretative Matrix (C-LIM). Use of this system is provided in the next section.

The effect that language differences have on test performance may be similar to that of acculturation in that more (rather than less) exposure to either is likely to enhance performance. In the context of assessment, a particular examinee's language proficiency may be limited by lack of exposure to the language and, consequently, be markedly different (and less well developed) than the proficiency level of the typical examinee in the standardization sample. Language exposure is limited for examinees with hearing deficits, those who speak English as a second language, and those who are less familiar with standard English. Consequently, these examinees may be penalized by highly loaded language tests, and the assumption of comparability regarding language may be invalid. For this reason, Figueroa (1990) cautioned examiners to take into account the language history of the examinee within his or her cultural context, not just the obvious language proficiency level. Language proficiency can be operationalized by available instruments (e.g., Woodcock–Muñoz Language Survey—Revised; Woodcock, Muñoz-Sandoval, Ruef, & Alvarado, 2005).

Ideally, test authors and publishers would create tests with standardization samples that are "leveled" by culture and language. That is, authors would use large and targeted samples, representative of all examinees in the population, selected on the basis of knowledge of various levels of acculturation and language proficiency. However, the U.S. population is much too diverse to make leveling feasible. In lieu of this approach, test authors have typically adopted one of two other solutions. They have either developed tests in the native language of examinees (e.g., the Bateria Woodcock–Muñoz III; Woodcock, Munoz-Sandoval, McGrew, & Mather, 2005) or language-reduced or nonverbal tests (e.g., UNIT). Both of these practical solutions are limited, but it is simply not practical to create tests in all the languages spoken by examinees in the United States, and there are not sufficient numbers of linguistically diverse examiners to administer the multitude of possible tests even if they could be developed and normed appropriately. Similarly, language-reduced or nonverbal tests cannot totally eliminate language and culture from the assessment. This point is made clear in the next section focusing on the use of C-LTC and C-LIM procedures.

Before discussing the use and benefits associated with C-LTC and C-LIM strategies, it is important to discuss the difference between cultural loading and bias. Any given test may be culturally loaded but not biased; it might also be biased but not culturally loaded. From a psychometric perspective, bias is a technical term, operationalized by results from particular statistical tests (e.g., correlational and related factor-analytic and model-testing comparisons across groups or use of the chi-square–based Mantel–Haenszel procedure). For example, a test that is biased against a particular group may predict less well for the marginalized group, may yield a factor structure for the marginalized group that is different from that for the mainstream group, and may

in general produce more error in the scores or the prediction for the marginalized group relative to the mainstream group.

A test that is biased against a particular population typically cannot be recommended for use with that population. Even a test that might not produce evidence of statistical bias in a technical sense, however, may be inappropriate for a particular examinee to the extent that its cultural loading leads to a lower score for that examinee. Similarly, a test that is highly language loaded may not be appropriate for examinees who have salient language limitations. How can an examiner know which tests or subtests are best suited for an individual whose scores may be negatively affected by cultural or language limitations? The next section provides some guidance.

Flanagan et al. (2007) recommend using C-LTC and C-LIM as part of the cross-battery assessment (XBA) process when the examiner believes a particular examinee's cognitive performance may be negatively influenced by cultural or language differences or deficits. The XBA approach relies on selecting and administering subtests from a primary battery and supplemental batteries, guided by the unique referral questions and specific criteria offered to ensure use of adequate, psychometrically sound, and relatively pure measures of the cognitive constructs of interest (e.g., auditory processing, long-term memory, short-term memory, visual processing, fluid intelligence, crystallized intelligence, quantitative reasoning). Appropriate use of XBA assessment for diverse individuals requires that the examiner consider how the examinee's unique characteristics may interact with the test content to influence the scores. Flanagan et al. cautioned examiners to keep four essential assumptions in mind: (a) All tests are culturally loaded and reflect the value, beliefs, and knowledge deemed important within the culture; (b) all tests, even nonverbal ones, require some form of language or communication from the examiner and the examinee, which will influence performance; (c) the language and culture loadings of tests vary significantly; and (d) interpretation of standardized tests results (using existing normative data) may be invalid for diverse individuals. The use of C-LTC and C-LIM ensures that these assumptions are addressed in a systematic and logical fashion.

Given that all tests and subtests require some form of language or communication and that all are influenced by the culture within which they are developed, how can these influences be characterized in a manner than guides examiner decision making? To determine the influence of language and culture on test results, Flanagan et al. (2007) relied on three strategies. First, they conducted a literature review of existing test data to determine the nature and extent of cultural and linguistic impact (of these tests). Second, they considered the distributions of scores on various tests for relevant samples. For example, they discovered that bilingual individuals score about 1 standard deviation below the mean of monolingual individuals on many available tests, based on a number of studies (e.g., Cummins, 1984; Mercer, 1979). They also noted that bilingual individuals are required to master the language and mores of two languages and cultures; consequently, they may not appear as proficient as their nonbilingual peers in either language or culture. These data led Flanagan et al. to make decisions regarding the attenuating effects of culture and language. Third, because many of the currently available tests have reported little or no data for bilingual or culturally diverse individuals, an expert consensus procedure was used to determine potential effects of either on the basis of the test characteristics. This classification scheme has since received some empirical support from the work of Nieves-Brull, Ortiz, Flanagan, and Chaplin (2006). From this information, Flanagan et al. created a 2×2 matrix for various tests and subtests, reflecting on one dimension the degree of (mainstream U.S.) cultural loading and on the other the degree of (English) language demand, with three levels of impact, low, moderate, and high.

To actually place tests on the cultural dimension, Flanagan et al. (2007) evaluated several test and subtest characteristics using a rational analyses, including emphasis on process (process dominant vs. product dominant), content (abstract or novel vs. culturally specific), and the nature of the response (use of conventional oral language vs. language-reduced strategies, gestures, pointing, pantomime, modeling). Tests that are more process oriented, that contain more novel content, and rely less on culture-laden content and communication strategies

are presumed to yield fairer scores (Jensen, 1974; Valdés & Figueroa, 1994). Similarly, specific placement on the linguistic demand dimension relied on consideration of several factors, including the extent to which expressive and receptive language is required from either the examiner or the examinee. The point is made that even nonverbal communication strategies such as nods and gestures require an understanding of a language-based system of communication and hence are neither truly linguistically nor culturally free.

More important, the C-LIM characterizations do not reflect the particular constructs presumed to underlie the tests and subtests. As noted in an earlier section, the influence of language and culture may contribute to construct-irrelevant variance. That is, the construct validity of the test may be compromised when the constructs of (level of) culture or language proficiency contribute significantly to the score rather than the constructs of interest. Examiners can guard against these unwanted effects by using the information provided by C-LTC and C-LIM (Flanagan et al., 2007) to guide selection of measures based on these dimensions.

In this chapter, the test-specific culture–language matrix classifications are reported in Table 4.1 for the nonverbal tests that Flanagan et al. (2007) characterized, that is, the UNIT, Leiter–R, and the nonverbal subtests of the Stanford–Binet, Fifth Edition (SB5; Roid, 2003). Also, in Table 4.1, classifications of other nonverbal tests reviewed in this chapter are made on the basis of application Flanagan et al.'s (2007) criteria. Examiners can use the information from Table 4.1 to estimate the extent to which results from an evaluation of a diverse individual may be the result of cultural differences versus a disorder or disability. Because an examinee's score will be reduced to the extent that his or her cultural and linguistic background or capability differs from the background of the individuals in the normative sample, lower performance may be the result of an experiential difference rather than an ability difference. In sum, when a significant experiential difference exists between an examinee's background and the test demands, the validity of the test results are suspect, and interpretation may be limited.

Examiners can determine the extent to which tests and subtests are affected by culture, language, or both by examining the entries in Table 4.1. Tests and subtests in the top leftmost cell are the least affected and those in the bottom rightmost cell are the most affected. According to Flanagan et al. (2007), in the context of XBA it is possible to operationalize this impact by calculating the average test and subtest scores within cells and comparing these average scores, assuming, of course, that the scores are reported (or converted to) the same metric. Impact is operationalized by the extent to which average cell scores increase (or not) as a function of their distance from the top left cell to the bottom right cell.

One of the characteristics of most nonverbal batteries and subtests is the use of limited verbal interactions required for administration. So, for the most part these tests and subtests can be characterized as having a low degree of linguistic demand. In fact, all the UNIT and Leiter–R subtests are contained in the low category, although the subtests vary considerably on the degree of cultural loading dimension. Examiners should remember that many nonverbal tests and subtests are, by their very nature, capable of assessing fluid reasoning, visual processing, and short-term memory and less capable of assessing crystallized ability, quantitative reasoning, and long-term memory. Consequently, fewer nonverbal (than verbal) measures are available for use in the XBA process; nonetheless, Wilhoit and McCallum summarized application of the XBA process for nonverbal tests and subtests in 2003. In the next section, specific comparisons of the eight nonverbal tests reviewed in this chapter are provided along several dimensions. Coupled with C-LIM information, the following content should provide examiners with useful guidelines for test selection and use.

AVAILABLE NONVERBAL TESTS

This chapter describes eight of the most commonly used language-reduced or nonverbal tests, including the CTONI–2 (Hammill et al., 2009); General Ability Measure for Adults (GAMA; Naglieri & Bardos, 1997); Leiter–R (Roid & Miller, 1997); NNAT–I (Naglieri, 2003); Nonverbal Scale of the SB5 (Roid,

TABLE 4.1

Matrix of Cultural Loading and Linguistic Demand Classifications of Subtests From Major Nonverbal Intelligence Tests

	Degree of linguistic demand		
Degree of cultural loading	**Low**	**Moderate**	**High**
Low	CTONI–2 Geometric Analogies Geometric Categories Geometric Sequences TONI–4 Form A and B SB5 NV Visual–Spatial Processing Leiter–R Design Analogies Repeated Patterns Paperfolding Figure Rotation Sequential Order UNIT Spatial Memory Cube Design Mazes WNV Object Assembly Recognition Spatial Span	GAMA Matching Construction Sequences Analogies SB5 NV Fluid Reasoning NNAT–I Spatial Visualization Pattern Completion Reasoning by Analogy Serial Reasoning	SB5 NV Working Memory
Moderate	Leiter–R Visual Coding Matching Attention Sustained UNIT Symbolic Memory WNV Matrices Coding	SB5 NV Quantitative Reasoning	
High	CTONI–2 Pictorial Analogies Pictorial Categories Pictorial Sequences Leiter–R Forward Memory Reverse Memory Spatial Memory Figure Ground Form Completion Picture Context Classification Associated Pairs Immediate Recognition Delayed Pairs Delayed Recognition	SB5 NV Knowledge (Levels 2–3)	SB5 NV Knowledge (Levels 4–6)

(*Continued*)

TABLE 4.1 (*Continued*)

Matrix of Cultural Loading and Linguistic Demand Classifications of Subtests From Major Nonverbal Intelligence Tests

	Degree of linguistic demand		
Degree of cultural loading	**Low**	**Moderate**	**High**
	UNIT Object Memory Analogic Reasoning WNV Picture Arrangement		

Note. CTONI–2 = Comprehensive Test of Nonverbal Intelligence—2; SB5 = Stanford–Binet Intelligence Scales, fifth edition; GAMA = General Ability Measure for Adults; TONI–4 = Test of Nonverbal Intelligence—4; NV = nonverbal; Leiter–R = Leiter International Performance Scale—Revised; NNAT–I = Naglieri Nonverbal Ability Test Individual Administration; UNIT = Universal Nonverbal Intelligence Test; WNV = Wechsler Nonverbal Intelligence Scale.

2003); TONI–4 (Johnsen et al., 2010); UNIT (Bracken & McCallum, 1998); and the WNV (Wechsler & Naglieri, 2006). The Raven Progressive Matrices (J. C. Raven, Raven, & Court, 1998) is another well-known nonverbal measure, with a rich history of worldwide use that dates to 1938.

The Raven Progressive Matrices is not a single test but a family of matrix reasoning tests, and it is one of the best researched nonverbal measures in the world; however, the Raven products were not included in this review because of the absence of U.S. norms (McCallum et al., 2001).

In this section, various characteristics of each test are presented, including a description of age ranges, subtests and scales, test administration procedures, and so on. In addition, psychometric properties of the tests are compared across tests and evaluated. Operationalizations of reliability include evidence of internal consistency and test stability, as reported in the respective test manuals. Various elements of validity are reported, as defined according to the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999), including evidence based on test content (e.g., what are the item characteristics and demands?), response processes (e.g., are cognitive requirements of items consistent with processes the test purports to measure?), internal structure (e.g., is the factor-analytic structure consistent with proposed test interpretation?), relationships with other variables (e.g., are converging, discriminant, and test-criterion relationships consistent with the proposed interpretation of the test?), and consequences of test use (e.g., can the test results be applied in the manner consistent with test authors' stated goals for improving instruction, making predictions?). Finally, scale properties (e.g., subtest floors and ceilings, item gradients) and test fairness characteristics (e.g., expert review, analyses addressing potential bias) are presented, compared across tests, and evaluated according to accepted standards in the field.

General Instrument Characteristics

Some of the general characteristics of the nonverbal tests reviewed in this chapter are shown in Table 4.2, including the test name, age range of examinees, and brief descriptions of the scales and subtests. Tests are included on the basis of their (a) clear identity as a nonverbal instrument, either by name or by author goals, and (b) utility, as defined by recency and representativeness of standardization data for the U S population.

The psychometric properties of the eight nonverbal tests are presented in tabular form, with elaboration in the narrative as needed. Relevant criteria for evaluating the tests are based on the generally accepted standards, either using Bracken's (1987) recommendations for minimal technical adequacy or the *Standards for Educational and Psychological Tests* (American Educational Research Association et al., 1999).

TABLE 4.2

General Characteristics of Tests Reviewed

Test	Age (years)	Scales or subtests
CTONI–2	6–89	Pictorial IQ, Geometric IQ, Nonverbal IQ; Analogic reasoning, categorical classifications, sequential reasoning
GAMA	18 and up	GAMA IQ; 4 subtests
Leiter–R	2–20	Brief IQ Screener, FSIQ; 10 Visualization/Reasoning and 10 attention–memory subtests
NNAT–	5–17	Total test score; Pattern Completion, Reasoning by Analogies, Serial Reasoning, Spatial Visualization
SB5	2 and up	Abbreviated Battery IQ, FSIQ; 5 verbal IQ and 5 nonverbal IQ subtests
TONI–4	6–89	Total test score
UNIT	5–17	Abbreviated Battery IQ, FSIQ, Memory Quotient, Reasoning Quotient, Symbolic Quotient, and Nonsymbolic Quotient Extended Battery; 6 subtests
WNV	4–21	FSIQ; 6 nonverbal IQ subtests

Note. CTONI–2 = Comprehensive Test of Nonverbal Intelligence—2; GAMA = General Ability Measure for Adults; Leiter–R = Leiter International Performance Scale—Revised; NNAT–I = Naglieri Nonverbal Ability Test Individual Administration; SB5 = Stanford–Binet Intelligence Scales, fifth edition; TONI–4 = Test of Nonverbal Intelligence—4; NV = nonverbal; UNIT = Universal Nonverbal Intelligence Test; WNV = Wechsler Nonverbal Intelligence Scale; FSIQ = full-scale IQ.

Reliability. Reliability coefficients provide operationalizations of the systematic variance assessed by a test and are critical sources of information for determining psychometric integrity. Within classical test theory, the reliability coefficient is assumed to represent the percentage of systematic variance contributed by the test, and test error is defined as 1 − reliability. Reliability can be defined by coefficients reflecting the internal consistency of a test, or subtest, usually obtained by some variation of the split-half procedure using the Spearman-Brown or the alpha statistic, or by test–retest data, typically collected within a short period of time, most often no more than 2 to 4 weeks. For the purposes of summarizing multiple reliability coefficients, mean or median reliabilities are sometimes reported.

Median subtest internal consistency. In Table 4.3 median internal consistency reliabilities across subtests are shown, by age, as available. Typically, subtests are not used for making high-stakes decisions but are combined to create global scores, which may then be used to determine eligibility for placement within programs (e.g., special education) and diagnoses (e.g., intellectual disability, learning disability). Consequently, the minimum recommended reliability estimate for subtests is less rigorous (i.e., .80) than for global scores (i.e., .90) according to experts (e.g., Bracken, 1999; Salvia & Ysseldyke, 2004).

Examination of Table 4.3 reveals that the average subtest reliability scores across the eight tests meet the .80 criterion for most ages, with some exceptions. Some of the tests have only reported total scores, so median subtest reliabilities are reported for them. Of the six tests reporting subtest scores, only the CTONI–2 and the SB5 median subtest reliabilities are above .80 for every age, and of the 19 values reported for the CTONI–2, 15 are .90 or above. The GAMA yields the lowest median values, with only two of the 11 median values at .80 or above.

Total test internal consistency indices. In Table 4.4 total internal consistency indices are shown, by age. Psychometric experts (e.g., Bracken, 1987; Saliva & Ysseldyke, 2004) have recommended a minimum coefficient of .90, assuming high-stakes use of the test, such as placement purposes, diagnostic use, and so on. As is apparent, the average reliability coefficients all meet this criterion, ranging from .90 to .97. Most also meet the criterion at the individual ages, with some exceptions. For example, coefficients for GAMA are slightly below the .90 value

TABLE 4.3

Median Subtest Internal Consistency Coefficients by Age Groupings

Instrument	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–89	*Mdn*
CTONI–2	N/A	N/A	N/A	N/A	.85	.87	.87	.90	.90	.90	.90	.90	.90	91	.91	.92	.94			.90		.90		.90		.88		.93			.91		.90
GAMA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.76			.78		.77		.82		.80		.78	.67	.76	.67	.61	.58	.73
Leiter–R																																		
VR	.89	.85	.82	.81	.80	.74	.78	.79	.81	.82	.82		.83		.76		74			N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.81
AM	.83		.85		.80		.76				.82					.73				N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.81
NNAT–I[a]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SB5[b]	.87	.90	.87	.87	.81	.84	.83	.83	.81	.82	.86	.81	.86	.87	.86		.88			.90		.89		.88		.89		87			.89		.84	.86
TONI– 4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
UNIT	N/A	N/A	N/A	.82	.79	.79	.83	.80	.82	.79	.82	.82	.81	.79	.81		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.80
WNV	N/A	N/A	.82	.83	.83	.83	.77	.82	.79	.76	.78	.84	.81	.85	.83		.79			.82		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.81

Note. CTONI–2 = Comprehensive Test of Nonverbal Intelligence—2; GAMA = General Ability Measure for Adults; Leiter–R = Leiter International Performance Scale—Revised; VR = visualization/reasoning subtests; AM = attention/memory subtests; NNAT–I = Naglieri Nonverbal Ability Test Individual Administration; SB5 = Stanford–Binet Intelligence Scales, fifth edition; TONI–4 = Test of Nonverbal Intelligence—4; UNIT = Universal Nonverbal Intelligence Test; WNV = Wechsler Nonverbal Intelligence Scale; N/A = not applicable. From *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (2nd ed., p. 564), by D. P. Flanagan and P. L. Harrison (Eds.), 2005, New York, NY: Guilford Press. Copyright 2005 by Guilford Press. Adapted with permission.

[a]Only total test scores provided. [b]Nonverbal subtests only.

TABLE 4.4

Total Test Internal Consistency Coefficients by Age Groupings

Instrument	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–89	+ *Mdn*
CTONI–2[a]	N/A	N/A	N/A	N/A	.95	.96	.96	.96	.97	.97	.97	.97	.97	.97	.97	.97	.97			.96		.96		.96		.96		.97			.97		.97
GAMA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A		.91		.91	.90		.94		.92		.92	.88	.91	.86	.85	.79	.90
Leiter–R[b]			.92				.91						.93					.93		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.90
NNAT–I	N/A	N/A	N/A	.89	.90	.89	.89	.89	.90	.88	.92	.92	.94	.94	.95		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SB5[a]	.95	.96	.95	.95	.94	.94	.94	.95	.94	.94	.94	.93	.95	.96	.96		.96			.96		.95		.97		.97		.97			.97		.96	.95
TONI–4 (Form A)	N/A	N/A	N/A	N/A	.89	.93	.90	.89	.89	.91	.93	.93	.93	.93	.93	.92	.91			.94		.95		.93		.90		.95			.95		.97	.93
TONI–4 (Form B)	N/A	N/A	N/A	N/A	.94	.93	.94	.95	.95	.96	.95	.95	.96	.97	.96	.96	.95			.97		.97		.97		.97		.96			.97			
UNIT[c]	N/A	N/A	N/A	.92	.91	.91	.92	.91	.93	.92	.94	.94	.93	.93	.94		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.93
WNV[c]	N/A	N/A	.91	.90	.92	.91	.91	.91	.87	.88	.92	.90	.92	.92	.92		.93			.92		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.91

Note. CTONI–2 = Comprehensive Test of Nonverbal Intelligence—2; GAMA = General Ability Measure for Adults; Leiter–R = Leiter International Performance Scale—Revised; NNAT–I = Naglieri Nonverbal Ability Test Individual Administration; SB5 = Stanford–Binet Intelligence Scales, fifth edition; TONI–4 = Test of Nonverbal Intelligence—4; UNIT = Universal Nonverbal Intelligence Test; WNV = Wechsler Nonverbal Intelligence Scale; N/A = not applicable. From *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (2nd ed., p. 565), by D. P. Flanagan and P. L. Harrison (Eds.), 2005, New York, NY: Guilford Press. Copyright 2005 by Guilford Press. Adapted with permission.

[a]Coefficients are based on Nonverbal IQ. [b]Coefficients are based on Full Scale IQ. [c]Coefficients are based on Extended Battery Full Scale IQ

for ages 60–64 and for age categories above 70, for the NNAT–I at ages 5, 7, 8, 9, and 11, and for the TONI–4 (Form A) at ages 6, 9, and 10. More important, even when the indices do not make the .90 criterion, they are typically very close, at .88 or .89.

Total test stability indices. For determining the stability of the nonverbal measures, authors have reported test–retest data; the test–retest interval ranged from 1 day to 49 days, but most ranged from 1 to 3 weeks. Again the .90 value has been adopted by some experts (e.g., Bracken, 1987) as the recommended test stability criterion for high-stakes use (for global scores). There is more variability across these coefficients, relative to the internal consistency indices. These coefficients range from .67 to .96, although most range from the mid-.80s to the mid-.90s. Only the CTONI–2 and the Leiter–R indices are above .90 for every age group reported across all forms. The GAMA yielded the lowest coefficient, .67; the NNAT–1 yielded indices ranging from .68 to .78. The UNIT yielded one coefficient below .80: .78 for 5- to 7-year-old examinees. All the other coefficients are .85 and better.

Scale characteristics. Subtest and total scores can be characterized using a number of criteria. Perhaps the most important two for these nonverbal intelligence tests include the extent to which they possess (a) adequate floors and ceilings and (b) reasonable difficulty gradients. The floor and ceiling reflect the ability of the tests to assess the youngest most delayed and the oldest most capable examinees, respectively. The difficulty gradient reflects the extent to which the test discriminates examinees within the distribution of scores produced, that is, whether the test is capable of producing fine differentiations among the adjacent scores after raw scores are transferred to standard scores.

Floors and ceilings. To provide a sensitive assessment of a construct such as intelligence, a test must yield scores that discriminate the lowest and highest 2% of the population from the remaining 98% (Bracken, 1987). Braden and Athanasiou (2005) noted that this criterion corresponds (roughly) to 2 standard deviations from the mean, or a z score of either $+2.0$ or -2.0; using this process, they recommended calculation of z scores for the lowest and highest standard scores possible by age. For instruments that included multiple norm tables for each year, they used an average of the subtests for each age interval. Using this method, all measures are capable of distinguishing the top and bottom 2% when total test scores are used.

Not all of the nonverbal tests reviewed have reported subtest scores. Most of the nonverbal tests reporting scores at the subtest level have ample subtest floors and ceilings, but there are some exceptions. For example, floors for the CTONI–2 at ages 6, 7, and 8 are -1.22, -1.56, and -1.83, respectively. The Leiter–R and SB5 floors for age 2 are -1.33 and -1.91, respectively. For the GAMA, floors for examinees older than 65 are also less than the criterion z of 2.0; for examinees ages 65–69, the z score is -1.83, for those 70–74, the z score is -1.75, for those 75–79, the z score is -1.58, and for those 80 and older, the z score is -1.58. Neither the UNIT nor the WNV produced subtest scores below the criterion, suggesting that these instruments can differentiate the lowest 2% of the population. The situation was similar for ceilings. Only the CTONI–2 produced inadequate ceilings, that is, a z score of 2.0 or greater. Examinees ages 30–39 and 40–49 yielded a z score of 1.89.

Difficulty gradients. Not only should tests be sensitive enough to register changes of abilities at the extremes, but they should be capable of distinguishing modest changes throughout the range of ability assessed. According to Bracken (1987), three (or more) raw score points should be associated with every standard score standard deviation change. Apparently, all the nonverbal tests reviewed are sensitive enough to reflect small ability gains. To reach this conclusion, the total number of raw score points available was divided by the number of standard deviations spanned by the subtest, at each age; when necessary, averaged multiple scores within a 1-year age level were used. The median of quotients across subtests was determined, and any quotient greater than 3 was assumed to meet the minimum criterion.

Validity. Evidence for the validity of the commonly used nonverbal instruments can be found in their respective test manuals and in the literature.

Validity is multifaceted and cannot be referenced by a particular numerical value or statistical analysis. Rather, the evidence accrues over time and is cumulative and certainly may include empirical data as well as theoretical underpinnings. The *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) provide perhaps the most comprehensive framework for defining validity; within this framework are five basic sources of evidence, including test content, response processes, internal structure, relationship to other variables, consequences of test use (see Table 4.5).

Test content. Evidence of validity based on test content refers to the format of test items, nature of the examinee tasks, administration procedures, relationship between the item content and the underlying theoretical constructs presumably assessed, and the extent to which there is expert agreement regarding the appropriateness of the content.

Response process. Evidence of validity based on response processes requires either theoretical or empirical support that the examinees use the hypothesized processes assessed, as described by test authors.

Internal structure. Evidence of validity based on internal structure requires a match (consistency) between the test items, subtests, and scales and the constructs presumably measured by these test components.

Relationships to other variables. Evidence of validity based on the relationship between the test and other variables subsumes convergent and divergent data with similar and dissimilar measures and predictive data with desirable and hypothesized outcomes.

Consequences of testing. Evidence based on consequences of test use requires that desirable outcomes are documented.

These five sources of evidence have been operationalized and summarized by Braden and Athanasiou (2005) in their excellent review for most of the currently used instruments, and their tables have been adapted, and updated, as necessary (see Table 4.5).

Table 4.5 provides a summary of the evidence obtained from the test manuals. Although all of the tests provide significant evidence for validity for most categories, data are scarce for three of the categories: (a) empirical basis for operative cognitive processes, (b) experimental intervention studies, and (c) evidence based on consequences of testing. These are salient limitations. As Braden and Athanasiou (2005) noted, it is difficult to establish the specific nature of cognitive processes engaged during completion of particular tasks because of the limitations associated with verifying the processes actually used. Also, not all examinees will engage the same processes. This limitation is not unique to nonverbal tests, as Braden and Niebling (2005) noted. As brain-based research techniques (e.g., functional magnetic resonance imaging, positron emission tomography scans) become less expensive and less difficult to actually use (and interpret), test authors will likely begin to use them to establish central nervous system functioning sites related to task completion (see Wilkie, 2009). Such efforts may soon provide some support for the theoretical links currently explicated for the measurement strategies used.

Similarly, evidence in support of test scores to guide intervention studies is difficult to obtain for a variety of reasons, some methodological and some practical. The Leiter–R is the only nonverbal test to provide some evidence in support of this category, but the data are not particularly strong. Data from traditional verbally laden tests are increasingly becoming available in support of their ability to support interventions (e.g., Kaufman & Kaufman, 1983; Mather & Jaffe, 2002), and test authors of nonverbal tests will be under increasing pressure to provide this type of evidence. Related to this point, none of the test manuals provided positive predictive evidence to support of test consequences, that is, that examinees benefit academically from test results, although there are related data. Most of the tests have reported positive correlation coefficients between test scores and academic performance. Presumably, examinees who have been placed into classes for special needs students on the basis of test outcomes are better off than they were before the test data were applied to the decision-making process. However, this argument is more speculative than empirical.

Test Fairness

All test authors aspire to develop fair tests, but because nonverbal tests will be used primarily to

TABLE 4.5

Evidence for Validity Across Instruments

Category and criterion	CTONI–2	GAMA	Leiter–R	NNAT–I	SB5	TONI–4	UNIT	WNV
Evidence based on test content								
Item relevance	x	x	x	x	x	x	x	x
Item adequacy or relevance					x	D	x	
Detailed content	x	x	x	x	x	x	x	x
Expert judgment			x		x		x	x
Expert judges and procedures used					x		x	x
Evidence based on response processes								
Theoretical basis for operative cognitive processes	x		x		x	x	x	x
Empirical basis for operative cognitive processes				D				
Evidence based on structure								
Age trends differentiation	E	E	E		E	E		E
Group differentiation	E	E	E	E	E	E	E	E
Item analysis	E	D	D	D	D	D	D	
Differential item functioning	E	D	E		E	D	D	E
Factor analysis	E		E		E	E	E	E
Item–subtest correlations		D		N/A		N/A		
Item–total correlations		D	D		D	D		
Subtest intercorrelations	E	E	E	N/A	E	N/A	E	E
Subtest–total correlations		E	E	N/A	E	N/A	E	E
Experimental intervention studies			E					
Evidence based on relations to other variables								
Correlations with other measures of ability	E	E	E	E	E	E	E	E
Correlations with other nonverbal measures of ability	E		E		E	E	E	E
Correlations with achievement measures		E	E	E	E	E	E	E
Cross-battery factor analysis			E		E			
Evidence based on consequences of testing								
Prediction analyses								

Note. CTONI–2 = Comprehensive Test of Nonverbal Intelligence—2; GAMA = General Ability Measure for Adults; Leiter–R = Leiter International Performance Scale—Revised; NNAT–I = Naglieri Nonverbal Ability Test Individual Administration; SB5 = Stanford–Binet Intelligence Scales, fifth edition; TONI–4 = Test of Nonverbal Intelligence—4; UNIT = Universal Nonverbal Intelligence Test; WNV = Wechsler Nonverbal Intelligence Scale; x = instrument meets criterion; D = discussed in test manual; E = empirical evidence provided in test manual. From *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (2nd ed., p. 568), by D. P. Flanagan and P. L. Harrison (Eds.), 2005, New York, NY: Guilford Press. Copyright 2005 by Guilford Press. Adapted with permission.

assess individuals with native language limitations, authors of nonverbal tests have even more reason to attend to that goal. Minimizing language use on the part of the examiner and examinee alone will not ensure a fair test for these populations. Nor will ensuring good psychometric properties, although technical adequacy is a necessary (although not sufficient) condition. As Athanasiou (2000) pointed out,

test fairness is based on several general indicators, including appropriate test development foundation, or model, content, test administration procedures, and empirical and statistical data. Specifically, do the test authors address fairness by choosing a model that is reasonable, given the populations to be assessed; is the content culturally and linguistically reduced; do the authors create appropriate and user-friendly administration and response formats; and do the authors report statistical data in support of fair use for the intended populations? Within the more general categories, specific categories exist, as is apparent from the information reported later. Attention to these criteria will help examiners reduce potential bias that might otherwise be part of the assessment process for these populations, and the statistical analyses will help to provide evidence showing the extent to which the efforts have been successful.

Fairness related to test development. Test fairness begins with consideration of the test structure, typically related to some theoretical model, and the extent to which the test model lends itself to nonbiased assessment. The process may continue by requesting that experts review the model and the test materials for evidence of bias. Table 4.6 provides information regarding these issues for the eight nonverbal tests reviewed.

As previously noted, nonverbal tests can be divided into two general categories: unidimensional and multidimensional (McCallum et al., 2001). Unidimensional tests primarily assess reasoning or visual processing, typically via a matrix analogies format. Multidimensional tests are more comprehensive, and may assess visual processing, reasoning, short-term memory, long-term memory, attention, fluid intelligence, and processing speed. Of all the nonverbal tests reviewed in this chapter, only the Leiter–R, SB5, UNIT, and WNV are multidimensional; all the others are unidimensional. Unidimensional tests are more subject to construct underrepresentation than multidimensional tests, that is, they typically provide only one operationalization of intelligence rather than several.

As a rule, the multidimensional tests are based on an explicated theoretical model, which guides

TABLE 4.6

Indices of Fairness Related to Test Development Across Instruments

Instrument	Foundation	Content review
CTONI–2	No specific theoretical model; unidimensional	No expert review
GAMA	No specific theoretical model; unidimensional	No expert review
Leiter–R	Carroll (1993) and Gustafsson (1988) models; multidimensional	No expert review
NNAT–I	No specific theoretical model; unidimensional	No expert review
SB5	Cattell–Horn–Carroll model; multidimensional	Eight expert reviewers examined items for potential bias; expert reviews conducted from perspectives of diverse groups
TONI–4	Founded on no specific theoretical model but presumed to assess generalized intelligence and fluid reasoning; unidimensional	No expert review
UNIT	Jensen's (1980) parsimonious two-factor model; multidimensional	Expert reviews from consultants representing diverse backgrounds reviewed items, artwork, manipulables
WNV	No specific theoretical model; multidimensional	Reviews from practitioners and researchers; reviewed aesthetics of artwork, usefulness of directions, and procedures

Note. CTONI–2 = Comprehensive Test of Nonverbal Intelligence—2; GAMA = General Ability Measure for Adults; Leiter–R = Leiter International Performance Scale—Revised; NNAT–I = Naglieri Nonverbal Ability Test Individual Administration; SB5 = Stanford–Binet Intelligence Scales, fifth edition; TONI–4 = Test of Nonverbal Intelligence—4; UNIT = Universal Nonverbal Intelligence Test; WNV = Wechsler Nonverbal Intelligence Scale. From *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (2nd ed., p. 570), by D. P. Flanagan and P. L. Harrison (Eds.), 2005, New York, NY: Guilford Press. Copyright 2005 by Guilford Press. Adapted with permission.

interpretation by examiners. The authors of the Leiter–R, SB5, and the UNIT provided discussions of the models used to construct the instruments within the manuals. For example, the Leiter–R is based on the work of Carroll (1993), Horn and Cattell (1966), and Gustafsson (1988). It assumes a hierarchical design with *g* at the apex, and the four broad factors of fluid reasoning, fundamental visualization, attention, and memory at the second level. The SB5, as with the Leiter–R, is also based on the work of Carroll (1993) and Horn and Carroll (1966) but more specifically relies on the Cattell–Horn–Carroll model, as described by McGrew and Flanagan (1998). This model explicates a number of broad cognitive abilities (e.g., fluid reasoning, crystallized knowledge, visual processing, auditory processing) and many more narrow abilities. The nonverbal portion of the SB5 assesses only some of these broad abilities, including fluid reasoning, knowledge, quantitative reasoning, visual–spatial reasoning, and working memory, and also provides a global score when both the nonverbal and the verbal subtests are administered, yielding an operationalization of *g*. The UNIT was developed to provide a psychometrically strong operationalization of *g*, defined by a full-scale score. Composite scores are also available for three memory and three reasoning subtests, based on the Level I versus Level II cognitive model described by Jensen (1980). Using a crossed design task, three of the six UNIT subtests can be more easily solved by relying on symbolic convert mediation; three, by relying on nonsymbolic processing. In addition to memory and reasoning composites, examiners may calculate composite scores for the symbolic and nonsymbolic subtests. Although the authors of the WNV indicated that the test is multidimensional in that it assesses visual–spatial ability, recall of sequenced information, and pencil-and-paper skills as well as general ability, no unifying model is described in the manual.

The content of nonverbal tests and subtests can be evaluated in terms of their psychometric properties (e.g., reliability, validity, bias) and appropriateness for the target population. Typically, statistical procedures are used to determine the psychometric properties; expert panels are used to offer judgments regarding the appropriateness of the items, materials, and so forth. The Flanagan et al. (2007) C-LTC model is described in the Use of C-LTC and C-LIM section of this chapter. They relied on both experts and statistical data to describe the cultural and linguistic effects on test data. The strategies they describe led to the categorization in Table 4.1 for the nonverbal tests reviewed in this chapter. Experts from diverse backgrounds also reviewed the SB5, UNIT, and WNV content and materials during development of the tests to ensure inclusion of appropriate content.

Fairness related to test administration procedures. Administration procedures vary considerably across the eight tests reviewed in this chapter, with most test authors describing procedures and accommodations within the respective manuals for special populations, required item presentation and response modes, use of time limits, and the extent to which the test directions rely on teaching items (see Table 4.7).

When possible, a cognitive test should be administered using the examinee's primary communication mode or language. To accommodate examinees who primarily use other languages, some nonverbal tests provide translations when spoken language is required or used (NNAT–1, SB5, TONI–4, WNV). Similarly, the WNV includes a section describing the considerations for testing examinees who are deaf or hard of hearing using several modes of communication (e.g., American Sign Language, sign-supported English, cued speech, auditory/oral technique), based significantly on the work of Hardy-Braz (2003). However, as discussed earlier in this chapter, the U.S. population is so diverse that test authors could not possibly create translations or standardizations for all non-English-speaking examinees. Consequently, nonverbal tests are more likely to be used than translated tests, except perhaps when high-incidence languages are used (e.g., Spanish, in the United States).

According to some experts, a nonverbal test should not require spoken language on the part of the examiner or examinee (Bracken & McCallum, 1998; Jensen, 1980). Administration procedures that require spoken or written language can limit understanding of task demands for individuals who use a different language and hence make the test results

TABLE 4.7

Indices of Test Fairness Related to Test Administration Across Instruments

Instrument	Presentation of instructions	Response modes	Use of time limits	Practice or teaching items
CTONI–2	Pantomimed or verbal instructions	Pointing; clicking mouse if computerized version used	No time limits	Unscored sample items
GAMA	Directions written in English in test booklet and read aloud	Marking bubbles on answer sheet	No time limits on items; time limit of 25 minutes for entire test	Unscored sample items
Leiter–R	Pantomimed instructions, but examiner can supplement with spoken language if necessary	Pointing, placing or arranging cards, marking in test booklet	Bonus points for speed and accuracy on three subtests; pacing procedure for slow responders	Scored reaching items
NNAT–I	Spoken English; Spanish and French translations included in manual	Speaking or pointing	No time limits	Unscored sample items
SB5	Spoken English	Handing examiner items, pointing, completing puzzles, finding objects, pantomiming actions, tapping blocks, arranging geometric pieces	Time limits on one nonverbal subtest	Unscored sample items and scored teaching items
TONI–IV	Pantomimed or verbal instructions; translations in English, Spanish, French, German, Chinese, Vietnamese, Korean, and Tagalog	Pointing or simple gestures	No time limits	Unscored training items
UNIT	Pantomimed	Pointing, placing chips on grid, arranging cards, completing drawn mazes, building cube designs	Time limits on one subtest; bonus points for speed on one subtest	Unscored demonstration and sample items; scored checkpoint items
WNV	Pantomime; pictorial or verbal instructions; French, Spanish, Chinese, German, and Dutch translations included in the manual	Pointing, drawing symbols, arranging puzzle pieces, tapping blocks	Time limits on one subtest; bonus points for speed on one subtest	Unscored demonstration and sample items

Note. CTONI–2 = Comprehensive Test of Nonverbal Intelligence—2; GAMA = General Ability Measure for Adults; Leiter–R = Leiter International Performance Scale—Revised; NNAT–I = Naglieri Nonverbal Ability Test Individual Administration; SB5 = Stanford–Binet Intelligence Scales, fifth edition; TONI–4 = Test of Nonverbal Intelligence—4; UNIT = Universal Nonverbal Intelligence Test; WNV = Wechsler Nonverbal Intelligence Scale. From *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (2nd ed., p. 571), by D. P. Flanagan and P. L. Harrison (Eds.), 2005, New York, NY: Guilford Press. Copyright 2005 by Guilford Press. Adapted with permission.

less valid for those individuals. More important, even tests that that can be administered without spoken language may not be free of cultural "baggage," and some communication modes may be more culturally laden than others. For example, use of gestures may still require some minimal cultural familiarity with the particular gestures used (see Ehrman, 1996; Greenfield, 1998) and may be more

culturally and linguistically laden than simply using pantomime, modeling, and other types of teaching items, with demonstration. Nonetheless, tests that can be administered without spoken language are less likely to disenfranchise and penalize examiners relative to those that do (use spoken language).

Some tasks are more likely to be influenced by cultural or linguistic familiarity than others. For example, tasks using pictures of concrete objects that exist within a particular culture may be more culturally or linguistically laden than pictures of abstract figures. The Picture Arrangement subtest of the WNV represents an example of a "nonverbal" task that demands a high level of nonverbal receptive language (Flanagan et al., 2007). All others things being equal, tasks that require less receptive or spoken language will be fair(er) than those that require more receptive or spoken language; similarly, those that use fewer gestures will be less culturally and linguistically laden than those that require more; and finally, those tests that require only demonstration and modeling of the actual task demands may be the least affected by language and culture, relative to those that require more complex directions (to communicate task demands).

Test authors can take important steps to reduce potential cultural and linguistic influences of communication modes that are used by examiners. For example, using narrative and audiovisual formats, the authors of one nonverbal test (Bracken & McCallum, 1998) described use of eight specific administration gestures, which were used in the standardization of the test. They reduced the language and cultural demands of the gestures by encouraging examiners to communicate the meaning of the (eight specific) gestures to examinees using demonstration and by using the examinees' unique language and communication skills (Bracken & McCallum, 1998). Thus, even diverse examinees should understand the meaning of the gestures before beginning the test, without having to rely on their familiarity with U.S. culture and the gestures that may be somewhat unique to it. Not all tests provide this level of specificity regarding the use of gestures. For example, the directions for use of gestures on the Leiter–R have been described as sometimes broad and confusing (McCallum et al., 2001), which may contribute to less systematic use of the gestures and could inadvertently contribute to test error.

In spite of this discussion admonishing examiners to limit spoken communication with examinees, there is caveat. Spoken language can be used to build rapport and provide a context for the administration for those examinees who share a common language. Examiners who do not share a common language may enlist the aid of someone who does for this purpose. In Table 4.8, test procedures are summarized in terms of task presentation modes, task response modes, use of time limits, and use of teaching tasks.

Fairness related to task presentation and response modes. Of the eight tests reviewed, five can be administered without use of spoken language: the CTONI–2, Leiter–R, TONI–4, UNIT, and the WNV, although the instructions for the Leiter–R, TONI–4 and WNV allow for use of some verbal instructions (see Table 4.7). Instructions for the SB5 and NNAT–1 are presented orally. The GAMA instructions are provided in written and oral English. The performance of diverse examinees may be influenced negatively to the extent that the examiner relies on spoken or written language; consequently, the GAMA and SB5 may be most affected.

As with task presentation modes, task response modes for the tests vary considerably, although none require spoken language (see Table 4.7). Some tests require only a pointing response (i.e., CTONI–2, TONI–4, and NNAT–1). With the exception of the GAMA, all the other tests allow a pointing response; the GAMA requires that examinees use a pencil to shade an oval in a bubble format, which may be an unfamiliar response format for some diverse examinees. Typically, the multidimensional tests require multiple response formats, including manipulating or arranging cubes or chips, arranging cards, using a pencil to draw, pantomiming, and finger tapping. According to Braden and Athanasiou (2005), using varied responses may actually increase examinees' assessment motivation.

Fairness related to use of time limits. Individuals within some cultures prioritize and value speeded performance. However, speeded performance is not valued similarly across all cultures (Harris,

TABLE 4.8

Indices of Fairness Related to Statistical Analyses Across Instruments

Instrument	Group comparison	Item functioning	Reliability	Internal structure	Consequences of testing
CTONI–2	Comparisons included; no matched controls[a]	Three-parameter item response theory and delta scores[a]	Internal consistency for sexes, three minority groups, one non-native group, and subjects with deafness and English as a second language status[b]	No information	Correlations only[a]
GAMA	No comparisons	No information	No information	No information	No information
Leiter–R	Minimal comparisons; no matched controls[a]	Rasch item analysis[b]	No information	No information	Regression analysis for European Americans and African Americans[b,c]
NNAT–I	Comparisons with matched controls□	Item response bias analysis (p. 34)[b]	No information	No information	No information
SB5	No comparisons	Mantel–Haenszel procedure[b]	Internal consistency estimates for three minority groups[a]	Chi-square tests of significance on subtest correlation matrices[b]	Equivalence of regression slopes across sexes and racial and ethnic groups[b]
TONI–4	Comparisons included; no matched controls[a, d]	Differential item functioning analyses	Internal consistency estimates for sexes, three minority groups, and seven special groups[a]	No information	Correlations and logistic regression
UNIT	Comparisons with matched controls included[a]	Mantel–Haenszel procedure[b]	Internal-consistency estimates for sexes and two minority groups	Confirmatory factor analyses on sexes and two minority samples	Equivalence of regression slopes for sexes, races, and Sex × Race
WNV	Comparisons with matched controls included[a]	No information	Internal consistency coefficients for nine special groups	No information	No information

Note. CTONI–2 = Comprehensive Test of Nonverbal Intelligence—2; GAMA = General Ability Measure for Adults; Leiter–R = Leiter International Performance Scale—Revised; NNAT–I = Naglieri Nonverbal Ability Test Individual Administration; SB5 = Stanford–Binet Intelligence Scales, fifth edition; TONI–4 = Test of Nonverbal Intelligence—4; UNIT = Universal Nonverbal Intelligence Test; WNV = Wechsler Nonverbal Intelligence Scale. From *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (2nd ed., p. 573), by D. P. Flanagan and P. L. Harrison (Eds.), 2005, New York, NY: Guilford Press. Copyright 2005 by Guilford Press. Adapted with permission.

[a]Statistical evidence presented in manual. [b]Information discussed in manual; no statistical evidence provided. [c]Reader is referred to Woodcock, McGrew, and Mather (2001). [d]Descriptive statistics only.

Reynolds, & Koegel, 1996; Padilla, 2001). Therefore, tests that yield scores highly influenced by timed performance may unfairly penalize examinees from cultures that do value speed. Table 4.7 shows the test characteristics on this dimension. Most of these nonverbal tests rely very little on speed and are not sensitive to this potential influence. The CTONI–2, NNAT–1, and TONI–4 have no time limits. The GAMA has a time limit on the entire test only, but not for individual items. The Leiter–R

provides bonus points for speed on three subtests, and examiners can encourage slow examinees to hurry. The SB5, UNIT, and WNV provide time limits on one subtest. The UNIT and WNV provide bonus points for speed on one subtest. Examiners who believe that speed should not be reflected in scores will pick and choose tests and subtests accordingly.

Fairness related to use of teaching tasks. Because nonverbal tests typically rely less on spoken administration directions than do verbally laden tests, authors of nonverbal tests rely more on demonstration, practice, and teaching items to convey task demands to examinees. In fact, all the nonverbal tests rely on unscored teaching or sample items to convey task demands (see Table 4.7). Both the UNIT and the WNV include unscored demonstration and sample items; in addition, the UNIT includes a checkpoint item, which is scored but allows for feedback (if the item is failed).

In general, task demands are less complex for unidimensional tests than for multidimensional tests; most unidimensional tests rely on a single-matrix analogies format or a modified-matrix analogies format for all items and require a pointing response only. The multidimensional batteries include multiple item formats and require a variety of response demands, as mentioned earlier. Teaching tasks may be less critical for the GAMA and SB5 for English-speaking examinees because of the reliance on English-language formats to convey task demands on these tests

Fairness related to statistical analyses. Although intelligence test scores have been used for years to aid educators and mental health professionals in making decisions about the eligibility of students for receipt of special education and related services, the use of these tests increased after 1974, when the Education for all Handicapped Children Act of 1975 mandated identification of all children eligible for special education. This "child find" effort created a greater need for intelligence testing, in part because the law encouraged the use of tests, including intelligence tests, in completing this process (Kamphaus, 2001). However, as the use of intelligence tests increased, they came under more intense scrutiny, primarily because many consumers became aware that cognitive ability tests often yielded lower scores for minority students than for nonminority students. Because these scores were used in determining eligibility, they were considered culpable in leading to overrepresentation of minority students in special education classes. Also, as a result of the *Larry P. v. Wilson Riles et al.* case, school districts in California were prohibited from using intelligence test scores in the assessment of African American students referred for special education services. Several cases followed (*PASE [Parents in Action Special Education] et al. v. Hannon et al.* in 1980 and *Marshall v. Georgia* in 1984), and the results have varied from case to case. These cases, and the resulting publicity generated by them, have positively influenced test users to develop the skills, expertise, and competencies required to assess diverse students (Gopaul-McNicol & Thomas-Presswood, 1998) and test authors to develop better means for assessing diverse students (McCallum, 2003). In particular, test users and authors began to scrutinize the potential sources of cultural and language bias within the tests and to develop tests that are less vulnerable to these criticisms. Typically, authors have become increasingly willing to evaluate their tests using empirical and statistical methods to ensure that they are as free from bias as possible, a practice endorsed by Jensen as early as 1980. As is apparent from this section, several specific statistical analyses can be used to assess the extent to which a test is biased against a particular group. Table 4.8 provides comparisons of the nonverbal tests reviewed primarily on the basis of the data reported in the respective test manuals. The table also includes rational or theoretical arguments used by authors to rule out bias in the use of their tests.

Fairness related to minority–nonminority group comparisons. In the early 1970s consumers of tests realized minority students often obtained lower IQs than nonminority students, as mentioned in the preceding section, and observed that this difference must mean that the tests were biased against minority students. Jensen (1980) and Reynolds (1982) addressed (and rebutted) this assertion. Jensen characterized the assumption that all groups should yield the same score on intelligence tests as the *egalitarian fallacy*; he noted that no a priori reason exists to

assume that all groups should be the same on any variable, including intelligence. Jensen and others such as Reynolds reviewed the existing empirical and statistical literature and concluded, as did Ackerman (1992), that mean score differences across groups are not a valid indication of bias, assuming that the test is detecting differences solely related to the constructs under scrutiny. Nonetheless, when a test yields mean-difference scores and the lower scores characterize a minority population or populations, the test becomes suspect, and the burden of proof shifts to the test authors to demonstrate that these differences do not represent bias.

Table 4.8 provides the extent to which the nonverbal test manuals describe mean-difference comparisons. Mean-difference data, using matched controls, are presented for the NNAT–1, UNIT, and WNV; the WNV shows matched comparisons with clinical (but not minority) groups. The CTONI–2, Leiter–R, and TONI–4 also include mean-difference comparison, but without the use of matched controls. The GAMA and SB5 manuals showed no group comparisons. Braden and Athanasiou (2005) mentioned the controversial use of mean-difference comparisons (also see R. T. Brown, Reynolds, & Whitaker, 1999) and noted that they may lead to misunderstanding of the data, in part because within-group variability is typically greater than between-group variability and because many variables can contribute to the differences, some of which may not be related to group membership. They recommended the inclusion of correlations among variables used for statistical control and between these variables and the tests scores. None of the reviewed tests provided such data.

It is important to note that the magnitude of the minority versus nonminority mean differences seems to be decreasing over time, based on some comparisons between African Americans and European Americans and between Hispanic Americans and European Americans (Bracken & McCallum, 1998; Kaufman & Kaufman, 1983). By using the methods mentioned earlier in this chapter to reduce potential sources of bias (e.g., expert panels, including more representative samples in standardization), authors have created fair(er) tests.

Fairness related to item characteristics. Another strategy to ensure fair(er) test development focuses on the molecular examination of item performance as a function of group membership. Presumably, nonbiased items should mean the same thing to members of one group as to another group when overall performance on the test is controlled. So, examinees of similar overall ability in one group should have the same probability of passing an item as examinees of similar ability in a second group. The CTONI–2, Leiter–R, NNAT–1 SB5, TONI–4, and UNIT describe some analyses of item functioning by group, typically either item response theory–based analyses or the Mantel–Haenszel procedure (see Table 4.8).

Fairness related to reliability. Ideally, a test would show similar reliability indices for all examinees. That is, reliability estimates for minority groups should be similar to those for nonminority groups. Manuals for the CTONI–2, SB5, TONI–4, and UNIT provided internal consistency data for two or more minority groups, in addition to the standardization sample The WNV provides internal consistency indices for a wide variety of clinical groups, but not for minority groups. The GAMA, Leiter–R, and NNAT–1 did not report data in this category. For the most part, reliabilities are similar across minority and nonminority groups (see Table 4.8).

Fairness related to internal structure. The internal structure of a test should be about the same for diverse examinees as for mainstream examinees, that is, the constructs measured by the test should be the same or very similar (Reynolds & Lowe, 2009). As a rule, factor-analytic and related model-testing analyses are used to determine similarity across groups. The UNIT provides confirmatory factor-analytic data for two minority samples. None of the other tests provided this level of analysis, although the SB5 provided subtest correlation matrices and related chi-square statistics testing whether the pattern of correlations was consistent across groups. These data showed consistency of patterns, none of which were statistically significant. Data from the UNIT and SB5 are assumed to reflect lack of construct validity bias for the minority samples examined.

Fairness related to consequences of testing. Fairness related to the consequences of testing is critical if a test is to be used for multiple groups, minority and nonminority, and for predicting real-world outcomes. In 1990, Reynolds and Kaiser noted that tests should predict outcomes for various subgroups with a consistent degree of accuracy, essentially meaning that error in prediction should not differ as a function of group membership. Stated in a more general form, claims for test outcomes should be empirically determined and the evidence weighted. For example, if a test purports to predict school achievement, one of the earliest raisons d'être for intelligence tests, and it is intended for use with diverse examinees, it should predict achievement as well and with no more error for minority examinees than for mainstream examinees. Typically, the power of the prediction is determined by examining coefficients of correlations, and the error is determined by examining the standard error of the prediction estimates. The Leiter–R, SB5, and UNIT provide perhaps the most rigorous data, showing equivalence of regression slopes for certain races and ethnic groups, with similar regression slopes found across groups. Correlational data are reported as well for the CTONI–2 and TONI–4. The WNV reported correlations for clinical samples, but not for minority–nonminority groups. In general, these data showed little or no evidence of bias associated with use of these instrument (see Table 4.8). However, much more research is needed to establish consequences-of-testing data for these instruments, and the literature providing such evidence will grow slowly and with test use or scrutiny (see Bracken & McCallum, 2009; Hammill & Pearson, 2009; Naglieri & Brunnert, 2009; Roid, Pomplum, & Martin, 2009; Roid & Tippin, 2009)

CONCLUSIONS

The state of the art or science of nonverbal intelligence testing is sound, in spite of certain inherent limitations (e.g., the somewhat limited breadth of nonverbal assessment). Current nonverbal tests represent significant improvements over their predecessors. For the most part, the authors of current tests report evidence of strong psychometric and technical properties, reasonable (and sometimes sophisticated) theoretical grounding, representative standardization samples, and empirical and theoretical support for fair(er) and (less) biased use relative to their older counterparts. Experts continue to provide strategies for more informed use of nonverbal and related measures for multicultural, bilingual, and language-impaired populations (Bracken, & McCallum, 2001; Flanagan et al., 2007; Suzuki, Ponterotto, & Meller, 2001; Wilhoit & McCallum, 2003).

Nonverbal tests can be used as stand-alone batteries, as part of a multitest battery, or in an XBA interpretative strategy, and ample guidelines are available advising examiners of the characteristics of nonverbal tests on a number of dimensions. Currently, examiners have explicit guidelines for choosing nonverbal tests wisely, depending on the examinees' particular characteristics and needs and the goals of assessment. For example, examiners can now operationalize the extent to which linguistic or cultural demands affect the test chosen for an assessment and the extent to which an examinee is culturally or linguistically limited, and they can vary the assessment depending on the extent to which these demands should be minimized. Examiners have more options than ever, and they can choose tests on the basis of which subconstructs of intelligence should be assessed and can take advantage of the XBA approach.

Authors of nonverbal tests continue to refine existing tests and increase the capacity of new tests to assess a broader range of abilities. For example, the UNIT is now under revision and restandardization. The UNIT 2 will not only assess memory and reasoning within a symbolic and nonsymbolic context, as before, but will also provide operationalizations of an additional construct: quantitative reasoning. Quantitative reasoning can be assessed either symbolically or nonsymbolically. Examiners who need to minimize the verbal content of a cognitive assessment but who value multidimensional theoretically based operationalizations, which not only produce strong predictions of academic success but have implications for guided interventions, can choose from a growing collection of psychometrically strong nonverbal tests.

Although the state of the art and science of nonverbal assessment is stronger today than ever, controversies and problems still need to be resolved. Not all experts agree that nonverbal intelligence tests measure general intelligence, period, as opposed to some subconstruct called nonverbal intelligence, despite the literature showing limited support for a verbal–nonverbal ability dichotomy (e.g., Wasserman & Lawhorn, 2003). Others have evaluated the available data differently (e.g., Rourke, 2000), and there is no well-defined methodology to address this question (Braden & Athanasiou, 2005). Similarly, no well-defined methodology exists for determining which cognitive processes are engaged while solving problems nonverbally, as opposed to solutions based on overt verbal mediation. The response process research is limited, but with the increased accessibility and sensitivity of brain imaging capabilities, answers to these questions should become more accessible. Researchers may use the familiar functional magnetic resonance imaging process to address these questions and also now the related but newer mechanism called diffusion tensor imaging (Wilkie, 2009), which maps white matter connectivity tissue using a water-sensing technique.

Despite the limitations of nonverbal assessment, in many cases it provides an advantage over verbal testing. Nonverbal tests can corroborate or disconfirm language-loaded test results in some instances, which may help in differential diagnosis of certain kinds of presenting problems. For example, in some cases it is difficult to determine sources of observed limitations of examinees from other cultures and those with speech, language, or linguistic deficits; hearing loss; and certain kinds of psychopathology. Use of nonverbal tests will likely continue to grow in use because of their utility for assessing such individuals, particularly as the United States' linguistic and cultural diversity increases.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91. doi:10.1111/j.1745-3984.1992.tb00368.x

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychological Association. (1991). *Guidelines for providers of psychological services to ethnic, linguistic, and culturally diverse populations.* Washington, DC: Author.

Arthur, G. (1943). *A Point Scale of Performance Tests: Clinical manual.* New York, NY: Commonwealth Fund. doi:10.1037/11315-000

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 5*, 313–326. doi:10.1177/073428298700500402

Bracken, B. A. (1999). Assessing diverse populations with nonverbal tests of intelligence. *Trainers' Forum, 17*(1), 6–8.

Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test.* Itasca, IL: Riverside.

Bracken, B. A., & McCallum, R. S. (2001). Assessing intelligence in a population that speaks over two hundred languages. In L. A. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (2nd ed., pp. 405–431). New York, NY: Jossey-Bass.

Bracken, B. A., & McCallum, R. S. (2009). Universal Nonverbal Intelligence Test (UNIT). In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (pp. 291–313). Hoboken, NJ: Wiley.

Braden, J. P., & Athanasiou, M. S. (2005). A comparative review of nonverbal measures of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 557–577). New York, NY: Guilford Press.

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (2010). *Tests of Nonverbal Intelligence* (4th ed.). Los Angeles, CA: Western Psychological Services.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since *Bias in Mental Testing. School Psychology Quarterly, 14*, 208–238. doi:10.1037/h0089007

Burgemeister, B., Blum, H., & Lorge, I. (1972). *The Columbia Mental Maturity Scale.* San Antonio, TX: Psychological Corporation.

Carrey, N. J. (1995). Itard's 1828 memoire on "Mutism caused by a lesion of the intellectual functions": A historical analysis. *Journal of the American Academy of Child and Adolescent Psychiatry, 34*, 1655–1661. doi:10.1097/00004583-199512000-00016

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511571312

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic, interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism, 19*, 121–129.

Cummins, J. (1984). *Bilingualism and special education.* Clevedon, England: Multilingual Matters.

Education for All Handicapped Children Act of 1975, Pub. L. No. 94–142, 20 U.S.C. § 1400 *et seq.*.

Ehrman, M. E. (1996). *Understanding second language learning difficulties.* Thousand Oaks, CA: Sage.

El Nasser, H., & Overberg, P. (2010, June 11). Diversity grows as majority dwindles: Minorities make up almost half of births. *USA Today*, p. 1A.

Figueroa, R. A. (1990). Best practices in the assessment of bilingual children. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (pp. 93–106). Washington, DC: American Psychological Association.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). *Essentials of cross-battery assessment* (2nd ed.). Hoboken, NJ: Wiley.

Gardner, H. (1999). *Intelligence reframed.* New York, NY: Basic Books.

Goodenough, F. L. (1926). *Measurement of intelligence by drawings.* New York, NY: World.

Gopaul-McNicol, S., & Thomas-Presswood, T. (1998). *Working with linguistic and culturally-different children: Innovative clinical and educational approaches.* Boston, MA: Allyn & Bacon.

Greenfield, P. M. (1998). The cultural evolution of IQ. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 81–123). Washington, DC: American Psychological Association.

Gustafsson, J. E. (1988). Hierarchical models of individual difference in cognitive abilities. In R. J. Sternberg (Ed.), *Psychology of human intelligence* (Vol. 4, pp. 35–71). Hillsdale, NJ: Erlbaum.

Hammill, D. D., & Pearson, N. A. (2009). Comprehensive test of nonverbal intelligence. In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (2nd ed., pp. 233–264). Hoboken, NJ: Wiley.

Hammill, D. D., Pearson, N. A., & Wiederholt, L. (2009). *Comprehensive Test of Nonverbal Intelligence* (2nd ed.). San Antonio, TX: Pearson Assessments.

Hardy-Braz, S. T. (2003, April). *Enhancing school-based psychological services: Assessments and interventions with students who are deaf or hard of hearing.* Workshop presented at the meeting of the National Association of School Psychologists, Toronto, Ontario, Canada.

Harris, A. M., Reynolds, M. A., & Koegel, H. M. (1996). Nonverbal assessment: Multicultural perspectives. In L. A. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (pp. 223–252). San Francisco, CA: Jossey-Bass.

Hoffman, B. (1962). *The tyranny of testing.* New York, NY: Crowell-Collier.

Horn, J. L., & Cattell, R. B. (1966). Refinement of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*, 253–270. doi:10.1037/h0023816

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf–Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). New York, NY: Guilford Press.

Individuals With Disabilities Education Improvement Act of 2004, Pub. L. 108–446, 20 U.S.C. § 1400 *et seq.*

Jackson, C. D. (1975). On the report of the Ad Hoc Committee on Educational Uses of Tests With Disadvantaged Students: Another psychological view from the Association of Black Psychologists. *American Psychologists, 30*, 86–90.

Jensen, A. R. (1974). Cumulative deficit: A testable hypothesis. *Developmental Psychology, 10*, 996–1019. doi:10.1037/h0037246

Jensen, A. R. (1980). *Bias in mental testing.* New York, NY: Free Press.

Jensen, A. R. (1984). Test validity: *g* versus the specificity doctrine. *Journal of Social and Biological Structures, 7*, 93–118. doi:10.1016/S0140-1750(84)80001-9

Kamphaus, R. W. (2001). *Clinical assessment of child and adolescent intelligence* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children (K-ABC).* Circle Pines, MN: American Guidance Service.

Knox, H. A. (1914). A scale based on the work at Ellis Island for estimating mental defect. *JAMA, 62*, 741–747. doi:10.1001/jama.1914.02560350001001

Larry P. et al. v. Wilson Riles et al. No. C 71 2270, slip op. (N.D. Calif. 1979).

Leiter, R. G. (1948). *International Performance Scale.* Chicago, IL: Stoelting.

Lopez, E. C. (1997). The cognitive assessment of limited English proficient and bilingual children. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 503–516). New York, NY: Guilford Press.

Marshall v. Georgia, No. CV 482–233 (S. D. Ga. 1984).

Mather, N., & Jaffe, L. E. (2002). *Woodcock–Johnson III: Reports, recommendations, and strategies.* New York, NY: Wiley.

McCallum, R. S. (Ed.). (2003). *Handbook of nonverbal assessment.* New York, NY: Plenum Press. doi:10.1007/978-1-4615-0153-4

McCallum, R. S., Bracken, B. A., & Wasserman, J. (2001). *Essentials of nonverbal assessment* (A. S. Kaufman & N. L. Kaufman, Series Eds.). New York, NY: Wiley.

McGrew. K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf–Gc cross-battery assessment.* Boston, MA: Allyn & Bacon.

Mercer, J. R. (1979). *System of Multicultural Pluralistic Assessment (SOMPA): Conceptual and technical manual.* San Antonio, TX: Psychological Corporation.

Naglieri, J. A. (2003). *Naglieri Nonverbal Ability Test.* New York, NY: Harcourt.

Naglieri, J. A., & Bardos, A. N. (1997). *The General Ability Measure of Adults.* New York, NY: Pearson Assessments.

Naglieri, J. A., & Brunnert, K. (2009). Wechsler Nonverbal Scale of Ability (WNV). In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (pp. 315–338). Hoboken, NJ: Wiley.

National Institutes of Health. (2010a). *Hearing disorders and deafness.* Retrieved from http://www.nlm.nih.gov/medlineplus/hearingdisordersanddeafness.html

National Institutes of Health. (2010b). *Speech and communication disorders.* Retrieved from http://www.nlm.nih.gov/medlineplus/speechandcommunicationdisorders.html

Nieves-Brull, A. I., Ortiz, S. O., Flanagan, D. P., & Chaplin, W. F. (2006). *Evaluation of the Culture–Language Matrix: A validation of test performance in monolingual English speaking and bilingual English/Spanish speaking populations.* Unpublished doctoral dissertation, St. John's University, Jamaica, NY.

Ortiz, S. O. (2002). Best practices in nondiscriminatory assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 1321–1336). Bethesda, MD: National Association of School Psychologists.

Ortiz, S. O., & Ochoa, S. H. (2005). Advances in cognitive assessment of culturally and linguistically diverse individuals. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 234–250). New York, NY: Guilford Press.

Padilla, A. M. (2001). Issues in culturally appropriate assessment. In L. A. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment* (2nd ed., pp. 5–27). San Francisco, CA: Jossey-Bass.

PASE (Parents in Action Special Education) et al. v. Hannon et al. No. 74C 3586, slip op. (N. E. Ill., 1980).

Petti, V. L., Voelker, S. L., Shore, D. L., & Hayman-Abello, S. E. (2003). Perception of nonverbal emotion cues by children with nonverbal learning disabilities. *Journal of Developmental and Physical Disabilities, 15*, 23–36.

Porteus, S. D. (1915). Mental tests for the feebleminded: A new series. *Journal of Psycho-Asthenias, 19*, 200–213.

Raven, J. C., Raven, J. E., & Court, J. H. (1998). *Progressive matrices.* Oxford, England: Oxford Psychologists Press.

Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 545–576). New York, NY: Plenum Press.

Reynolds, C. R., & Kaiser, S. (1990). Test bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (2nd ed., pp. 487–525). New York, NY: Wiley.

Reynolds, C. R., & Lowe, P. A. (2009). The problem of bias in psychological assessment. In T. Gutkin & C. Reynolds (Eds.), *The handbook of school psychology* (4th ed., pp. 332–374). Hoboken, NJ: Wiley.

Roach, A. T., & Elliott, S. N. (2006). The influence of access to the general education curriculum on the alternate assessment performance of students with significant cognitive disabilities. *Education Evaluation and Policy Analysis, 28*, 181–194. doi:10.3102/01623737028002181

Roid, G. H. (2003). *Stanford–Binet Intelligence Scales* (5th ed.). Itasca, IL: Riverside.

Roid, G. H., & Miller, L. J. (1997). *Leiter International Performance Scale—Revised.* Wooddale, IL: Stoelting.

Roid, G. H., Pomplum, M., & Martin, J. F. (2009). Nonverbal intellectual and cognitive assessment with the Leiter International Performance Scale—Revised (Leiter–R). In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (pp. 265–290). Hoboken, NJ: Wiley.

Roid, G. H., & Tippin, S. M. (2009). Assessment of intellectual strengths and weaknesses with the Stanford–Binet Intelligence Scales—Fifth Edition (SB5). In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (pp. 127–152). Hoboken, NJ: Wiley.

Rourke, B. P. (1991). Human neuropsychology in the 1990s. *Archives of Clinical Neuropsychology, 6*, 1–14.

Rourke, B. P. (2000). Neuropsychological and psychosocial subtyping: A review of investigations within the University of Windsor laboratory. *Canadian Psychology/Psychologie Canadienne, 41*, 34–51. doi:10.1037/h0086856

Rourke, B. P., & Conway, J. A. (1997). Disabilities of arithmetic and mathematical reasoning: Perspective from neurology and neuropsychology. *Journal of Learning Disabilities, 30*, 34–46. doi:10.1177/002221949703000103

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment* (9th ed.). New York, NY: Houghton Mifflin.

Sattler, J. M. (2001). *Assessment of children* (4th ed.). San Diego, CA: Author.

Seguin, E. (1907). *Idiocy and its treatment by the physiological method.* New York, NY: Teachers College, Columbia University.

Suzuki, L. A., Ponterotto, J. G., & Meller, P. J. (Eds.) (2001). *Handbook of multicultural assessment.* (2nd ed.). San Francisco, CA: Jossey-Bass.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias.* Norwood, NJ: Ablex.

Wasserman, J. D., & Lawhorn, R. M. (2003). Nonverbal neuropsychological assessment. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 315–360). New York, NY: Kluwer Academic/Plenum. doi:10.1007/978-1-4615-0153-4_15

Wechsler, D. (1939). *The measurement of adult intelligence.* Baltimore, MD: Williams & Wilkins. doi:10.1037/10020-000

Wechsler, D. (1949). *Wechsler Adult Intelligence Scale for Children.* New York, NY: Psychological Corporation.

Wechsler, D., & Naglieri, J. A. (2006). *Wechsler Nonverbal Scale of Ability.* New York, NY: Pearson Assessments.

Wilhoit, B., & McCallum, R. S. (2003). Cross-battery analysis of the UNIT. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 63–86). New York, NY: Kluwer Academic/Plenum Press.

Wilkie, D. (2009). Coming soon to a scanner near you. *Monitor on Psychology, 40*(3), 45–46.

Williams, R. L. (1971). Abuses and misuses in testing Black children. *The Counseling Psychologist, 2*, 62–73. doi:10.1177/001100007100200314

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities.* Itasca, IL: Riverside.

Woodcock, R. W., Muñoz-Sandoval, A. F., McGrew, K. S., & Mather, N. (2005). *Bateria III Woodcock–Muñoz.* Itasca, IL: Riverside.

Woodcock, R. W., Muñoz-Sandoval, A. F., Ruef, M., & Alvarado, C. G. (2005). *Woodcock–Muñoz Language Survey, Revised.* Itasca, IL: Riverside.

CHAPTER 5

INDIVIDUAL ASSESSMENT OF ACADEMIC ACHIEVEMENT

Nancy Mather and Bashir Abu-Hamour

This chapter focuses on the individualized academic assessment of reading, written language, and mathematics for students with special needs enrolled in preschool to postsecondary settings. The chapter is organized into six parts. The first part focuses on individually administered standardized academic assessments and includes a brief discussion of administration, types of scores, and test formats and demands. The second presents an overview of curriculum-based measures and assessments. The third reviews special considerations and concerns regarding the assessment of certain types of students (e.g., students with specific learning disabilities). The remaining three parts address the assessment of reading, written language, and mathematics.

Individualized academic assessment is used within school settings for a variety of different purposes, including determining present academic performance levels, identifying academic strengths and weaknesses, comparing performance to age or grade peers, comparing intelligence test scores to achievement, investigating the need for special services, monitoring educational progress across the school years, and assisting with instructional planning (Mather & Wendling, 2009). Within both public and private school settings, academic assessment is a key component of special education eligibility and services, in terms of both identifying students with disabilities and monitoring their academic progress throughout the years. In some situations, such as a reevaluation for special education services, academic testing alone may be deemed to be sufficient. In other situations, academic testing is one important component of a comprehensive evaluation that includes an assessment of intellectual and cognitive processes, academic performance, social–emotional functioning, and consideration of the environmental factors affecting performance (Kavale, Kaufman, Naglieri, & Hale, 2005).

Comprehensive evaluations are often conducted when a student is first referred for an evaluation because more in-depth information is needed as well as an explanation of the reasons why the student is struggling in school. The evaluator integrates information from multiple sources of data, including observations, interviews, and a variety of instruments to support or clarify the diagnostic hypotheses (Kaufman, 1994; Kaufman, Lichtenberger, & Naglieri, 1999; Lichtenberger, Mather, Kaufman, & Kaufman, 2004). In explaining the findings, the evaluator addresses the referral questions and provides clinical observations of test-taking behaviors, detailed error analyses, and an analysis of strengths and weaknesses (Kaufman, 1994). Then, on the basis of the specific diagnostic findings, the evaluator recommends specific interventions and accommodations to address the identified areas of need. Academic assessment involves the use of both individual standardized measures of achievement and informal or formative measures of academic

We thank June E. Downing, professor emerita, California State University, Northridge, for assistance with information regarding the assessment of individuals with intellectual disabilities and Lynne E. Jaffe, adjunct professor, Department of Disability and Psychoeducational Studies, University of Arizona, Tucson, for assistance with the assessment of individuals with sensory impairments.

DOI: 10.1037/14049-005
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

competence, including curriculum-based measurements (CBMs) and curriculum-based assessments (CBAs) that provide measures of progress over time.

STANDARDIZED ACADEMIC ASSESSMENTS

Broad-based individualized tests of academic achievement are often used to get overall estimates of how well a student is performing in reading, writing, and mathematics as well as on a narrower subset of skills, such as word identification, spelling, or math problem solving. Commonly used individual achievement tests include subtests that assess a variety of specific skills and abilities. Many of these standardized achievement tests include similar measures. For example, the Woodcock–Johnson III Tests of Achievement (WJ III ACH; Woodcock, McGrew, & Mather, 2007), the Kaufman Test of Educational Achievement II (Kaufman & Kaufman, 2004b), and the Wechsler Individual Achievement Test—III (Wechsler Individual Achievement Test—III; Psychological Corporation, 2009) all include measures of nonword reading, word reading, spelling, math calculation, and math problem solving. Different batteries, however, provide greater emphasis on certain aspects of achievement as well as different formats and ways of assessing these skills. Table 5.1 lists examples of several widely used measures of academic performance. More in-depth academic testing is often done when a person exhibits markedly poor performance in a specific area, such as low reading or math performance.

Administration of all norm-referenced tests requires standardized practices and procedures, including factors such as ensuring a quiet, distraction-free test environment; establishing rapport with the student; having all of the necessary testing materials; and striving for a brisk but accurate administration. Examiners are expected to follow the standardized procedures for all tests they are using and must guard against applying their own

TABLE 5.1

Commonly Used Individual Achievement Tests

Test name	Age range	Abilities
Basic Achievement Skills Inventory (Bardos, 2004)	8–80 years	Reading, written language, and math skills
Diagnostic Achievement Battery, 3rd ed. (Newcomer, 2001)	6 years, 0 months–14 years, 11 months	Listening, speaking, reading, writing, and mathematics
Kaufman Test of Educational Achievement, 2nd ed. (Kaufman & Kaufman, 2004a)	4 years, 6 months–90+ years	Reading, math, written language, and oral language
Peabody Individual Achievement Test—Revised—Normative Update (Markwardt, 1997)	5 years, 0 months–22 years, 11 months	Reading, mathematics, written language, and general information
Process Assessment of the Learner: Diagnostics for Reading and Writing, 2nd ed. (Berninger, 2007b)	Kindergarten–6th grade	Diagnostics for reading and writing
Wechsler Individual Achievement Test, 3rd ed. (Psychological Corporation, 2009)	4 years, 0 months–19 years, 11 months	Oral language, reading, written expression, and mathematics
Wide Range Achievement Test, 4th ed. (Jastak & Jastak, 2005)	5 years, 0 months–94 years	Reading, spelling, and mathematics
Woodcock–Johnson III Tests of Achievement—3rd ed.—Normative Update (Woodcock, McGrew, & Mather, 2007)	2 years, 0 months–90+ years	Oral expression, listening comprehension, written expression, basic reading skills, reading comprehension, math calculation skills, and math reasoning
Young Children's Achievement Test (Hresko et al., 2000)	4 years, 0 months–7 years, 11 months	General information, reading, mathematics, writing, and spoken language

rules or rules that they have learned from the use of another test (e.g., different tests have different types of basal [starting points] and ceiling [stopping points] rules).

Types of Scores

To determine the derived scores, some standardized achievement tests provide both age and grade norms. Some have age norms that are reported for each month (e.g., from ages 2 years, 0 months, through 18 years, 11 months) and grade norms for each 10th of the year (e.g., from the beginning of kindergarten through 18.0, the median score of graduate students starting a second year of graduate school), whereas others provide comparisons at only two points in the school year, such as having fall and spring norms. Clearly, a continuous-year procedure that provides month-by-month norms increases the precision of the scores.

Individually administered achievement tests provide several types of derived scores that include raw scores (the number correct), age equivalents, grade equivalents, relative proficiency indexes (unique to the Woodcock Reading Mastery Test—Revised—Normative Update [Woodcock, 1998] and the WJ III ACH), percentile ranks, and standard scores. Although most of these scores are familiar to psychologists and diagnosticians, they are not to most parents or teachers. Some scores, however, such as the relative proficiency index, may be unfamiliar to test users and therefore require additional explanation. The score information from individually administered standardized achievement tests can be grouped into an interpretive framework with four hierarchical levels of information (Woodcock & Johnson, 1989). Each level provides different types of information about the examinee's performance, and information from one level is not interchangeable with information from another. Table 5.2 provides a brief explanation of the four levels of interpretive information and an example of the different types of scores.

A skilled evaluator integrates information from these four levels, as well as from other sources, to develop a more complete understanding of the student's academic performance. When making

TABLE 5.2

Types of Test Scores

Type of information	Score and description	Example
Qualitative level (criterion-referenced)	Error analysis, behavioral observations, and informal data	On the Nonword Reading test, Kasey had difficulty pronouncing words that began with a consonant blend.
Developmental level (norm-referenced)	Age equivalent	Kasey's reading age equivalent was 6 years, 4 months.
Instructional level (norm-referenced)	Grade equivalent	Kasey's reading grade equivalent was first grade, second month.
Proficiency (criterion-referenced)	Relative proficiency index	Kasey's relative proficiency index of 4/90 on the Spelling test indicates that when average grade mates are having 90% success when spelling, Kasey will have approximately 4% success.
	Instructional zone	His grade scores on the instructional zone indicate that an easy level of reading for Kasey is mid-first grade, whereas a frustration level is beginning second grade.
Relative group standing (norm-referenced)	Standard score	Kasey obtained a standard score of 66. His performance compared with average age peers is very low.
	Percentile rank	When Kasey's score is compared with his age peers, only 1 of 100 people would obtain a score as low or lower on the Word Identification test (percentile rank = 1).

high-stakes decisions, such as deciding whether a student is eligible for special education, evaluators often use cluster or factor scores that combine two or more tests, because of the higher reliabilities and broader coverage of the construct. The narrow abilities measured by the individual tests, however, often reveal the student's specific strengths and weaknesses. If an examiner only looks at performance on the broad clusters or factor scores that are composed of numerous abilities, the specific nature of the problem may be overlooked, such as a problem only in spelling or in reading rate. Thus, when planning an appropriate instructional program, information from the clusters or factors, from the individual tests, and even at the item level is useful for interpreting and explaining performance (Mather & Wendling, 2009).

Test Formats and Demands

Different batteries provide greater emphasis on certain aspects of achievement as well as different formats and ways of assessing these skills. As an example, the WJ III ACH has one measure of math fluency (a combination of addition, subtraction, and multiplication), whereas the Wechsler Individual Achievement Test—III has three separate measures (addition, subtraction, and multiplication); the Wechsler Individual Achievement Test—III includes both sentence and essay writing, whereas the WJ III ACH measures various aspects of sentence construction as well as writing fluency. Thus, the assessment results and findings may differ depending on the choice of the test battery as well as the format and task demands of the specific subtests. For example, timed tests require more sustained attention than tests that measure the accuracy of performance alone. Math problem-solving tests place more of a demand on oral language and reasoning abilities than do math computation tests. Thus, achievement tests are often more factorially complex than their names suggest; reading comprehension involves language comprehension, nonsense word reading requires phonological awareness, and math problem solving involves reasoning and language development (McGrew, 1997).

In addition, although tests are designed to measure certain academic abilities, a student's performance can be influenced by other factors. For example, a student's score on a timed test may be affected by attention, anxiety, cognitive response style (reflective or impulsive), perfectionistic tendencies, or visual–motor coordination, if marking or writing is involved (McGrew, 1994). When assessing written expression, an evaluator should also consider the various constraints affecting writing, such as limited instruction, specific cognitive or linguistic weaknesses, limited cultural experiences, and poor motivation because these factors can help inform the type and extent of accommodations and instruction needed (Berninger, 1996). Thus, as with the interpretation of intelligence tests, competent interpretation of achievement tests requires careful consideration of multiple influences and explanations of performance (McGrew, 1997).

CURRICULUM BASED

Some researchers have expressed the viewpoint that standardized norm-referenced tests are not particularly useful as a basis for making instructional decisions (e.g., Marston, 1989; Salvia, Ysseldyke, & Bolt, 2010). Instead, they have recommended the use of two alternative assessment procedures: CBMs and CBAs, two types of formative assessments. Formative assessment is not a test per se but instead a process by which teachers use test-elicited evidence to revise instruction for students and help students adjust their own learning strategies (Popham, 2009). Both CBMs and CBAs have been used by teachers and school psychologists for more than 3 decades and have been shown to provide reliable and valid indicators of students' achievement in reading, writing, and mathematics (for reviews, see Deno, 1985; Deno, Fuchs, Marston, & Shin, 2001).

Within academic settings, CBMs and CBAs are useful for (a) establishing norms for screening and identifying students in the need of special education services, (b) identifying students for special education evaluation who demonstrate a low level of performance and inadequate rate of improvement, (c) monitoring student progress, and (d) planning effective instruction in the general education classroom (Stecker, Fuchs, & Fuchs, 2005). Both CBMs

and CBAs are considered to be types of authentic assessment practices that are designed to provide prevention and intervention services to students (Hoover & Mendez-Barletta, 2008).

Curriculum-Based Measurements

CBMs are widely used informal assessments tied directly to the classroom curriculum that are developed by teachers or designed as standardized benchmark and progress-monitoring measures (e. g., AIMS Web, 2008). Because these measures are brief and curriculum based, they provide a good tool for directly measuring the effectiveness of instruction and intervention (Deno, 1985). CBM procedures are standardized and conducted regularly to monitor progress, adjust interventions, and measure academic gains. Instructional decisions are made using the data collected from the CBM probes with the criterion for goals and progress rates determined by comparison to a normative group (Deno et al., 2001). The National Center on Student Progress Monitoring provides materials, references, and research regarding the use and implementation of CBM procedures on its website (http://www.studentprogress.org/weblibrary.asp).

Curriculum-Based Assessments

Similar to CBMs, CBAs use brief measures of academic skills administered on a frequent basis to assess the material taught in the classroom. With CBAs, teachers use classroom-based tasks to determine student capabilities and then plan and modify instruction (Hoover & Mendez-Barletta, 2008). CBAs are designed to (a) align assessment practices with curriculum instruction; (b) permit continuing assessment of student progress; and (c) be sensitive to task variability, task demand, and the pace of instruction to ensure student success (Gickling & Rosenfield, 1995). CBAs evaluate both the environment and the learner and help determine whether an individual is receiving instruction that is appropriately challenging, effective, and delivered with fidelity. Unlike CBMs, CBAs provide data on both individual student performance and the difficulty level of the material that can then help guide intervention (Burns, Dean, & Klar, 2004).

SPECIAL CONSIDERATIONS AND CONCERNS

Before reviewing the various standardized and formative academic assessments for reading, writing, and mathematics, we discuss some issues and concerns that are relevant to the academic assessment of certain types of students. The most common referrals for academic assessments are students who are suspected of having a specific learning disability (SLD) as well as students who are suspected of having attention-deficit/hyperactivity disorder (ADHD) because these two groups often exhibit depressed academic functioning compared with their peers without disabilities (Demaray, Schaefer, & Delong, 2003; Gregg, 2007; Kaufman & Kaufman, 2004a). In addition, high comorbidity exists between these two disorders. Children who come from impoverished backgrounds are also at risk for academic difficulties. Thus, some reasons for low achievement are considered to be intrinsic (e.g., SLD or ADHD), whereas others are considered to be extrinsic (e.g., poverty or limited or inadequate instruction).

In conducting academic assessments, evaluators must consider the testing conditions, the types of assessments that will be most appropriate, and how to interpret performance for students of different ages, backgrounds, abilities, and disabilities. Special considerations are often required when evaluating individuals with SLD, individuals with ADHD, individuals with oral language impairments, English language learners (ELLs), preschool children, students with sensory impairments, students who are gifted, and students with intellectual disabilities.

Individuals With Specific Learning Disability

The category of SLD encompasses a heterogeneous group of disorders that have an adverse impact on the development of some aspect of academic functioning and proficiency. In essence, the basic defining component of nearly all SLD definitions is that learning disabilities are specific disorders in one or more of the basic psychological processes involved in learning. Because development is uneven, with some abilities being more advanced than others, a discrepancy exists between a set of intact cognitive

processes and one or more disordered academic processes (Bell, McCallum, & Cox, 2003; Hale, Naglieri, Kaufman, & Kavale, 2004). The difficulty acquiring academic skills is attributed to one or more cognitive processes that mediate achievement; thus, these individuals are often described as displaying unexpected underachievement because certain areas of achievement are lower than predicted by their overall cognitive or intellectual abilities (Boada, Riddle, & Pennington, 2008; Johnson & Myklebust, 1967).

Individuals with dyslexia, the most common type of learning disability or neurobiological disorder affecting children, have their lowest scores on standardized measures of reading achievement (Shaywitz & Shaywitz, 2003). Because of their reading difficulties, these individuals may avoid tasks that involve sustained reading, resulting in further delay in and interference with the development of word recognition skill and fluency (Wiznitzer & Scheffel, 2009). Thus, an evaluator must consider the impact of the disability in reading on other areas of performance, such as the acquisition of knowledge and the development of vocabulary.

The process of identification in public schools often differs from more clinical perspectives of what constitutes SLD. In public school settings, four different types of procedures have been implemented to identify students with SLD: (a) ability–achievement discrepancy, (b) response to intervention (RtI), (c) low achievement, and (d) intraindividual variations. These procedures, which all require some type of academic testing, are discussed briefly next.

Ability–achievement discrepancies. This procedure requires a specific difference between a student's predicted performance, based on the results of an intelligence test, and his or her actual school achievement, based on measures of oral expression, listening comprehension, written expression, basic reading skills, reading fluency, reading comprehension, mathematical calculation, or mathematical problem solving. Presently, with the reauthorization of the Individuals With Disabilities Education Act, the Individuals With Disabilities Education Improvement Act of 2004, states may not require the use of a severe discrepancy, but they may permit its use, as well as other alternative research-based procedures, for determining SLD.

Response to intervention. As specified in the Individuals With Disabilities Education Improvement Act of 2004, states may now permit a process that examines whether a student responds to scientific, research-based intervention as part of an SLD evaluation procedure. RtI models often use various types of CBM procedures to monitor and document student progress. Although an in-depth discussion of the role of RtI in the SLD identification process is beyond the scope of this chapter, models that rely solely on RtI for SLD identification appear to threaten the validity of the SLD concept and may result in inaccurate identification and potential legal challenges (McKenzie, 2009).

Low achievement. Some professionals have suggested that SLD should be defined simply as low achievement and that the evaluation of cognitive processes is unnecessary. For example, Dombrowski, Kamphaus, and Reynolds (2004) proposed that learning disability should be viewed as a developmental delay that can be determined by a standard score cutoff on an achievement test, such as a standard score of 85 or lower. In this type of diagnostic model, SLD becomes synonymous with underachievement and developmental delay rather than with a disorder in psychological processes that affects academic development. Sole reliance on a standard cutoff score of 85 on an achievement test has the potential to significantly increase the number of students identified as having SLD but not increasing the validity of the diagnoses (Mather & Gregg, 2006).

Intraindividual variations. One key characteristic of individuals with SLD is that a pattern of strengths and weaknesses exists among a person's abilities. For an individual with SLD, academic performance is often affected in some specific areas but not in others. This concept of specificity is not new. For example, Travis (1935) described students with a special disability in which a striking disparity exists between achievement in one area and achievement in another. Examples would include a student who cannot read but who can comprehend material read aloud, or

a student who excels in reading and writing but struggles with mathematics. Thus, examining intra-individual variations among a person's achievement scores can provide a piece of confirmatory evidence for the diagnosis of SLD. What often distinguishes individuals with a specific reading disability (dyslexia) from other poor readers is that their listening comprehension ability is significantly higher than their abilities to decode words and comprehend what they read (Aaron, Joshi, Palmer, Smith, & Kirby, 2002; Rack, Snowling, & Olson, 1992). Regardless of what type of model or combination of models of SLD identification are adopted (e.g., ability–achievement discrepancy, RtI, low achievement, pattern of strengths and weaknesses), some type of achievement testing will play a central role.

Individuals With Attention-Deficit/Hyperactivity Disorder

As with SLD, an assessment for ADHD must be accompanied by a thorough developmental, behavioral, and emotional evaluation (Goldstein & Cunningham, 2009). For students with ADHD who are taking medication, academic testing should occur while they are on their medication so that the results provide a more accurate assessment of their actual competence, as opposed to their distractibility or impulsivity (Lichtenberger, Sotelo-Dynega, & Kaufman, 2009). Although poor academic performance is associated with ADHD, not all individuals with ADHD experience academic difficulties (Frazier, Youngstrom, Glutting, & Watkins, 2007). Students with the most severe ADHD and the most impaired achievement, however, are the least likely to complete high school and pursue postsecondary education, and if they do attend college, they rarely complete their degrees (Frazier et al., 2007; Weiss & Hechtman, 1993). Barkley (2006) found that adolescents with hyperactivity were 3 times more likely to have failed a grade and 8 times more likely to have been expelled or dropped out of school. Unfortunately, students with ADHD have a long-standing history of lacking academic persistence (Goldstein, 1997).

With an individual with ADHD, a pattern of strengths and weaknesses may also be present, but often the weaknesses are on academic tasks that require the most sustained attention. Unlike with SLD, the impaired functioning and academic achievement problems can be partially explained by poor self-regulation as well as by disruptive, impulsive, and inattentive behaviors. Virtually every student with ADHD has trouble completing homework, studying, taking tests, organizing materials, and listening in class; furthermore, many individuals with ADHD also have a SLD in reading, writing, or mathematics that further complicates their school difficulties (Robin, 2006). As with students with dyslexia, students with ADHD appear to obtain their lowest scores on standardized tests of reading achievement (Frazier et al., 2007). The presenting academic problems may, however, vary as a function of age, educational setting, or both. For example, among university students with SLD or ADHD-related problems, Cellucci, Remsperger, and McGlade (2007) found that difficulties in mathematics were more common than reading problems.

Individuals With Oral Language Impairments

Students who have difficulty understanding or using spoken language also have difficulty with aspects of reading, writing, and mathematics that require language-specific processes and involve higher order cognitive activities. Across the life span, verbal abilities and acquired knowledge have a strong and consistent relationship with reading (Evans, Floyd, McGrew, & Leforgee, 2002), written expression (McGrew & Knopik, 1993), and mathematical problem solving (Floyd, Evans, & McGrew, 2003). Thus, poor oral language abilities affect performance in reading comprehension, written language, and mathematical problem solving. In most instances, the comprehension of spoken and written language appears to be independent of word-reading ability (Aaron & Simurdak, 1991). Oral language, reading, and writing, however, all form an integrated system with reciprocity in development: Oral language provides the knowledge base for reading and writing, and what students learn from reading and writing enhances their oral language development (Lerner & Kline, 2005).

Both reading comprehension and written expression depend on background knowledge to understand and create the messages, familiarity with

sentence structures, verbal reasoning abilities, and the possession of a broad and deep vocabulary (McCardle, Scarborough, & Catts, 2001). Thus, words and the concepts they represent provide the foundation for advanced literacy (Cunningham, Stanovich, & Wilson, 1990; Perfetti, Marron, & Foltz, 1996). In some cases, students with SLD also have underlying oral language impairments that contribute to low scores on measures of oral language ability as well as low scores on measures of reading comprehension and written expression; in other cases, students with SLD have adequate to advanced verbal abilities but severe problems acquiring word recognition and spelling skills (Carlisle, 1993; Carlisle & Rice, 2002; Fletcher, Lyon, Fuchs, & Barnes, 2007).

English Language Learners

ELLs have unique assessment needs because they represent a wide range of demographic characteristics and educational experiences as well as a variety of academic, socioeconomic, cultural, linguistic, and ethnic backgrounds (Lenski, Ehlers-Zavala, Daniel, & Sun-Irminger, 2006). Historically, ELLs have lagged behind their peers in academic achievement (Hoover & Mendez-Barletta, 2008). When assessing ELLs, an evaluator should consider the student's language background, socioeconomic factors, and language use in both the home and school. Ideally, ELL assessments are conducted by qualified bilingual evaluators who (a) use assessment tools that are standardized and validated in the student's first language; (b) have oral and written language skills in English as well as the student's first language; and (c) have knowledge and understanding of the student's cultural background (Kraemer, 2010).

Often, assessment features that have been developed and designed for native English speakers can be problematic for ELLs who are unfamiliar with certain words, more complex linguistic structures, or both (Hoover & Mendez-Barletta, 2008). If a student is primarily monolingual in a language other than English, the examiner would need to administer standardized or informal assessments in the person's native language rather than attempting to translate the test. In some cases, information from academic assessments can be useful for determining how quickly students are acquiring English. If the goal is to assess a student's current academic functioning in English, then it is appropriate to test the student in English and report that the results indicate the student's current level of ability to understand and use English (Salvia et al., 2010).

As with all types of testing, the evaluation measures used must be culturally appropriate. The fundamental principle is that the evaluation materials should measure students' knowledge, skills, or abilities, not their limited ability to understand and use English (Salvia et al., 2010). Evaluators must always consider whether limited English proficiency is a factor affecting understanding of specific test items. If an achievement test is administered in a language that a student does not fully understand, the test becomes a measure of language rather than a measure of content knowledge or skill (Hoover & Mendez-Barletta, 2008). As noted by Lenski et al. (2006), "all assessments in English are also assessments *of* English" (p. 29). In interpreting the results of achievement tests, measures of vocabulary and acquired knowledge are always subject to cultural influences and experiences. Thus, ELLs may obtain low scores on content-based assessments in math, science, and social studies, even though they may actually know as much as their non-ELL peers (Hoover & Mendez-Barletta, 2008).

Before beginning an assessment, an examiner must be knowledgeable about the student's academic performance in the first language as well as issues regarding bilingualism and the process of second language acquisition. For example, the evaluator would want to know whether any differences exist in the reading systems; for example, English is read from left to right, whereas Arabic is read from right to left. In addition, some languages have more regular phoneme–grapheme correspondence and are easier to learn to read and spell than English. In English, there are five to seven vowel letters and 15 vowel sounds, whereas in Spanish there are only five vowel letters and five vowel sounds. Although the vowels in English and Spanish look the same, they represent different sounds (Klingner & Geisler, 2008). In addition, the evaluator would want to know what sounds and letters differ from English in the native language or do not exist in the native

language so as to identify specific misunderstandings and confusions that may affect reading and spelling development. For example, most Spanish dialects do not include the short vowel sound for the letter *i* or the /sh/ sound, which makes it difficult for the student who is not accustomed to hearing these sounds to distinguish them from other sounds (Klingner & Geisler, 2008). Also, when asking ELLs to summarize stories and events, they may, depending on their first language, use different organizational structures than Western European text structures; the world knowledge, vocabularies, and discourse structures of students from different cultural and socioeconomic backgrounds often differ significantly from what they encounter in school (Snyder, Caccamise, & Wise, 2005).

One difficult task for both evaluators and teachers is to distinguish between the normal effects of the second language acquisition process and the diagnosis of SLD because the characteristics associated with language acquisition can mirror SLD (Klingner & Geisler, 2008). The Individuals With Disabilities Education Improvement Act of 2004 specifies that special education services cannot be provided if a student's low academic performance can be attributed to limited English proficiency. Efforts must be made to ensure that a student's learning characteristics are interpreted appropriately to avoid misclassification of a student by confusing language acquisition with SLD; thus, CBAs and additional authentic assessments, such as portfolios that provide a collection of classroom work, are often necessary to differentiate the existence of a disability from issues related to second language learning (Abedi, 2006; Hoover & Mendez-Barletta, 2008). It is, however, important to keep in mind that although many schools take a wait-and-see approach and services are delayed or not provided, some ELLs truly do have SLD and would benefit from special education services that provide explicit, intensive academic interventions (Klingner & Geisler, 2008).

Preschool Children

For many decades, developmental researchers have focused on how to best prepare young children to meet the academic demands of school (Ryan, Fauth, & Brooks-Gunn, 2006) as well as on how to accurately identify young children who are at risk for academic difficulties. To address these concerns, several preschool academic achievement tests have been developed that focus on domain-specific areas related to the acquisition of emergent literacy and numeracy skills (Duncan et al., 2007; Konold & Pianta, 2005). Early readiness skills, such as the abilities to manipulate phonemes and recognize letters and letter sounds predict later reading achievement (Bradley & Bryant, 1983; Bryant, MacLean, Bradley, & Crossland, 1990; Duncan et al., 2007; Lonigan, Burgess, & Anthony, 2000; Wagner & Torgesen, 1987), whereas early numeracy skills, including counting, number knowledge, estimation, and number pattern facility, predict later mathematical competence (Duncan et al., 2007; Geary, 2003; Geary, Hoard, & Hamson, 1999; Jordan, Kaplan, Ola'h, & Locuniak, 2006). Table 5.3 presents examples of commonly used measures of early literacy.

Even though instruments are available, the accurate assessment of academic achievement in young children can present unique challenges for several reasons. Many academic skills have yet to be developed. A number of examiner, examinee, and environmental factors must also be considered. Young children are rarely accustomed to the more structured atmosphere of a standardized testing situation. In addition, their spontaneity, activity level, wariness of strangers, inconsistent performance in new environments, and other developmental characteristics can pose challenges for even the most experienced examiner. With knowledge of early childhood development, however, reliable and valid early achievement assessments may be conducted. Most achievement test manuals provide a discussion of appropriate testing procedures for preschool-age children (Ford & Dahinten, 2005).

Sensory Impairments

One difficulty when evaluating individuals with sensory impairments is that few instruments have been adapted for use with these individuals. Thus, evaluators often find that they have to pick and choose among the various subtests within a battery, make specific adaptations to the testing material, or both (e.g., enlarging the print or presenting information orally rather than with a compact disc).

TABLE 5.3

Commonly Used Standardized Measures of Early Literacy

Test name	Age range	Abilities
Assessment of Literacy and Language (Lombardino, Lieberman, & Brown, 2005)	Prekindergarten–1st grade	Emergent literacy, language, phonological awareness, and phonological–orthographic
Early Reading Assessment (Hammill, Pearson, Hresko, & Hoover, 2012)	4 years, 0 months–7 years, 11 months	Print knowledge, phonological awareness, and receptive vocabulary
Early Reading Diagnostic Assessment, 2nd ed. (Jordan, Kirk, & King, 2003)	Kindergarten–3rd grade	Brief vocabulary, reading comprehension, listening comprehension, phonological awareness, word reading, pseudoword decoding, rapid automatized naming, and passage fluency
Early Reading Success Indicator (Psychological Corporation, 2004)	5 years, 0 months–10 years, 0 months	Rapid automatizing naming–letters, phonological processing, speeded naming, word reading, and pseudoword decoding
Diagnostic Assessments of Reading, 2nd ed. (Roswell, Chall, Curtis, & Kearns, 2005)	5 years, 0 months–adult	Print awareness, letters and sounds, word recognition, oral reading accuracy and fluency, silent reading comprehension, spelling, and word meaning
Dynamic Indicators of Basic Early Literacy Skills, 6th ed. (Good et al., 2003)	Kindergarten–3rd grade	Initial sound fluency, letter naming fluency, phoneme segmentation fluency, nonsense word fluency, oral reading fluency, oral retelling fluency, and word use fluency
Test of Early Reading Ability, 3rd ed. (Reid, Hresko, & Hammill, 2001)	3 years, 6 months–8 years, 6 months	Alphabet, conventions, and meaning
Test of Preschool Early Literacy (Lonigan, Wagner, Torgesen, & Rashotte, 2007)	3 years, 0 months–5 years, 11 months	Print knowledge, definitional vocabulary, and phonological awareness

Students with hearing impairments. When assessing an individual with a hearing impairment, the primary mode of communication is more important than the degree or type of hearing loss. The main communication modes may be grouped as follows: (a) American Sign Language, a complete visual–spatial language with its own semantics, syntax, and pragmatics; (b) manually coded English, the use of signs in English word order, sometimes including English parts of speech that do not exist in American Sign Language; (c) sign-supported speech, the use of spoken English with sign used simultaneously; and (d) aural–oral English, the use of spoken English without sign, usually aided by some form of auditory amplification.

The individual's primary communication mode will determine what test adaptations and accommodations are needed. A professional who is familiar with the student and has expertise in the different communication modes should determine the communication mode. If the individual uses any degree of signed communication, an examiner who is not fluent in sign should work with a certified interpreter. Additionally, background noise and visual distractions should be minimized, and the student's amplification (e.g., hearing aid, cochlear implant) must be checked to ensure it is turned on and working properly. When testing an individual who uses American Sign Language, the signs used for instructions and some items will depend more on the intent of the task than on the English sentences being translated. Consequently, the examiner should familiarize the interpreter with the test content before testing.

When administering audiorecorded tests to individuals with a mild hearing loss, the examiner may use an amplification system that feeds the recording directly into the subject's hearing aids. For individuals with more severe hearing losses, the examiner should administer audiorecorded tests orally, facing the individual to facilitate speech reading and attempting to imitate the recording as closely as possible. When administering phonological awareness or spelling tests, it is helpful to have the student repeat the stimulus words before responding. Then, when analyzing errors, the examiner will be able to

determine whether the error resulted from the individual's hearing loss. For example, when asked to count the sounds in a word, the student will likely count only those sounds he or she produces. When testing a student who signs, the examiner needs to be aware of any substantial change in the task demands. For example, on a word identification test, a hearing person is given credit for pronouncing the word correctly and does not have to know the meaning. Because a sign represents the meaning of a word, the same test for a person who signs is a reading vocabulary test. Also, not all English words have signs, and some are routinely finger spelled (e.g., *bank, the, air*). In these cases, the examiner will need to accept any explanation that indicates that the student understands the word. Generally, aural–oral students whose amplified hearing and speech discrimination are normal or near normal should be able to take most tests according to standardized procedures, in which case, use of the derived scores is probably valid. In other cases, the examiner must judge the validity of the scores on the basis of the number and degree of adaptations made. Kamphaus (1993) aptly described the situation as follows:

> The examiner who is unfamiliar with hearing-impaired children and the issue of hearing impairments in general may be able to get a score, perhaps even an accurate score. The central issues, however, are *interpretation* of that score and treatment plan design. An examiner with greater expertise related to the child's referral problem will simply be able to better understand the etiology, course, and treatments. It's a matter similar to seeing a psychiatrist for heart problems. While the psychiatrist can perhaps obtain relevant EKG and other test scores, I personally would feel better in the hands of a cardiologist! (p. 400)

Students with visual impairments. Visual impairment, a particularly complex and nonunitary condition with multiple causes and manifestations, is typically discussed in two categories: low vision and blindness. A common description of a person with low vision is one who has "difficulty accomplishing visual tasks, even with prescribed corrective lenses, but who can enhance his or her ability to accomplish these tasks with the use of compensatory visual strategies, low vision and other devices, and environmental modifications" (Corn & Koenig, 1996, p. 4). Before selecting tests to administer, the evaluator must learn as much as possible about the student's eye condition, how it affects the student's vision (e.g., blurred vision, nystagmus, central or peripheral vision only), how the student uses vision, any visual devices the student uses, and the accommodations typically provided. This type of information is usually provided in a recent functional vision evaluation report. For educational purposes, a student who is blind (also termed *functionally blind*) does not use vision as a major channel for learning and instead uses the auditory and tactual senses. If able, these students typically use braille as their reading medium.

Because few assessments have been developed or specifically adapted for testing students with visual impairments, evaluators often find that they have to make adaptations to the test materials. Adaptations are done in collaboration with a vision specialist who knows the student and is familiar with his or her eye condition. For students with low vision, the adaptations or accommodations required may be minimal, such as the use of a focus lamp, slant board, or both or extensive, such as retyping text in a larger sans serif font or using a closed-circuit television system. In making adaptations or providing accommodations, an examiner must carefully consider whether they might significantly alter the task demands. For example, substantially enlarging a reading passage on a closed-circuit television breaks up the flow of the text, interfering with fluency and possibly comprehension.

When selecting tests for students with low vision as well as for those who are blind, an evaluator must ensure that any adaptation or accommodation does not make the task more difficult or complicated. This is also a concern when translating a written test into braille. To date, the WJ III ACH—Braille Adaptation (Jaffe, 2009; Jaffe & Henderson, 2009; Schrank & Woodcock, 2007) is the only standardized academic achievement battery that has been specifically adapted for braille readers.

Students Who Are Gifted

Because of limited federal funding, the results of standardized group achievement testing are most often used to identify individuals with advanced academic performance in one or more domains. One problem with gifted assessment on both group and individualized measures is that some tests do not have enough "top" items, or items that are difficult enough to measure very advanced achievement accurately. In other words, because the test ceilings are too low, a student may even get all of the items correct so an accurate performance level is not determined.

In some cases, gifted students do participate in individualized academic assessments, particularly in the case of individuals who are suspected of being "twice exceptional," or having both high abilities and SLD. A student may struggle with reading but be highly competent in another academic area. Monroe (1932) observed that "the children of superior mental capacity who fail to learn to read are, of course, spectacular examples of specific reading difficulty since they have such obvious abilities in other fields" (p. 23). The instructional needs of twice exceptional, high-functioning students can be overlooked because their high abilities may allow them to compensate and, as a result, they are not recognized as being gifted or as gifted with SLD (National Education Association, 2006).

One common misunderstanding regarding the educational achievement of gifted individuals is that they have high abilities in all academic domains. Because variation exists within and among people in neurodevelopment, it is perfectly normal to have significant strengths and weaknesses within the same person; a child who struggles with reading may have spectacular gifts in math (Gilger & Hynd, 2008). Thus, evaluators need to pay attention to individual differences and create appropriate instructional goals for students who are identified as being advanced or gifted in one academic domain, but not in another.

Students With Intellectual Disabilities

When assessing students with moderate to severe intellectual disabilities, standardized assessments are not considered best practice for the following reasons: (a) They lack sensitivity to measure small gains and do not provide a positive or comprehensive picture of the true abilities of a given student and are thus not very useful for planning educational interventions (Siegel & Allinder, 2005; Taylor, 2008); (b) the students may not progress in the same developmental order as the majority of students and may have skills that, according to a normal acquisition process, they should not have (Tindal et al., 2003); (c) these types of assessments are out of the student's natural context or routine (not familiar and there is no real reason to perform or comply); (d) they depend on a student having good verbal skills (at least receptive language); and (e) they rely on the student being motivated to perform well (which may be a false assumption; Downing & Demchak, 2008; Siegel & Allinder, 2005). Instead, for evaluating academic performance, the recommended practice is to use CBAs, criterion-referenced tests (individualized for each student), or observational assessments using ecological inventories. These assessments must be designed to measure small increments of progress in natural environments in which the student knows the routine and expectations and has the supports needed to be successful (Downing, 2010; Snell, 2002).

READING

The reading process is complex, incorporating at least two facets of performance: word recognition, which includes both accuracy and rate, and comprehension (Aaron, Joshi, & Williams, 1999). Each of these areas can develop, or fail to develop, independently of each other. Thus, reading difficulties can be a result of weaknesses in word recognition skills (decoding) or weaknesses in listening and language comprehension that underlie the development of reading comprehension (Torgesen & Miller, 2009). Results from a recent meta-analysis of reading research indicated that problems in several distinct areas contribute to the differences in reading abilities among adults with and without reading disabilities (Swanson & Hsieh, 2009). An evaluator can determine whether an individual's poor reading stems from decoding weaknesses, limited vocabulary, poor reading comprehension, or an inadequate

home environment. By identifying the source of the reading difficulties, instruction can then be carefully planned to address an individual's specific needs.

Diagnostic reading tests are used to assess basic reading skills, fluency, vocabulary, and comprehension, all important aspects of reading. Many standardized measures, such as the Gray Oral Reading Test, 5th ed. (Wiederholt & Bryant, 2012), include measures of oral reading fluency and rate, whereas others include measures of comprehension based on vocabulary, sentence completion, or paragraph construction (e.g., Test of Reading Comprehension—4 [Brown, Hammill, & Wiederholt, 2008] and Woodcock Reading Mastery Test—Revised—Normative Update [Woodcock, 1998]) or measures of basic reading skills such as phonemic awareness, phonics knowledge, or irregular and regular word reading (e.g., WJ III Diagnostic Reading Battery [Woodcock, Mather, & Schrank, 2004]; Test of Irregular Word Reading Efficiency [Reynolds & Kamphaus, 2007]; Test of Word Reading Efficiency—2 [Torgesen, Wagner, & Rashotte, 2012]). Some of the measures are timed (e.g., Test of Word Reading Efficiency—and Test of Irregular Word Reading Efficiency), whereas others measure word reading accuracy (e.g., Woodcock Reading Mastery Test—Revised—Normative Update). Evaluators select the appropriate reading assessment to administer on the basis of the referral question or diagnostic hypotheses. Table 5.4 lists commonly used standardized measures of reading performance.

To assess and monitor reading progress, many teachers also use informal reading inventories in which students read aloud graded passages and answer comprehension questions. These inventories provide a system for analyzing specific reading errors. By using a series of graded passages, teachers can estimate an appropriate instructional level for the reading materials, both for word decoding and comprehension. These mastery levels often include

TABLE 5.4

Commonly Used Standardized Measures of Reading

Test name	Age range	Abilities
Gray Oral Reading Tests, 5th ed. (Wiederholt & Bryant, 2012)	6 years, 0 months–23 years, 11 months	Oral reading skills
Gray Silent Reading Tests (Wiederholt & Blalock, 2000)	7 years, 0 months–25 years, 0 months	Silent reading comprehension ability
Gray Diagnostic Reading Tests, 2nd ed. (Bryant, Wiederholt, & Bryant, 2004)	6 years, 0 months–13 years, 11 months	Decoding, comprehension, general reading, listening vocabulary, rapid naming, and phonological awareness
Nelson-Denny Reading Test (Brown, Fishco, & Hanna, 1993)	9–16 years	Vocabulary, comprehension, and reading rate
Test of Irregular Word Reading Efficiency (Reynolds & Kamphaus, 2007)	3 years, 0 months–94 years	Irregular word reading
Test of Reading Comprehension, 4th ed. (Brown, Hammill, & Wiederholt, 2008)	7 years, 0 months–17 years, 11 months	Reading comprehension, vocabulary, and contextual fluency
Test of Silent Reading Efficiency and Comprehension (Wagner, Torgesen, Rashotte, & Pearson, 2010)	1st–12th grades	Silent reading efficiency (i.e., speed and accuracy) and comprehension
Test of Silent Word Reading Fluency (Mather, Hammill, Allen, & Roberts, 2004)	6 years, 6 months–17 years, 11 months	Word recognition fluency
Test of Word Reading Efficiency—2 (Torgesen, Wagner, & Rashotte, 2012)	6 years, 0 months–24 years, 11 months	Regular and irregular word reading accuracy and fluency
Woodcock–Johnson III Diagnostic Reading Battery (Woodcock, Mather, & Schrank, 2004)	2 years, 0 months–80+ years	Reading achievement and ability
Woodcock Reading Mastery Tests—Revised—Normative Update (Woodcock, 1998)	5 years, 0 months–75+ years	Visual–auditory learning, letter and word identification, word attack, vocabulary, and passage comprehension

the independent level (easy), the instructional level (with support), and the frustration level (too difficult). Many informal reading inventories also provide measures of both reading and listening comprehension that can help document differences among these abilities.

Phonological Awareness

One important aspect of early literacy development, as well as a causal factor of reading failure, is phonological awareness. Many poor readers, whether having a reading disability or not, have weaknesses in phonology or a generalized deficit in phonological processing (Swanson, Mink, & Bocian, 1999). Although phonological awareness is an oral language ability, it is often assessed as part of a reading evaluation because of its close relationships with reading outcomes. As students learn to read and write an alphabetic language such as English, a critical first step is becoming aware that speech can be separated and sequenced into a series of discrete sounds or phonemes, the smallest units of sound. In most instances, this awareness develops gradually during the preschool and early elementary years. Phonological awareness, particularly the ability to blend phonemes together, provides the basis for the development of phonics, a method of pronouncing words through the conversion of a letter or letters (graphemes) into their corresponding phonemes. Table 5.5 presents several commonly used measures of phonological awareness.

Basic Reading Skills

Standardized assessments of reading include measures of word recognition accuracy that involve reading a list of unrelated real words as well as measures of nonword reading that involve reading nonsense words that conform to English spelling patterns (e.g., *flib*). Real-word reading allows evaluators to measure accuracy as well as the amount of print exposure. Nonword reading tasks help the evaluator assess phonic skills. If a student has poor word identification skills, an evaluator also attempts to determine the effect on the speed of both word perception and reading comprehension. It is important to try and separate the accuracy and speed of

TABLE 5.5

Commonly Used Standardized Measures of Phonological Awareness

Test name	Age range	Abilities
Comprehensive Test of Phonological Processing (Wagner, Torgesen, & Rashotte, 1999)	5 years, 0 months–24 years, 0 months	Phonological awareness (elision, blending words, sound matching), phonological memory (memory for digits, nonword repetition), and rapid naming
Lindamood Auditory Conceptualization Test, 3rd ed. (Lindamood & Lindamood, 2004)	5 years, 0 months–18 years, 11 months	Isolated phoneme patterns, tracking phonemes, counting syllables, tracking syllables, tracking syllables and phonemes
Phonemic-Awareness Skills Screening (Crumrine & Lonegan, 2000)	1st–2nd grades	Rhyming, sentence segmentation, blending, syllable segmentation, deletion, phoneme isolation, phoneme segmentation, and substitution
Phonological Awareness Literacy Screening (Pre-K; Invernizzi, Sullivan, Meier, & Swank, 2004)	3 years, 0 months–5 years, 0 months	Name writing, alphabet knowledge, and print and word awareness
Pre-Reading Inventory of Phonological Awareness (Dodd, Crosbie, McIntosh, Teitzel, & Ozanne, 2003)	4 years, 0 months–6 years, 11 months	Rhyme awareness, syllable segmentation, alliteration awareness, sound isolation, sound segmentation, and letter–sound knowledge.
Test of Phonological Awareness, 2nd ed. PLUS (Torgesen & Bryant, 2004)	5 years, 0 months–8 years, 0 months	Recognize phonemes in spoken words and the relationship between letters and phonemes
Test of Phonological Awareness in Spanish (Riccio, Imhoff, Hasbrouck, & Davis, 2004)	4 years, 0 months–10 years, 11 months	Initial sounds, final sounds, rhyming words, and deletion
Test of Phonological Awareness Skills (Newcomer & Barenbaum, 2003)	5 years, 0 months–10 years, 11 months	Rhyming, incomplete words, sound sequencing, and phoneme deletion

reading from the ability to comprehend text. Some students have developed accurate word pronunciation skills, but still read slowly (Mastropieri, Leinart, & Scruggs, 1999).

Reading Fluency

Reading fluency encompasses the speed or rate of reading as well as the ability to read materials with expression. *Fluency* has been defined as "the ability to read connected text rapidly, smoothly, effortlessly, and automatically with little conscious attention to the mechanics of reading, such as decoding" (Meyer & Felton, 1999, p. 284). The concept of *automaticity*, the key to skilled reading, refers to a student's ability to recognize words rapidly by sight with little attention required to the word's appearance (Ehri, 1998).

Standardized measures of reading fluency. A variety of formats exist for measuring reading rate, fluency, or both. Some standardized tests measure reading fluency by having the student read lists of words as quickly as possible. For example, the Test of Word Reading Efficiency—2 measures how many real words and nonsense words a student can read aloud within two 45-second periods. The Test of Silent Word Reading Fluency (Mather, Hammill, Allen, & Roberts, 2004) requires that the student read words silently for 3 minutes and place slashes between words that are presented with no spaces (e.g., *rundogtoy*). The Gray Oral Reading Test (5th ed.; Wiederholt & Bryant, 2012) measures the accuracy of reading combined with the amount of time it takes to read a story, whereas the Nelson Denny (Brown, Fishco, & Hanna, 1993) assesses reading rate within the first minute of the first passage of the Comprehension test.

Oral Reading Fluency (CBM). The most widely researched CBM measure is oral reading fluency. For this measure, students are given a reading passage at their grade or instructional level and asked to read the passage aloud for 1 minute. Their performance is then based on the number of words read correctly. The rate of reading is then used to set both short-term goals and a long-range goal to be obtained at end of the monitoring period (Fuchs & Fuchs, 1999). Using different reading passages, the teacher then collects frequent data probes to determine whether the student is making adequate progress. Results from a recent meta-analysis indicated a significant, strong overall correlation among the oral reading fluency CBM and other standardized tests of reading achievement. The differences in correlations occurred as a result of the different test and administration formats (Reschly, Busch, Betts, Deno, & Long, 2009). Table 5.6 presents samples of CBM probes for reading and written expression.

Reading Comprehension

Many reading comprehension measures, including both standardized tests and CBMs, measure only a limited type of comprehension and do not demand in-depth interpretation of text (Snyder et al., 2005). As with reading fluency, a variety of different procedures exist for measuring reading comprehension; thus, tests and scores will differ for individuals depending on how comprehension is assessed.

TABLE 5.6

Samples of Curriculum-Based Measurement Probes for Reading and Written Expression

Area	Teacher copy	Student copy
Reading fluency	It was raining outside. There 5 was nothing for Norman to do. 11	It was raining outside. There was nothing for Norman to do.
Reading comprehension (maze task)	Stuart has nice parents. They didn't embarrass him in [glad/front/yellow] of his friends.	Stuart has nice parents. They didn't embarrass him in [glad/front/yellow] of his friends.
Written expression	(same as student's)	One day, I was out sailing. A storm carried me far out to sea and wrecked my boat on a desert island.

Standardized measures of reading comprehension. Different formats involve varied task demands that measure different skills (Keenan, Betjemann, & Olson, 2008). Some tests use a cloze format in which the person must fill in a missing word in a sentence (e.g., WJ III ACH); others use a paragraph followed by specific questions (e.g., Kaufman Test of Educational Achievement II). Some tests measure silent reading; others measure oral reading.

Another important consideration is the length of the passages. Short passages are more influenced by decoding skills, whereas on longer passages, a student can reduce the effect of decoding problems through the use of background knowledge and context (Keenan et al., 2008). Thus, the shorter the passage, the greater the emphasis on accurate decoding; the longer the passage, the easier it is to use context. One problem that has been common to some tests of reading comprehension is that students can answer some of the test items through prior knowledge without having read the passage (Coleman, Lindstrom, Nelson, Lindstrom, & Gregg, 2010; Snow, 2003).

The reasons for comprehension problems also vary. For some individuals, poor comprehension results from poor word recognition; for other individuals, comprehension is more affected by limited oral language abilities and world knowledge. To more fully understand the nature of a student's comprehension problems, the evaluator must analyze and compare the student's performances on measures of wording reading, reading fluency, and listening comprehension (Snyder et al., 2005).

Curriculum-based measurement procedures. One commonly used CBM procedure for reading comprehension is to ask students to read passages and then retell or write about what they recall. Another CBM procedure that has been used to assess reading comprehension is asking students to read short selections and complete a maze task (Shin, Deno, & Espin, 2000). For the maze procedure, words are deleted from the text, and the student marks the word that belongs from the three options (see Table 5.6). Although maze procedures have been described as valid ways to assess reading comprehension, they may not measure the higher level abilities of drawing inferences and constructing connections across text (Snyder et al., 2005).

WRITTEN LANGUAGE

Writing disabilities are complex and multifaceted because writing requires the linking of language, thought, and motor skills. The writer must write legibly, spell words correctly, and translate thoughts into writing. Difficulty in any aspect of writing can contribute to difficulty in another. For example, motor difficulties may have a direct impact on handwriting and spelling performance, and then poor handwriting and spelling may have an impact on the quality and quantity of written output. Thus, writing is a highly complex, integrated task that has been described as "an immense juggling act" (Berninger & Richards, 2002, p. 173). Table 5.7 includes several widely used standardized measures of written expression and spelling.

Handwriting

Difficulties with handwriting are typically determined through an analysis of writing errors. In analyzing handwriting, an evaluator considers overall legibility, letter formation errors, and writing rate. Legibility is often best determined by attempting to read a student's papers. Letter formation errors are identified by examining words more closely to determine any problematic letters or problems with the joining of letters in cursive writing. Writing speed is often measured by asking a student to copy a short passage for 1 minute and then comparing his or her performance to those of classmates. Few standardized assessments exist for handwriting evaluation.

Basic Writing Skills

Although basic writing skills also include punctuation and capitalization rules, the main component is spelling. As with basic reading skills, students with poor spelling often show weaknesses in the phonological aspects of language (Bruck, 1993; Moats, 1995). Spelling ability is primarily related to phonemic segmentation, the ability to break apart the sounds in words, another aspect of phonological awareness. In addition, orthography, knowledge of and ability to recall spelling patterns and letters

TABLE 5.7

Commonly Used Standardized Measures of Written Expression and Spelling

Test name	Age range	Abilities
Illinois Test of Psycholinguistic Abilities, 3rd ed. (Hammill, Mather, & Roberts, 2001)	5 years, 0 months–12 years, 11 months	Oral and written language
Oral and Written Language Scales (Carrow-Woolfolk, 1995)	3 years, 0 months–21 years, 0 months	Listening comprehension, oral expression, written expression
Test of Adolescent and Adult Language, 4th ed. (Hammill, Brown, Larsen, & Wiederholt, 2007)	12 years, 0 months–24 years, 11 months	Includes subtests on written language abilities
Test of Early Written Language, 2nd ed. (Hresko, Herron, & Peak, 1996)	3 years, 0 months–10 years, 11 months	Basic writing, contextual writing, and global writing
Test of Orthographic Competence (Mather, Roberts, Hammill, & Allen, 2008)	6 years, 0 months–17 years, 11 months	Basic writing conventions, spelling, and speed of letter and word perception
Test of Written Expression (McGhee, Bryant, Larsen, & Rivera, 1995)	6 years, 6 months–14 years, 11 months	Written achievement and writing samples
Test of Written Language, 4th ed. (Hammill & Larsen, 2008)	9 years, 0 months–17 years, 0 months	Written language skills
Tests of Written Spelling, 4th ed. (Larsen, Hammill, & Moats, 1999)	6 years, 0 months–18 years, 11 months	Spelling
Word Identification and Spelling Test (Wilson & Felton, 2004)	7 years, 0 months–18 years, 11 months	Word identification, spelling, and sound–symbol knowledge
Written Language Observation Scale (Hammill & Larsen, 2009)	9 years, 0 months–17 years, 11 months	Students' daily classroom writing behaviors
Writing Process Test (Warden & Hutchinson, 1992)	8–19 years	Written products using the writing process

strings, and morphology, knowledge of the meaning units of language, are also necessary for sequencing letters and adding word parts and endings. Students with reading disabilities often have poor spelling because of fundamental weaknesses in phonological, orthographic, and morphological awareness. In addition, older students with dyslexia often have difficulty recalling the motor and orthographic patterns necessary for spelling (Gregg, 2009; Gregg, Coleman, Davis, & Chalk, 2007). Figure 5.1 illustrates the attempted spellings of irregular words (i.e., words in which a part of the word does not conform to common English spelling rules, such as the "ai" in the word *said*) by Mark, a sixth-grade student. As can be noted, Mark spells words the way they sound relying on phonology rather than on the way the word looks, which requires the use of orthography. When spelling the dictated words, Mark commented that he only knew how to spell the word *ocean* and *island* because they were on the classroom spelling test the week before and that to remember the spelling of *island,* he pronounced it as "is-land."

Standardized tests. Most standardized tests include a separate measure of spelling in which the examiner dictates words for the student to spell. On a few tests, such as the Peabody Individual Achievement Test—Revised—Normative Update (Markwardt, 1997), the student is asked to recognize the correct spelling of a word from several plausible choices (e.g., *redy, reddy, ready,* or *retty*). Although these standardized measures can provide a global estimate of the current level of spelling skill, CBM procedures are more useful for documenting improvement in spelling skill.

Curriculum-based measurement procedures. In a typical CBM, the evaluator reads aloud a list of words for the students to try to spell at a prescribed pace, such as one every 7 to 10 seconds for a total of 2 minutes. For first- through third-grade students, 12 grade-level words are administered at the rate of one word every 10 seconds. For older students in Grades 4 to 8, 17 words are administered at the rate of one word every 7 seconds. Some school districts

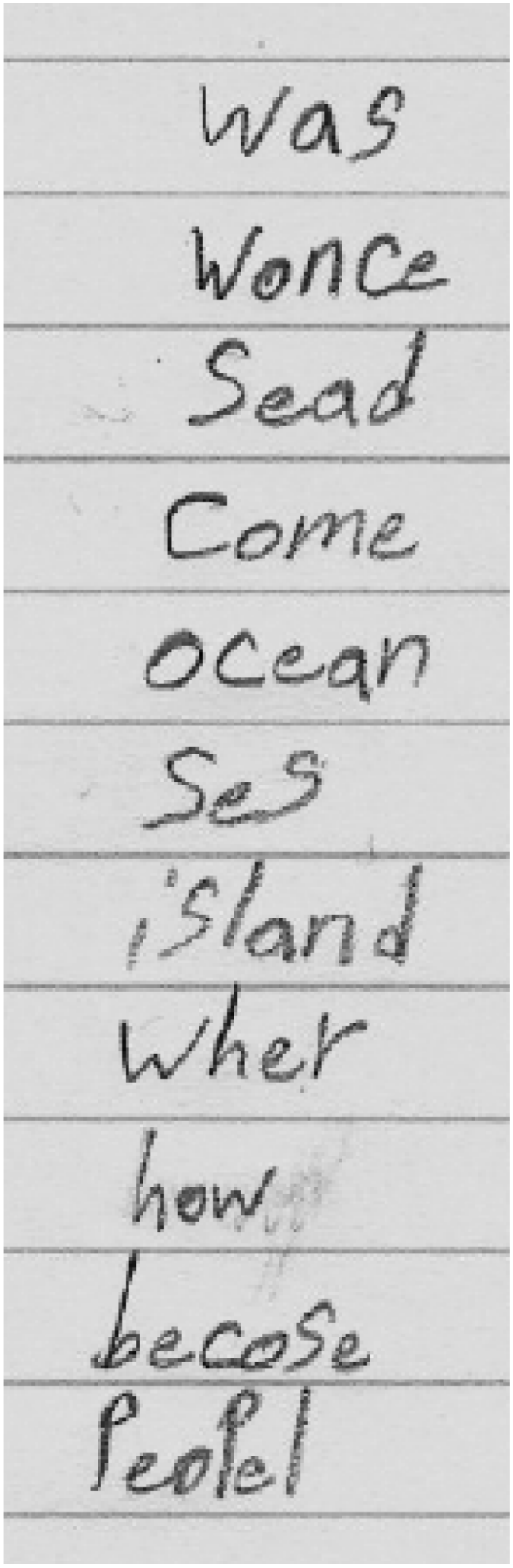


FIGURE 5.1. Mark's spelling list.

have created their own graded spelling lists for this purpose, whereas others use a commercial program, such as the Aimsweb Spelling Curriculum-Based Measurement (Psychological Corporation, 2008; http://www.aimsweb.com) that provides graded spelling lists for Grades 1 through 8.

Administering spelling probes is simple and often done with groups. Students are told how much time they will have to write each word (e.g., 7 or 10 seconds) and that they will get credit for each letter they write correctly. They are also told to go on to the next word once it is given, even if they have not finished the previous word. The score can be based on two elements: (a) the number of words spelled correctly, (b) the number of correct letter sequences. *Correct letter sequences* are pairs of letters in a word that are written in the correct order. Scoring correct letter sequences gives partial credit for words that are misspelled and helps pinpoint growth and monitor progress. The maximum credit for correct letter sequences is always 1 point more than the number of letters in the word: a four-letter word has five correct letter sequences. A blank or empty space that occurs both before and after the word creates five possible letter sequences (e.g., *_s_a_i_d_*). One method used when marking correct letter sequences is to place a caret above each correct letter pair or space and letter. In Figure 5.1, Mark spelled the word *said* as "sead." This would be scored as ^s e a d^ for a score of 2 correct letter sequences.

Written Expression

Because reciprocal influences exist between oral and written language, oral language abilities affect an individual's ability to compose written text (Berninger & Wolf, 2009). Thus, when evaluating a student's ability to express ideas, an evaluator should include measures of both receptive and expressive oral language. Common errors found in older struggling writers include limited variety in use of sentence structures and errors in producing grammatical structures, such as omitting words or phrases, reversing word order, substituting related words, overusing simple sentence structures, and using incorrect verb tenses (Gregg, Coleman, Stennett, & Davis, 2002).

Standardized tests. Standardized tests of written expression typically require the student to write a story to a picture or a prompt within a specified time limit (e.g., 15 minutes). When assessing written expression, it is important to analyze several longer pieces of text (e.g., essays) rather than relying solely on the results obtained from standardized tests, which often include only one writing sample or several short samples completed within one time frame. In addition, because many standardized writing tests require that the student complete the story within a limited time frame, the student does not have much time for planning and constructing a cohesive plot.

Curriculum-based measurement procedures. CBMs can also be used to monitor in growth in written expression. As with spelling, CBM writing probes can be administered to groups or individuals. A common method is to provide a grade-appropriate story starter sentence printed at the top of a sheet of lined composition paper (see Table 5.6). Often, the first time CBM writing tasks are administered, the evaluator will collect three samples during the same session or on consecutive days. The purpose of getting three samples is to determine the student's median score. The median, or middle score, becomes the first data point or baseline for progress monitoring of that student's writing performance.

When administering the story starter probe, the student is allowed 1 minute to think about the story and then 3 minutes to write the story or 7 to 10 minutes at the secondary level. The following types of criteria may be used to score the probes: (a) the total number of words written, (b) the words spelled correctly, (c) the total number of letters written, or (d) the number of writing units in correct sequence. For the total words written criterion, the evaluator counts all the words written, including the title, misspelled words, or nonsense words. For the total letters written criterion, the evaluator simply counts up all the letters written, including those in the misspelled words. To determine the words spelled correctly, the evaluator counts each correctly spelled word in the writing sample in isolation rather than in the context of the sentence. For the writing units in correct sequence, the evaluator considers the accuracy of spelling, grammar, capitalization, and punctuation. Although this criterion requires more time and effort to score, it is worthwhile because of the additional information it provides about the quality of the writing. To determine the writing units in correct sequence, the evaluator starts at the beginning of the writing sample and considers each successive pair of writing units, which is defined as a word or an essential punctuation mark, such as a period or question mark. As with scoring correct letter sequences in spelling, a caret is used to mark between each correct writing sequence.

MATHEMATICS

Compared with reading, not much is known about low math performance (Monuteaux, Faraone, Herzig, Navsaria, & Biederman, 2005), even though between 5% and 8% of school-age children in the United States are estimated to have a SLD in mathematics (Badian, 1983; Geary, 2004). Although the prevalence of a SLD in math appears to be comparable to the prevalence of reading disability, less attention has been devoted to understanding these problems (Jordan, Levine, & Huttenlocher, 1995; Mazzocco & Myers, 2003). This lack of attention may be partially because of the complexity associated with the study of mathematics (Landerl, Bevan, & Butterworth, 2004). As with reading disabilities, specific cognitive processes are often linked to the SLD. For example, growth in working memory has been shown to be an important predictor of children's abilities to solve mathematical problems (Swanson, Jerman, & Zheng, 2008). The description of the math performance problems of children is further complicated by the effect of oral language on math development; students with accompanying language-based disorders frequently find reading and interpreting word problems difficult because of the language comprehension involved.

Children with an SLD in math are actually quite heterogeneous with regard to the types of problems they experience (Ackerman & Dykman, 1995; Geary et al., 1999; Geary, Hamson, & Hoard, 2000). Within the federal regulations specifying the operational criteria for identifying children with SLD, two separate types of math disabilities are described: mathematics calculation and mathematics problem

solving. Although these two areas are recognized, compared with reading and written language, fewer standardized math assessments exist. The most widely used test is the KeyMath3 (Connolly, 2007), which is appropriate for use up to the age of 21. Table 5.8 lists commonly used individualized measures of mathematics.

Basic Math Skills

Psychologists, teachers, and educational consultants are familiar with the wide range of behavioral symptoms frequently seen in students with math disabilities. Examples are difficulty completing a sequence of steps in a multistep problem (such as a long-division problem), difficulty aligning numbers, using incorrect regrouping procedures in computation, difficulty comprehending the differing value of a digit according to place value, trouble comprehending fractional concepts (including ratios, percentages, and decimals), and trouble memorizing and recalling basic math facts (Mercer & Pullen, 2005). Although no one feature is a marker of an SLD in mathematics, weak recall of basic number facts is often cited as one of the most common characteristics (Geary, 1994; Geary, Hoard, & Hamson, 1999; Greene, 1999). Individuals with poor computational abilities often have difficulty with both the representation and the retrieval of math facts, implicating weaknesses within the storage and retrieval process (Geary, 1993, 2007).

Some students with reading problems also have problems in mathematics calculation (Swanson, Hoskyn, & Lee, 1999); for others, calculation skills are higher than both reading and spelling scores (Bonafina, Newcorn, McKay, Koda, & Halperin, 2000). Calculation skills depend on attentional resources and working memory capabilities—both vital skills for mental mathematics. Geary (1993) discussed the importance of the speed of reasoning, which may underlie the development of accurate representations of math facts that is necessary for automatic fact recall.

Standardized tests. Most standardized tests include a separate assessment of computational or calculation skills, ranging from problems in simple addition to higher level algebraic equations. Some tests have a time limit (e.g., Wide Range Achievement Test, 4th ed.; Jastak & Jastak, 2005); others do not (e.g., Kaufman Test of Educational

TABLE 5.8

Commonly Used Standardized Measures of Mathematics

Test name	Age range	Abilities
Comprehensive Mathematical Abilities Test (Hresko, Schlieve, Herron, Swain, & Sherbenou, 2003)	7 years, 0 months–18 years, 11 months	Comprehension (reasoning), calculation, and application
Early Math Diagnostic Assessment (Psychological Corporation, 2002)	Prekindergarten–3rd grade	Math reasoning and numerical operations
KeyMath3 Diagnostic Assessment (Connolly, 2007)	4 years, 6 months–21 years, 11 months	Conceptual knowledge, computational skills, and problem solving
Process Assessment of the Learner, Diagnostics for Math 2nd ed. (Berninger, 2007a)	Kindergarten–6th grade	Math skills and related cognitive processes
STAR Math(r), Version 2.0. (Renaissance Learning, 2002)	1st–12th grade	Numeration concepts, computation, problem solving, estimation, data analysis and statistics, geometry, measurement, and algebra
Test of Early Mathematics Ability 3rd ed. (Ginsburg & Baroody, 2003)	3 years, 0 months–8 years, 11 months	Number skills, number-comparison facility, numeral literacy, number facts, calculation skills, and understanding of concepts.
Test of Mathematical Abilities 2nd ed. (Brown, Cronin, & McEntire, 1994)	8 years, 0 months–18 years, 11 months	Story problems and computation

Achievement II; WJ III ACH). In addition, the Wechsler Individual Achievement Test—III and WJ III ACH contain measures of math fluency that are timed tests involving the calculation of simple addition, subtraction, and multiplication problems. As with timed reading measures, math fluency tests are designed to explore mastery of and automaticity with basic math facts.

Curriculum-based measurement procedures. For math CBM procedures, a worksheet is developed with various computations. Five main types of basic CBM math probes are used: quantity array, number identification, quantity discrimination, missing number, and computation. All of these types of math probes are administered individually and have been used as screening tools and for progress monitoring (Fuchs & Fuchs, 2005). For quantity array, students are asked to orally identify the number of dots in a box within 5 seconds. For number identification, the teacher marks the student's responses on the score sheet as the student says the answers aloud within 3 seconds. The quantity discrimination test requires the student to orally identify the bigger number from a pair of numbers within 3 seconds. The teacher marks the student's responses on the score sheet as the student says the answers aloud. The missing number test requires the student to orally identify the missing number in a sequence of numbers. Similar to the other procedures, correct answers within 3 seconds are scored as correct. Figure 5.2 illustrates the five different types of CBM probes.

When scoring CBM computation, students receive 1 point for each correct digit. Although one scoring procedure is the total number of correct problems, scoring the numbers of correct digits within each problem provides a more sensitive index of student change (Fuchs & Fuchs, 2005). Reversed or rotated digits are not counted as errors unless their change in position makes them into another digit (e.g., 9 and 6).

Math Problem Solving

Many students who have trouble with math problem solving either lack the prerequisite skills or do not apply the resources needed to solve word problems. Problem solving and word problems are an important part of mathematics programs in elementary schools because word problems help students apply formal mathematical knowledge and skills to real-world situations.

Standardized tests. Unless group administered, the problems on standardized tests of math problem

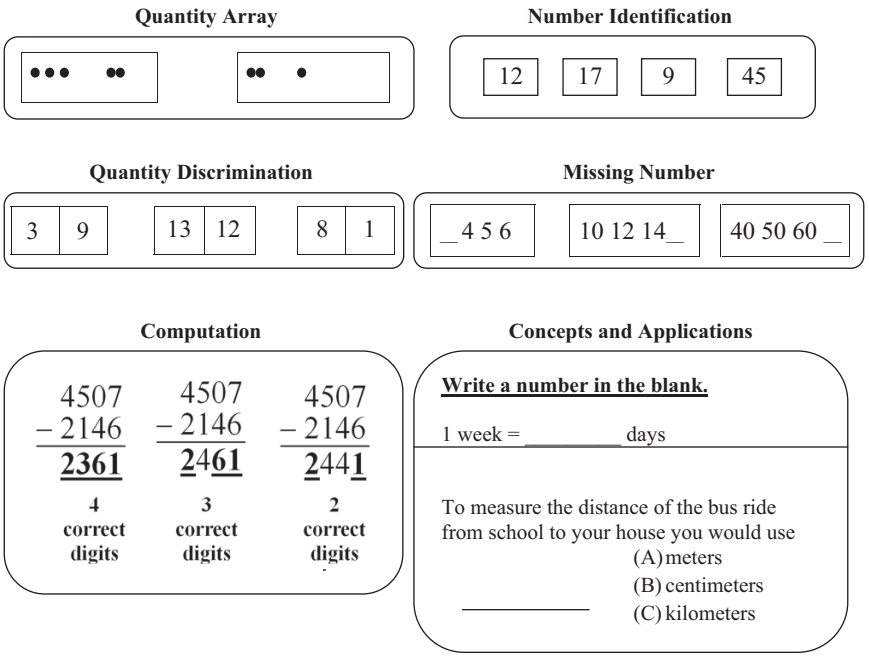


FIGURE 5.2. Math curriculum-based measurement probes.

solving are read to individuals. Although reading ability is eliminated as a confounding variable, listening comprehension is required. On most measures, however, students may ask for a problem to be repeated and are provided with pencil and paper, which reduces the demands on working memory.

Curriculum-based measurement procedures. Most of the CBM research in mathematics has focused on computational fluency in elementary school math rather than on early mathematics development (preschool through Grade 1) or problem solving (second through sixth grade; Jitendra, Czesniak, & Deatline-Buchman, 2005). CBM problem solving or concepts and applications can be administered to a group of students at one time. As with computation probes, the evaluator presents each student with a CBM concepts and applications test. Students answer the math problems, and their score is based on the total number of correct items within the set time limit (Fuchs & Fuchs, 2005).

CONCLUSIONS

The major purposes of individualized academic evaluations are to (a) establish present levels of achievement, (b) determine what the student can and cannot do, (c) pinpoint patterns of strengths and weaknesses, (d) identify ways to measure and monitor future academic progress, and (e) determine specific educational needs. Strengths and weaknesses need to be assessed so that treatment plans can be developed to address both (Gilger & Hynd, 2008). Both standardized and formative assessments are useful for planning interventions and monitoring the academic progress of students with low achievement. Standardized assessments help illuminate the nature and severity of a specific problem or disorder and provide clinicians with the opportunity to observe how different children approach different tasks (Lichtenberger et al., 2009). In fact, the way in which a student achieves a score is often more significant than the actual score itself (Wiznitzer & Scheffel, 2009). Formative assessments, such as CBMs and CBAs, also provide useful information about the rate at which a student is achieving and the effectiveness of the interventions used.

The results obtained from individualized academic assessments are useful for both diagnosis and educational planning. These measurement tools are designed to help evaluators understand a student's current levels of academic development and competence. As Stanger and Donohue (1937) observed,

> If these tests will give us a basis from which we can start to understand a child's difficulties, they will have justified the time spent on them. Anything which helps educators or parents to understand any phase of development or lack of development is of immeasurable value. (p. 189)

Care has to be taken, however, to ensure that the results from all academic assessments lead to better educational outcomes by enhancing the quality of instruction. Popham (2009) aptly noted that "too much testing time robs students of the teaching time they deserve" (p. 85). Thus, evaluators must strive to achieve a balance between assessment and instruction to ensure that the results from all academic assessments, both standardized and curriculum based, are used to inform and guide the delivery of instruction.

References

Aaron, P. G., Joshi, M., & Williams, K. A. (1999). Not all reading disabilities are alike. *Journal of Learning Disabilities, 32*, 120–137. doi:10.1177/002221949903200203

Aaron, P. G., Joshi, R. M., Palmer, H., Smith, N., & Kirby, E. (2002). Separating genuine cases of reading disability from reading deficits caused by predominantly inattentive ADHD behavior. *Journal of Learning Disabilities*, *35*, 425–436. doi:10.1177/00222194020350050301

Aaron, P. G., & Simurdak, J. (1991). Reading disorders: Their nature and diagnosis. In J. E. Obrzut & G. W. Hynd (Eds.), *Neuropsychological foundations of learning disabilities: A handbook of issues, methods, and practice* (pp. 519–548). San Diego, CA: Academic Press.

Abedi, J. (2006). Psychometric issues in ELL assessment and special education eligibility. *Teachers College Record, 108*, 2282–2303. doi:10.1111/j.1467-9620.2006.00782.x

Ackerman, P. T., & Dykman, R. A. (1995). Reading-disabled students with and without comorbid

arithmetic disability. *Developmental Neuropsychology, 11*, 351–371. doi:10.1080/87565649509540625

Badian, N. A. (1983). Dyscalculia and nonverbal disorders of learning. In H. R. Myklebust (Ed.), *Progress in learning disabilities* (pp. 235–264). New York, NY: Stratton.

Bardos, A. N. (2004). *Basic Achievement Skills Inventory*. San Antonio, TX: Pearson Assessments.

Barkley, R. A. (2006). ADHD in adults: Developmental course and outcome of children with ADHD, and ADHD in clinic-referred adults. In R. A. Barkley (Ed.), *Attention-deficit/hyperactivity disorder: A handbook for diagnosis and treatment* (3rd ed., pp. 248–296). New York, NY: Guilford Press.

Bell, S. M., McCallum, R. S., & Cox, E. A. (2003). Toward a research-based assessment of dyslexia: Using cognitive measures to identify reading disabilities. *Journal of Learning Disabilities, 36*, 505–516. doi:10.1177/00222194030360060201

Berninger, V. W. (1996). *Reading and writing acquisition: A developmental neuropsychological perspective.* Oxford, England: Westview Press.

Berninger, V. W. (2007a). *Process assessment of the learner: Diagnostics for math* (2nd ed.). San Antonio, TX: Pearson Assessments.

Berninger, V. W. (2007b). *Process assessment of the learner: Diagnostics for reading and writing* (2nd ed.). San Antonio, TX: Pearson Assessments.

Berninger, V. W., & Richards, T. (2002). *Brain literacy for educators and psychologists.* San Diego, CA: Academic Press.

Berninger, V. W., & Wolf, B. J. (2009). *Teaching students with dyslexia and dysgraphia: Lessons from teaching and science.* Baltimore, MD: Paul H. Brookes.

Boada, R., Riddle, M., & Pennington, B. F. (2008). Integrating science and practice in education. In E. Fletcher-Janzen & C. R. Reynolds (Eds.), *Neuropsychological perspectives on learning disabilities in the era of RTI: Recommendations for diagnosis and intervention* (pp. 179–191). Hoboken, NJ: Wiley.

Bonafina, M. A., Newcorn, J. H., McKay, K. E., Koda, V. H., & Halperin, J. M. (2000). A cluster analytic approach for distinguishing subgroups. *Journal of Learning Disabilities, 33*, 297–307. doi:10.1177/002221940003300307

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read—A causal connection. *Nature, 301*, 419–421. doi:10.1038/301419a0

Brown, J. I., Fishco, V. V., & Hanna, G. S. (1993). *Nelson-Denny Reading Test.* Rolling Meadows, IL: Riverside.

Brown, V. L., Cronin, M. E., & McEntire, E. (1994). *Test of Mathematical Abilities* (3rd ed.). Austin, TX: Pro-Ed.

Brown, V. L., Hammill, D. D., & Wiederholt, J. L. (2008). *Test of Reading Comprehension* (4th ed.). Austin, TX: Pro-Ed.

Bruck, M. (1993). Component spelling skills of college students with childhood diagnoses of dyslexia. *Learning Disability Quarterly, 16*, 171–184. doi:10.2307/1511325

Bryant, B. R., Wiederholt, J. L., & Bryant, P. B. (2004). *Gray Diagnostic Reading Tests* (2nd ed.). Austin, TX: Pro-Ed.

Bryant, P. E., MacLean, M., Bradley, L. L., & Crossland, J. (1990). Rhyme and alliteration, phoneme detection, and learning to read. *Developmental Psychology, 26*, 429–438. doi:10.1037/0012-1649.26.3.429

Burns, M. K., Dean, V. J., & Klar, S. (2004). Using curriculum-based assessment in the responsiveness to intervention diagnostic model for learning disabilities. *Assessment for Effective Intervention, 29*(3), 47–56. doi:10.1177/073724770402900304

Carlisle, J. F. (1993). Selecting approaches to vocabulary instruction for the reading disabled. *Learning Disabilities Research and Practice, 8*, 97–105.

Carlisle, J. F., & Rice, M. S. (2002). *Improving reading comprehension: Research-based principles and practices.* Baltimore, MD: York Press.

Carrow-Woolfolk, E. (1995). *Oral and written language scales: Listening comprehension/oral expression.* Austin, TX: Pro-Ed.

Cellucci, T., Remsperger, P., & McGlade, E. (2007). Psycho-educational evaluations for university students in one clinic. *Psychological Reports, 101*, 501–511. doi:10.2466/PR0.101.6.501-511

Coleman, C., Lindstrom, J., Nelson, J., Lindstrom, W., & Gregg, K. N. (2010). Passageless comprehension on the Nelson-Denny reading test: Well above change for university students. *Journal of Learning Disabilities, 43*, 244–249. doi:10.1177/0022219409345017

Connolly, A. J. (2007). *KeyMath-3 Diagnostic Assessment.* San Antonio, TX: Pearson Assessments.

Corn, A. L., & Koenig, A. J. (1996). *Foundations of low vision: Clinical and functional perspectives.* New York, NY: AFB Press.

Crumrine, L., & Lonegan, H. (2000). *Phonemic-awareness skills screening.* Austin, TX: Pro-Ed.

Cunningham, A. E., Stanovich, K. R., & Wilson, M. R. (1990). Cognitive variation in adult college students differing in reading ability. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 129–159). San Diego, CA: Academic Press.

Demaray, M. K., Schaefer, K., & Delong, K. (2003). Attention-deficit/hyperactivity disorder (ADHD): A

national survey of training and current assessment practices in school. *Psychology in the Schools, 40*, 583–597. doi:10.1002/pits.10129

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.

Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*, 507–524.

Dodd, B., Crosbie, S., McIntosh, B., Teitzel, T., & Ozanne, A. (2003). *Pre-Reading Inventory of Phonological Awareness.* San Antonio, TX: Psychological Corporation.

Dombrowski, S. C., Kamphaus, R. W., & Reynolds, C. R. (2004). After the demise of the discrepancy: Proposed learning disabilities diagnostic criteria. *Professional Psychology: Research and Practice, 35*, 364–372. doi:10.1037/0735-7028.35.4.364

Downing, J. E. (2010). *Academic instruction for students with moderate and severe intellectual disabilities in inclusive classrooms.* Thousand Oaks, CA: Corwin Press.

Downing, J. E., & Demchak, M. A. (2008). First steps: Determining individual abilities and how best to support students. In J. E. Downing, *Including students with severe and multiple disabilities in typical classrooms: Practical strategies for teachers* (3rd ed., pp. 49–90). Baltimore: Paul H. Brookes.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446. doi:10.1037/0012-1649.43.6.1428

Ehri, L. C. (1998). Grapheme–phoneme knowledge is essential for learning to read words in English. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 3–40). Mahwah, NJ: Erlbaum.

Evans, J. J., Floyd, R. G., McGrew, K. S., & Leforgee, M. H. (2002). The relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and reading achievement during childhood and adolescence. *School Psychology Review, 31*, 246–262.

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2007). *Learning disabilities: From identification to intervention.* New York, NY: Guilford Press.

Floyd, R. G., Evans, J. J., & McGrew, K. S. (2003). Relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and mathematics achievement across the school-age years. *Psychology in the Schools, 40*, 155–171. doi:10.1002/pits.10083

Ford, L., & Dahinten, V. S. (2005). Use of intelligence tests in the assessment of preschoolers. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 487–518). New York, NY: Guilford Press.

Frazier, T. W., Youngstrom, E. A., Glutting, J., & Watkins, M. W. (2007). ADHD and achievement: Meta-analysis of the child, adolescent, and adult literatures and a concomitant study with college students. *Journal of Learning Disabilities, 40*, 49–65. doi: 10.1177/00222194070400010401

Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom based assessment. *School Psychology Review, 28*, 659–671.

Fuchs, L. S., & Fuchs, D. (2005). *Using curriculum-based measurement for progress monitoring in math.* Retrieved from http://www.progressmonitoring.org

Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin, 114*, 345–362. doi:10.1037/0033-2909.114.2.345

Geary, D. C. (1994). *Children's mathematical development: Research and practical applications.* Washington, DC: American Psychological Association. doi:10.1037/10163-000

Geary, D. C. (2003). Learning disabilities in arithmetic: Problem-solving differences and cognitive deficits. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 199–212). New York, NY: Guilford Press.

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37*, 4–15. doi:10.1177/00222194040370010201

Geary, D. C. (2007). An evolutionary perspective on learning disabilities in mathematics. *Developmental Neuropsychology, 32*, 471–519. doi:10.1080/87565640701360924

Geary, D. C., Hamson, C. O., & Hoard, M. K. (2000). Numerical and arithmetical cognition: A longitudinal study of process and concept deficits in children with learning disability. *Journal of Experimental Child Psychology, 77*, 236–263. doi:10.1006/jecp.2000.2561

Geary, D. C., Hoard, M. K., & Hamson, C. O. (1999). Numerical and arithmetical cognition: Patterns of functions and deficits in children at risk for a mathematical disability. *Journal of Experimental Child Psychology, 74*, 213–239. doi:10.1006/jecp.1999.2515

Gickling, E., & Rosenfield, S. (1995). Best practices in curriculum-based assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (3rd ed., pp. 587–595). Washington, DC: National Association of School Psychologists.

Gilger, J. W., & Hynd, G. W. (2008). Neurodevelopmental variation as a framework for thinking about the twice exceptional. *Roeper Review, 30*, 214–228. doi:10.1080/02783190802363893

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability* (3rd ed.). Austin, TX: Pro-Ed.

Goldstein, S. (1997). *Managing attention and learning disorders in late adolescence and adulthood: A guide for practitioners.* New York, NY: Wiley.

Goldstein, S., & Cunningham, S. (2009). Current issues in the assessment of intelligence, specific learning disability, and attention-deficit/hyperactivity disorder. In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to intelligence and achievement* (pp. 11–23). Hoboken, NJ: Wiley.

Good, R. H., Kaminski, R. A., Moats, L. C., Laimon, D., Smith, S., & Dill, S. (2003). *Dynamic indicators of basic early literacy skills* (6th ed.). Frederick, CO: Sopris West.

Greene, G. (1999). Mnemonic multiplication fact instruction for students with learning disabilities. *Learning Disabilities Research and Practice, 14*, 141–148. doi:10.1207/sldrp1403_2

Gregg, N. (2007). Underserved and underprepared: Postsecondary learning disabilities. *Learning Disabilities Research and Practice, 22*, 219–228. doi:10.1111/j.1540-5826.2007.00250.x

Gregg, N. (2009). *Adolescents and adults with learning disabilities and ADHD: Assessment and accommodation.* New York, NY: Guilford Press.

Gregg, N., Coleman, C., Davis, M., & Chalk, J. C. (2007). Timed essay writing: Implications for high-stakes tests. *Journal of Learning Disabilities, 40*, 306–318. doi:10.1177/00222194070400040201

Gregg, N., Coleman, C., Stennett, R. B., & Davis, M. (2002). Discourse complexity of college writers with and without disabilities: A multidimensional analysis. *Journal of Learning Disabilities, 35*, 23–, 38, 56. doi:10.1177/002221940203500103

Hale, J. B., Naglieri, J. A., Kaufman, A. S., & Kavale, K. A. (2004). Specific learning disability classification in the new Individuals With Disabilities Education Act: The danger of good ideas. *School Psychologist, 58*(1), 6–13, 29.

Hammill, D. D., Brown, V. L., Larsen, S. C., & Wiederholt, J. L. (2007). *Test of Adolescent and Adult Language* (4th ed.). Austin, TX: Pro-Ed.

Hammill, D. D., & Larsen, S. C. (2008). *Test of Written Language* (4th ed.). Austin, TX: Pro-Ed.

Hammill, D. D., & Larsen, S. C. (2009). *Written Language Observation Scale.* Austin, TX: Pro-Ed.

Hammill, D. D., Mather, N., & Roberts, R. (2001). *Illinois Test of Psycholinguistic Abilities* (3rd ed.). Austin, TX: Pro-Ed.

Hammill, D. D., Pearson, N. A., Hresko, W. P., & Hoover, J. J. (2012). *Early Reading Assessment.* Austin, TX: Pro-Ed.

Hoover, J. J., & Mendez-Barletta, L. M. (2008). Considerations when assessing ELLs for special education. In J. K. Klingner, J. J. Hoover, & L. Baca (Eds.), *Why do English language learners struggle with reading?* (pp. 93–108). Thousand Oaks, CA: Corwin Press.

Hresko, W. P., Herron, S. R., & Peak, P. K. (1996). *Test of early written language* (2nd ed.). Austin, TX: Pro-Ed.

Hresko, W. P., Peak, P., Herron, S., & Bridges, D. (2000). *Young Children's Achievement Test.* Austin, TX: Pro-Ed.

Hresko, W. P., Schlieve, P. L., Herron, S. R., Swain, C., & Sherbenou, R. J. (2003). *Comprehensive Mathematical Abilities Test.* Austin, TX: Pro-Ed.

Individuals With Disabilities Education Improvement Act of 2004, Pub. L. 108–446, 20 U.S.C. § 1400 *et seq.*

Invernizzi, M., Sullivan, A., Meier, J., & Swank, L. (2004). *Phonological Awareness Literacy Screening.* Charlottesville: University of Virginia.

Jaffe, L. E. (2009). Supplementary manual: *Woodcock–Johnson III Tests of Achievement—Braille Adaptation.* Louisville, KY: American Printing House for the Blind

Jaffe, L. E., & Henderson, B. W. (with Evans, C. A., McClurg, L., & Etter, N). (2009). *Woodcock–Johnson III Tests of Achievement Normative Update—Braille Adaptation.* Louisville, KY: American Printing House for the Blind.

Jastak, J. F., & Jastak, S. (2005). *Wide Range Achievement Test* (4th ed.). Lutz, FL: Psychological Assessment Resources.

Jitendra, A. K., Czesniak, E., & Deatline-Buchman, A. (2005). An exploratory validation of curriculum-based mathematical word problem-solving tasks as indicators of mathematics proficiency for third graders. *School Psychology Review, 34*, 358–371.

Johnson, D. J., & Myklebust, H. R. (1967). *Learning disabilities: Educational principles and practices.* New York, NY: Grune & Stratton.

Jordan, N. C., Kaplan, D., Ola'h, L. N., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development, 77*, 153–175. doi:10.1111/j.1467-8624.2006.00862.x

Jordan, N. C., Levine, S. C., & Huttenlocher, J. (1995). Calculation abilities in young children with different patterns of cognitive functioning. *Journal of Learning Disabilities, 28*, 53–64. doi:10.1177/002221949502800109

Jordan, R. R., Kirk, D. J., & King, K. (2003). *Early Reading Diagnostic Assessment* (2nd ed.). San Antonio, TX: Psychological Corporation.

Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence: A handbook for professional practice.* Boston, MA: Allyn & Bacon.

Kaufman, A. S. (1994). *Intelligent testing with the WISC–III.* New York, NY: Wiley.

Kaufman, A. S., & Kaufman, N. L. (2004a). *Kaufman Assessment Battery for Children* (2nd ed.). Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (2004b). *Kaufman Test of Educational Achievement* (2nd ed.). San Antonio, TX: Pearson Assessments.

Kaufman, A. S., Lichtenberger, E. O., & Naglieri, J. A. (1999). Intelligence testing in the schools. In C. R. Reynolds & T. Gutkin (Eds.), *The handbook of school psychology* (3rd ed., pp. 307–349). New York, NY: Wiley.

Kavale, K. A., Kaufman, A. S., Naglieri, J. A., & Hale, J. (2005). Changing procedures for identifying learning disabilities: The danger of poorly supported ideas. *School Psychologist, 59*, 16–25.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills that they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*, 281–300. doi:10.1080/10888430802132279

Klingner, J. K., & Geisler, D. (2008). Helping classroom reading teachers distinguish between language acquisition and learning disabilities. In J. K. Klingner, J. J. Hoover, & L. Baca (Eds.), *Why do English language learners struggle with reading?* (pp. 57–73). Thousand Oaks, CA: Corwin Press.

Konold, T. R., & Pianta, R. C. (2005). Empirically-derived, person-oriented patterns of school readiness in typically-developing children: Description and prediction to first-grade achievement. *Applied Developmental Science, 9*, 174–187. doi:10.1207/s1532480xads0904_1

Kraemer, R. J. (2010). *Special education placement factors for Latino students*. Unpublished doctoral dissertation, University of Arizona, Tucson.

Landerl, K., Bevan, A., & Butterworth, B. (2004). Developmental dyscalculia and basic numerical capacities: A study of 8–9 year old students. *Cognition, 93*, 99–125. doi:10.1016/j.cognition.2003.11.004

Larsen, S. C., Hammill, D. D., & Moats, L. C. (1999). *Test of Written Spelling* (4th ed.). Austin, TX: Pro-Ed.

Lenski, S. D., Ehlers-Zavala, F., Daniel, M. C., & Sun-Irminger, X. (2006). Assessing English-language learners in mainstream classrooms. *Reading Teacher, 60*, 24–34. doi:10.1598/RT.60.1.3

Lerner, J. W., & Kline, F. (2005). *Learning disabilities and related disorders: Characteristics and teaching strategies* (10th ed.). Boston, MA: Houghton Mifflin.

Lichtenberger, E. O., Mather, N., Kaufman, N. L., & Kaufman, A. S. (2004). *Essentials of assessment report writing*. New York, NY: Wiley.

Lichtenberger, E. O., Sotelo-Dynega, M., & Kaufman, A. S. (2009). The Kaufman Assessment Battery for Children. In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (2nd ed., pp. 61–93). Hoboken, NJ: Wiley.

Lindamood, P. C., & Lindamood, P. (2004). *Lindamood Auditory Conceptualization Test* (3rd ed.). Austin, TX: Pro-Ed.

Lombardino, L. J., Lieberman, J., & Brown, J. C. (2005). *Assessment of Literacy and Language*. San Antonio, TX: Pearson Assessments.

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. *Developmental Psychology, 36*, 596–613. doi:10.1037/0012-1649.36.5.596

Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2007). *Test of Preschool Early Literacy*. Austin, TX: Pro-Ed.

Markwardt, F. C., Jr. (1997). *Peabody Individual Achievement Test—Revised—Normative Update*. San Antonio, TX: Psychological Corporation.

Marston, D. (1989). Curriculum-based measurement: What is it and why do it? In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York, NY: Guilford Press.

Mastropieri, M. A., Leinart, A., & Scruggs, T. E. (1999). Strategies to increase reading fluency. *Intervention in School and Clinic, 34*, 278–283. doi:10.1177/105345129903400504

Mather, N., & Gregg, N. (2006). Specific learning disabilities: Clarifying, not eliminating, a construct. *Professional Psychology: Research and Practice, 37*, 99–106. doi:10.1037/0735-7028.37.1.99

Mather, N., Hammill, D. D., Allen, E. A., & Roberts, R. (2004). *Test of Silent Word Reading Fluency*. Austin, TX: Pro-Ed.

Mather, N., Roberts, R., Hammill, D. D., & Allen, E. A. (2008). *Test of Orthographic Competence*. Austin, TX: Pro-Ed.

Mather, N., & Wendling, B. J. (2009). Woodcock–Johnson III Tests of Achievement. In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (pp. 503–535). New York, NY: Wiley.

Mazzocco, M. M. M., & Myers, G. F. (2003). Complexities in identifying and defining mathematics learning disability in the primary school-age years. *Annals of Dyslexia, 53*, 218–253. doi:10.1007/s11881-003-0011-7

McCardle, P., Scarborough, H. S., & Catts, H. W. (2001). Predicting, explaining, and preventing children's reading difficulties. *Learning Disabilities Research and Practice, 16*, 230–239. doi:10.1111/0938-8982.00023

McGhee, R., Bryant, B. R., Larsen, S. C., & Rivera, D. M. (1995). *Test of Written Expression*. Austin, TX: Pro-Ed.

McGrew, K. S. (1994). *Clinical interpretation of the Woodcock–Johnson Tests of Cognitive Ability—Revised.* Needham Heights, MA: Allyn & Bacon.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf–Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151–180). New York, NY: Guilford Press.

McGrew, K. S., & Knopik, S. N. (1993). The relationship between the WJ-R *Gf-Gc* cognitive clusters and writing achievement across the life span. *School Psychology Review, 22*, 687–695. doi:10.1177/073428299501300102

McKenzie, R. G. (2009). Obscuring vital distinctions: The oversimplification of learning disabilities within RTI. *Learning Disability Quarterly, 32*, 203–215.

Mercer, C. D., & Pullen, P. C. (2005). *Students with learning disabilities* (6th ed.). Upper Saddle River, NJ: Merrill.

Meyer, M. S., & Felton, R. H. (1999). Repeated reading to enhance fluency: Old approaches and new directions. *Annals of Dyslexia, 49*, 283–306. doi:10.1007/s11881-999-0027-8

Moats, L. C. (1995). *Spelling: Development, disability, and instruction.* Timonium, MD: York Press.

Monroe, M. (1932). *Children who cannot read.* Chicago, IL: University of Chicago Press.

Monuteaux, M. C., Faraone, S. V., Herzig, K., Navsaria, N., & Biederman, J. (2005). ADHD and dyscalculia: Evidence for independent familial transmission. *Journal of Learning Disabilities, 38*, 86–93. doi:10.1177/00222194050380010701

National Education Association. (2006). *The twice-exceptional dilemma.* Washington, DC: Author.

Newcomer, P. L. (2001). *Diagnostic Achievement Battery* (3rd ed.). Austin, TX: Pro-Ed.

Newcomer, P. L., & Barenbaum, E. (2003). *Test of Phonological Awareness Skills (TOPAS)*. Austin, TX: Pro-Ed.

Perfetti, C. A., Marron, M. A., & Foltz, P. W. (1996). Sources of comprehension failure: Theoretical perspectives and case studies. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 137–165). Mahwah, NJ: Erlbaum.

Popham, W. J. (2009). A process—not a test. *Educational Leadership, 66*, 85–86.

Psychological Corporation. (2002). *Early Math Diagnostic Assessment.* San Antonio, TX: Author.

Psychological Corporation. (2004). *Early Reading Success Indicator.* San Antonio, TX: Author.

Psychological Corporation. (2008). *AIMSweb* [Computer software]. San Antonio, TX: Author.

Psychological Corporation. (2009). *Wechsler Individual Achievement Test* (3rd ed.). San Antonio, TX: Author.

Rack, J. P., Snowling, M. J., & Olson, R. K. (1992). The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quarterly, 27*, 28–53. doi:10.2307/747832

Reid, K. D., Hresko, W. P., & Hammill, D. D. (2001). *Test of Early Reading* (3rd ed.). Austin, TX: Pro-Ed.

Renaissance Learning. (2002). *STAR Math, Version 2.0.* Wisconsin Rapids, WI: Author.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-Based Measurement Oral Reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*, 427–469. doi:10.1016/j.jsp.2009.07.001

Reynolds, C. R., & Kamphaus, R. W. (2007). *Test of Irregular Word Reading Efficiency*. Lutz, FL: Psychological Assessment Resources.

Riccio, C. A., Imhoff, B., Hasbrouck, J. E., & Davis, G. N. (2004). *Test of Phonological Awareness in Spanish.* Austin, TX: Pro-Ed.

Robin, A. L. (2006). Training families with adolescents with ADHD. In R. A. Barkley (Ed.), *Attention-deficit/hyperactivity disorder: A handbook for diagnosis and treatment* (3rd ed., pp. 499–546). New York, NY: Guilford Press.

Roswell, F. G., Chall, J. S., Curtis, M. E., & Kearns, G. (2005). *Diagnostic Assessments of Reading* (2nd ed.). Austin, TX: Pro-Ed.

Ryan, R. M., Fauth, R. C., & Brooks-Gunn, J. (2006). Childhood poverty: Implications for school readiness and early childhood education. In B. Spodek & O. N. Saracho (Eds.), *Handbook of research on the education of young children* (2nd ed., pp. 323–346). Mahwah, NJ: Erlbaum.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2010). *Assessment in special and inclusive education* (11th ed.). Belmont, CA: Wadsworth.

Schrank, F. A., & Woodcock, R. W. (2007). WJ III NU Compuscore and Profiles Program [Computer software]. Rolling Meadows, IL: Riverside.

Shaywitz, S. E., & Shaywitz, B. A. (2003). Neurobiological indices of dyslexia. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 514–531). New York, NY: Guilford Press.

Shin, J., Deno, S., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *Journal of Special Education, 34*, 164–172. doi:10.1177/002246690003400305

Siegel, E., & Allinder, R. M. (2005). Review of assessment procedures for students with moderate and severe disabilities. *Education and Training in Developmental Disabilities, 40*, 343–351.

Snell, M. E. (2002). Using dynamic assessment with learners who communicate nonsymbolically. *Augmentative and Alternative Communication, 18*, 163–176. doi:10.1080/07434610212331281251

Snow, C. (2003). Assessment of reading comprehension. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 192–206). New York, NY: Guilford Press.

Snyder, L., Caccamise, D., & Wise, B. (2005). The assessment of reading comprehension: Considerations and cautions. *Topics in Language Disorders, 25*, 33–50. doi:10.1097/00011363-200501000-00005

Stanger, M. A., & Donohue, E. K. (1937). *Prediction and prevention of reading difficulties.* New York, NY: Oxford University Press.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*, 795–819. doi:10.1002/pits.20113

Swanson, H. L., Hoskyn, M., & Lee, C. (1999). *Interventions for students with learning disabilities: A meta-analysis of treatment outcomes.* New York, NY: Guilford Press.

Swanson, H. L., & Hsieh, C. J. (2009). Reading disabilities in adults: A selective meta-analysis of the literature. *Review of Educational Research, 79*, 1362–1390. doi:10.3102/0034654309350931

Swanson, H. L., Jerman, O., & Zheng, X. (2008). Growth in working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology, 100*, 343–379. doi:10.1037/0022-0663.100.2.343

Swanson, H. L., Mink, J., & Bocian, K. M. (1999). Cognitive processing deficits in poor readers with symptoms of reading disabilities and ADHD: More alike than different? *Journal of Educational Psychology, 91*, 321–333. doi:10.1037/0022-0663.91.2.321

Taylor, R. L. (2008). *Assessment of exceptional students: Educational and psychological procedures* (8th ed.). Boston, MA: Allyn & Bacon.

Tindal, G., McDonald, M., Tedesco, M., Glasgow, A., Almond, P., Crawford, L., & Hollenbeck, K. (2003). Alternate assessments in reading and math: Development and validations for students with significant disabilities. *Exceptional Children, 69*, 481–494.

Torgesen, J. K., & Bryant, B. R. (2004). *Test of Phonological Awareness—Second Ed.: PLUS.* Austin, TX: Pro-Ed.

Torgesen, J. K., & Miller, D. H. (2009). *Assessments to guide adolescent literacy instruction.* Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). *Test of Word Reading Efficiency* (2nd ed.). Austin, TX: Pro-Ed.

Travis, L. E. (1935). Intellectual factors. In G. M. Whipple (Ed.), *The thirty-fourth yearbook of the National Society for the Study of Education: Educational diagnosis* (pp. 37–47). Bloomington, IL: Public School.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101*, 192–212. doi:10.1037/0033-2909.101.2.192

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2012). *Comprehensive Test of Phonological Processing (*2nd ed.*)*. Austin, TX: Pro-Ed.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *Test of Silent Reading Efficiency and Comprehension.* Austin, TX: Pro-Ed.

Warden, R., & Hutchinson, T. A. (1992). *Writing Process Test.* Austin, TX: Pro-Ed.

Weiss, G., & Hechtman, L. (1993). *Hyperactive children grown up* (2nd ed.). New York, NY: Guilford Press.

Wiederholt, J. L., & Blalock, G. (2000). *Gray Silent Reading Tests.* Austin, TX: Pro-Ed.

Wiederholt, J. L., & Bryant, B. R. (2012). *Gray Oral Reading Tests* (5th ed.). Austin, TX: Pro-Ed.

Wilson, B. A., & Felton, R. H. (2004). *Word Identification and Spelling Test.* Austin, TX: Pro-Ed.

Wiznitzer, M., & Scheffel, D. L. (2009). Learning disabilities. In R. B. David, J. B. Bodensteiner, D. E. Mandelbaum, & B. Olson (Eds.), *Clinical pediatric neurology* (pp. 479–492). New York, NY: Demos Medical.

Woodcock, R. W. (1998). *Woodcock Reading Mastery Tests—Revised Edition—Normative Update.* San Antonio, TX: Pearson Assessments.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson Psycho-Educational Battery—Revised.* Itasca, IL: Riverside.

Woodcock, R. W., Mather, N., & Schrank, F. A. (2004). *Woodcock–Johnson III Diagnostic Reading Battery.* Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). *Woodcock–Johnson III Tests of Achievement—Third Edition—Normative Update.* Rolling Meadows, IL: Riverside.

CHAPTER 6

BEHAVIORAL, SOCIAL, AND EMOTIONAL ASSESSMENT OF CHILDREN

Bridget V. Dever and Randy W. Kamphaus

The range of behavioral, social, and emotional assessment procedures used by psychologists working with school-age children has expanded rapidly over the past few decades (Barry, Frick, & Kamphaus, 2009). Thus, an overview of these assessment tools and trends is in order for any text dealing with child and adolescent psychological assessment. Current trends and developing issues in the realm of behavioral, social, and emotional assessment are provided in this chapter to ensure the relevance of its content for some years to come. The first section focuses on a broad introduction to the types of assessments frequently used by psychologists. The purposes of behavioral, social, and emotional assessment are reviewed, because the goal of an assessment process will influence the particular types of strategies and tools selected. Second, commonly used methods of assessment, including rating scales and interviews, are discussed. Throughout this first section, examples of references for each type of assessment are provided. The examples cited are meant to clarify each section and are not to be considered an exhaustive list of available assessment tools, which is a veritable impossibility given that entire textbooks and handbooks have been devoted to this topic (see, e.g., Frick, Barry, & Kamphaus, 2009).

The second section discusses salient issues to be considered when conducting a behavioral, social, and emotional assessment. First, information is reviewed regarding the choice of informants for an assessment. Second, a multiple-gate model of assessment is introduced. Finally, a concluding discussion of the future directions and challenges facing psychologists working in schools is provided, focusing particularly on the importance of selecting appropriate assessment tools for an increasingly diverse student population.

INTRODUCTION TO BEHAVIORAL ASSESSMENT

The first section of this chapter provides a broad overview of current practices in behavioral assessment. This introduction includes a review of common purposes and methods of assessment. Throughout this section, examples of some of the more frequently used assessment tools for each purpose are provided, along with citations for those seeking further information on these available tools.

Purposes of Assessment

This section provides an overview of three common purposes of behavioral, social, and emotional assessment by school psychologists: (a) diagnosis and classification, (b) screening, and (c) progress monitoring. These purposes are not all inclusive because school psychologists may also conduct assessments for alternative reasons, such as assessment of schoolwide trends in behavior by tracking office discipline referrals or other information gathering. Although all three purposes reviewed here serve important roles in designing assessments, screening and progress monitoring are emphasized as two goals that are beginning to receive more attention in the field of school psychological assessment.

DOI: 10.1037/14049-006
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

Assessment for diagnosis and classification. Traditionally, school psychologists have used behavioral, social, and emotional assessment tools for the purposes of diagnosis and classification (Dowdy, Mays, Kamphaus, & Reynolds, 2009; Merrell, 2008). *Classification* is a broad term often used in the fields of education, biology, and other disciplines to differentiate between categories (Merrell, 2008), such as species, weather patterns, and educational classifications (e.g., gifted and talented, English language learner). *Diagnosis,* the more narrow term borrowed from medicine, refers to the practice of identifying disease states (e.g., learning disability, emotional or behavioral disturbance). Realistically, most of the work of school psychologists is diagnostic in nature, in that much of their time is devoted to identifying children with a disorder, or differentiating between disorders, in an effort to make educational placement decisions. Accordingly, the topic of classification receives less mention hereinafter.

The diagnosis rendered is dependent on the diagnostic system used as the model informing the assessment. One diagnostic system commonly used by school psychologists is the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), which is currently in its fourth edition (text revision; *DSM–IV–TR;* American Psychiatric Association, 2000). Since its second edition (American Psychiatric Association, 1968), the manual has included and continued to expand on diagnoses considered to be disorders usually first diagnosed in infancy, childhood, or adolescence. The *DSM–IV–TR* currently lists a variety of disorders relevant to the behavioral, social, and emotional assessment of children, including major depression, conduct disorder, attention-deficit/hyperactivity disorder (ADHD), and social phobia. Despite improvements that have been made throughout *DSM* revisions, the reliability of the classifications applicable to children and adolescents is questionable and deserving of further attention (e.g., Langenbucher & Nathan, 2006). With each new edition of the *DSM,* the disorders of childhood and adolescence have changed substantially, and additional significant revisions are anticipated for the fifth edition (Achenbach, 2005).

The *DSM–IV* and *DSM–IV–TR* (American Psychiatric Association, 1994, 2000) do not recommend specific methods or tools for informing diagnostic decisions; the determination process is left to the practitioner's discretion. However, each edition has attempted to provide more objective, reliable criteria for making diagnostic decisions than previous versions of the manual (Kauffman, 2000); these criteria have been organized into structured diagnostic interviews for identifying the specific symptoms associated with disorders. For example, the National Institute of Mental Health has produced the fourth edition of the Diagnostic Interview Schedule for Children (DISC–IV) to coincide with the major classifications of the *DSM* applicable for children between the ages of 9 and 17 years (Fisher, Wicks, Shaffer, Piacentini, & Lapkin, 1992; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000). The DISC–IV essentially codifies the *DSM–IV–TR* criteria, thus providing a one-to-one relationship between measures and symptoms necessary to diagnose various disorders. Both the *DSM–IV–TR* and its associated measurement tool the DISC are dichotomous methods, in that a child either meets the diagnostic criteria or does not, and degree of severity cannot be assessed. For example, whether a child is found to have just enough symptoms to meet the criteria for ADHD combined type or has all of the symptoms of the disorder is irrelevant; the *DSM–IV–TR* has no diagnostic criteria for mild, moderate, or severe ADHD (Dowdy et al., 2009).

Independent of this categorical method, psychological assessment has focused on the creation of an alternative dimensional assessment paradigm, with the goal of assessing constructs or latent traits that exist along continua. By way of example, the *DSM–IV–TR* classifies depression when a child's symptoms exceed the criteria set for the disorder. However, a psychometric measure of a child's depression will determine the percentile rank associated with a child's depression (or melancholy, given that this construct was identified at least as far back as the ancient Greeks), thus allowing for a child to have "subsyndromal" depression (i.e., beneath diagnostic threshold), mild depression (meets some cut score determined by research, e.g., the 98th percentile rank in comparison to same-age peers), or severe depression (e.g., a percentile rank indicating that a child or adolescent has more symptoms than virtually all same-age peers).

The *DSM*–categorical and psychometric–dimensional approaches are not entirely orthogonal, nor are they highly correlated, and debate exists as to which approach is most appropriate for diagnosing disorders of childhood (Dowdy et al., 2009). For the purposes of this chapter, dimensional–psychometric methods that assign children a given percentile rank for a construct in comparison to same-age peers are emphasized because norm-referenced assessment is highlighted in the training of school psychologists. In contrast, history taking to identify symptoms such as those contained within the *DSM* method is emphasized in medical and psychiatric assessment.

Within the U.S. education system, school psychologists' practices are guided by the Individuals With Disabilities Education Improvement Act of 2004 (IDEIA), which is the reauthorization of the Individuals With Disabilities Education Act of 1997. To qualify for special education services within the U.S. public school system, students must first be identified as having a particular disability as defined by IDEIA. Students who meet the criteria for an emotional disturbance are eligible for special education and related services or special education placement, which is a categorical classification (eligible vs. not eligible) of the same ilk as the *DSM–IV–TR*. As with the *DSM–IV–TR*, some measures of this categorical classification system are available, such as the Clinical Assessment of Behavior, which includes two superordinate scales to differentiate serious emotional disturbance from social maladjustment (Bracken & Keith, 2004). Guidelines for appropriate assessment procedures are outlined, including the use of tests that are validated for the intended purpose, choosing assessments that are culturally and linguistically appropriate, and relying on more than one assessment or instrument to make placement and service decisions; however, the specific assessment plan is left to the discretion of the practitioner or school-based team, much as with the *DSM–IV–TR* (American Psychiatric Association, 2000).

Assessment for diagnosis assists in achieving important goals in the work of school psychologists (Dowdy et al., 2009). First, accurate diagnosis enhances the goal of improving and clarifying communications between mental health and related service providers (e.g., Blashfield, 1998). Once a specific diagnosis is determined, school psychologists can better communicate with other mental health professionals who are interested in the child's well-being. For example, although a psychologist might observe that a child is somewhat impulsive, hyperactive, and inattentive, the diagnosis of ADHD would denote further information, such as the fact that these problems began early in development, they have a chronic course and are not likely to abate without specific intervention, they are severe enough to cause impairment in schooling, and they are not explained by some other problem such as the temporary reaction to the loss of a loved one. Second, diagnosis serves a critical role in the provision of services and choice of interventions that could be useful for a particular condition or set of symptoms. For example, abundant research has shown that behavioral and somatic interventions for ADHD decrease hyperactivity and yet do not lead to long-term academic improvement (Barry, Frick, & Kamphaus, 2009). Thus, knowledge of the ADHD diagnosis will also dictate the need for academic interventions concurrent with behavioral interventions. Classification into a particular group, such as one of the groups eligible for special education under IDEIA, will inform the services that are provided, such as the balance between special and regular education services dictated by the severity of diagnosis. Finally, diagnosis assists both research and clinical endeavors (Frick et al., 2009). Having a common language and common set of criteria allows researchers and clinicians to better record, report, and compare information across students, settings, and larger studies (e.g., Scotti & Morris, 2000).

Although current diagnostic systems used by school psychologists have been criticized for problems such as failure to appreciate the academic impact of comorbidity by emphasizing differential diagnosis to isolate the primary problem, reliance on subjective judgments, and questionable reliability and validity of the diagnostic category per se (e.g., pervasive developmental disorder; see Kamphaus, Dowdy, Kim, & Chen, in press), the necessity of classification for the provision of services remains. Therefore, the continual improvement of existing assessment tools, introduction of new assessment tools, and collection of validity research to support

interpretation of test score inferences remain priorities for the field. The introduction of new and revised screening measures of child behavioral and emotional risk and disorder is but one example of a new and expanding set of assessment tools that fill a specific need of school psychologists.

Screening. Although assessment for diagnosis is an important function of school psychologists, considering the role of assessment in early prevention and intervention efforts is also critical. The failure to assess for behavioral, social, and emotional difficulties early in the course of their development, and to follow assessment with appropriate intervention services, leads to poor outcomes, including academic underachievement, special education placement, and school dropout (e.g., Gutman, Sameroff, & Cole, 2003; Rapport, Denney, Chung, & Hustace, 2001). In one study of academic outcomes, 75% of students with significant emotional and behavioral problems were achieving below expected grade levels in reading, and 97% were below expected levels in mathematics (Bradley, Doolittle, & Bartolotta, 2008). As a group, children classified as having emotional and behavioral disorders at school also often have higher rates of suspensions, expulsions (Wagner, Kutash, Duchnowski, & Epstein, 2005), and school absenteeism (Lane, Carter, Pierson, & Glaeser, 2006) than children without this special education classification. In the United States, more than half of all students identified as having significant emotional or behavioral problems leave the educational system by dropping out, and only about 42% of those who remain in school graduate with a diploma (Bradley et al., 2008; U.S. Public Health Service, 2000). In the United States, and likely in many other countries, most children in need of a behavioral or emotional assessment to identify their special needs either do not receive a psychological evaluation or do not receive one in a timely way when problems are first noticed.

Despite the dire consequences experienced by students with behavioral, social, and emotional difficulties, epidemiological studies have reported that children with the worst problems, those who meet diagnostic criteria for a mental health disorder, often go unidentified and untreated (Jamieson & Romer, 2005). Large-scale epidemiological studies, for example, have revealed that approximately 20% of children in the United States have a diagnosable disorder in any given year, but only 15% to 30% of these children receive mental health services (Costello, Mustillo, Erkanli, Keeler, & Angold, 2003; Ringel & Sturm, 2001; U.S. Public Health Service, 2000). Cross-national studies have identified similar prevalence rates and service delivery inadequacies worldwide (Belfer, 2008). The use of appropriate assessment tools is critical to the identification, diagnosis, and monitoring of children with behavioral, social, and emotional problems. Moreover, without an adequate model of assessment, children with such problems are likely to continue to go unidentified and underserved by the educational and mental health systems.

Research is often conducted to identify risk factors for particular diseases or disorders, and screening tools are then developed to identify these risk factors. In 1949, Dawber and his colleagues began a longitudinal study of heart disease in Framingham, Massachusetts, known now as the Framingham Study. The intention of this study was to follow participants over time to better understand how those who developed heart disease differed from those who did not (Dawber & Kannel, 1966). This study of heart disease development had its highest value not in curing those diagnosed with the disease but in being able to predict others who were most likely to develop the disease later. Consistent with this landmark study, the goal of screening for behavioral, social, and emotional difficulties is not to diagnose a specific disorder but rather to detect the risk for developing a disorder at a later time.

When implementing a screening procedure, defining the targeted population to screen for behavioral, social, and emotional risk is necessary. In a public health model of service delivery, a school psychologist may take a multitiered approach to identifying and serving those in need of additional supports (Kamphaus et al., in press). In this public health model, universal services are provided to all students to promote positive development; selected services are provided only to those students who are identified as being at risk; finally, indicated services are reserved for those students with the most need

for intervention (Durlak, 1997). This multitiered approach is also the basis for the response-to-intervention (RtI) approach to school psychological service delivery, which consists of periodic universal screening, determination of individual student needs, and provision of interventions on the basis of data gathered throughout the entire process (Reschly & Bergstrom, 2009).

Following this model, screening could be implemented at the universal, selected, and indicated levels. In the strictest sense, universal screening includes gathering risk-related information by screening every student in the population of interest. For a school psychologist assigned to one school, this task would translate into assessing each student at that given school for behavioral, social, and emotional risk. *Selected screening* would involve assessing only those who are already considered to be in a group at risk for behavioral, social, and emotional problems. For example, students who are experiencing stressors such as family disruption or poverty might be targeted because of the relationship of these risk factors to adjustment difficulties (e.g., Gutman et al., 2003; Werner, 1994). Finally, *indicated screening* suggests screening only those students who have previously been identified or diagnosed with a behavioral, social, or emotional disorder. In this instance, the purpose of screening would be to identify specific areas of concern on which to focus treatment rather than prevention (Kamphaus et al., in press). There is debate in the literature as to which approach—universal, selective, or indicated screening—is superior (Levitt, Saka, Romanelli, & Hoagwood, 2007). The ultimate decision of which approach to use will likely depend on the resources available and whether the goal of screening is focused on prevention, early intervention, or treatment.

In recent years, screening at the universal level has been given an increasing amount of attention as the importance of prevention has come to the forefront of discussions concerning behavioral, social, and emotional difficulties (Glover & Albers, 2007; Levitt et al., 2007). A preponderance of evidence now suggests that prevention and early intervention can eliminate or reduce the severity of socioemotional and behavioral difficulties of childhood and improve relevant outcomes (Atkins, Frazier, Adil, & Talbott, 2003; Catalano, Haggerty, Oesterle, Fleming, & Hawkins, 2004; McIntosh, Flannery, Sugai, Braun, & Cochrane, 2008).

To achieve these improved outcomes for children, screening instruments must be easy to complete, brief, affordable, and reasonably accurate (O'Connell, Boat, & Warner, 2009). Providing greater guidance for test developers and users, Glover and Albers (2007) suggested that a sound screening assessment is appropriate for the intended use, technically adequate, and usable. To determine whether a measure is appropriate, the screener should have evidence of use with the population of interest, align with the constructs of interest, and have theoretical and empirical support. Technical adequacy is demonstrated through sound psychometric evidence, including norms, reliability, validity of key score inferences, sensitivity, specificity, positive predictive value, and negative predictive value (DiStefano & Kamphaus, 2007). Finally, a screener is considered usable when the associated costs are reasonable, screening is feasible and acceptable to stakeholders, resources are available to carry out the screening procedure, and the outcomes are considered useful. O'Connell et al. (2009) also suggested that screening be implemented longitudinally because risk factors and early symptoms may reveal themselves over time; this developmental issue deserves further attention because longitudinal development of behavioral, social, and emotional risk, and the appropriate assessment tools for this purpose, need continued attention and refinement (e.g., Bracken & Reinties, 2010).

The impracticality of many universal screening measures has largely contributed to their lack of adoption for universal screening in both pediatric and school settings (Flanagan, Bierman, & Kam, 2003; Saunders & Wojcik, 2004; Schmitz, Kruse, Heckrath, Alberti, & Tress, 1999). Even the most popular comprehensive behavior rating scales are not feasible for widespread screening because of the time and monetary resources needed to assess thousands of children in a given school (Flanagan et al., 2003). Yet comprehensive behavior rating scales, which typically include 50 to 100 items or more and take 10 to 45 minutes to complete, are commonly

identified and used as screeners (Levitt et al., 2007; Najman et al., 2008).

The adoption of lengthy and time-intensive assessment methods has led to both sparse screening implementation overall and a lack of development of technically adequate and practical screening measures (DiStefano & Kamphaus, 2007; Kamphaus et al., 2007). When school districts screen for behavioral or emotional problems, they often do not screen every child (i.e., selected vs. universal screening); use measures with either unknown or poor reliability and validity evidence, such as adult nomination or referral, with local or nonexistent normative standards; or use lengthy measures and procedures that were originally designed for diagnostic purposes, thus making the financial and personnel costs of screening prohibitive (Romer & McIntosh, 2005). Therefore, recent efforts have been aimed at developing screeners that assess risk for disorder rather than the disorder per se so that children's risk can be ameliorated to prevent disorder (Kamphaus & Reynolds, 2007; Levitt et al., 2007), can be completed by teachers and students without training or instruction in fewer than 5 minutes per child (DiStefano & Kamphaus, 2007; Kamphaus et al., 2007), and have minimum internal consistency estimates of reliability of .90 and preferably higher than .95 (DiStefano & Kamphaus, 2007; Kamphaus et al., 2007).

A number of screeners for socioemotional and behavioral risk have been designed that meet many of the criteria set forth by Glover and Albers (2007) and O'Connell et al. (2009). The Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997, 1999, 2001) is a brief, 25-item behavioral, emotional, and social screening test for youths ages 11 through 17 years. Teacher, parent, and student self-report forms are available, and their use has generated some longitudinal screening research. Respondents rate items on a 3-point scale ranging from 0 (*not at all*) to 2 (*very much or all the time*). The SDQ contains five scales—Emotional Symptoms, Conduct Problems, Hyperactivity/Inattention, Peer Relationship Problems, and Prosocial Behavior—each consisting of five items.

Initial SDQ reliability and validity studies produced some evidence in support of several score inferences: Coefficient alpha coefficients ranged from .63 to .83; moderate to high correlations were found between the Child Behavior Checklist and SDQ, ranging from .59 to .87; and data suggested moderate sensitivity (.77) and specificity (.85; Goodman & Scott, 1999; Hysing, Elgen, Gillberg, Lie, & Lundervold, 2007). Some studies, however, have yielded low coefficient alphas, making the use of the SDQ questionable even as a preliminary screener. For example, in a study conducted with both urban and suburban samples in the northeastern United States, coefficient alphas for the total scores were .79 and .83, respectively. However, the SDQ subscales frequently produced alphas in the .40s, .50s, and .60s for the urban sample and slightly higher values for the suburban sample (Ruchkin, Jones, Vermeiren, & Schwab-Stone, 2008). Therefore, Ruchkin et al. (2008) deemed the SDQ to have inadequate psychometric characteristics.

The Student Risk Screening Scale (Drummond, 1994) is a seven-item teacher rating scale designed to detect antisocial behavior in students enrolled in kindergarten through sixth grades. It is a practical tool thanks to its brevity, requiring elementary school teachers 15 minutes to complete for an entire classroom. In a series of investigations, Lane and colleagues (Lane, Parks, Robertson Kalberg, & Carter, 2007; Lane, Robertson Kalberg, Parks, & Carter, 2008, Lane et al., 2009) extended the use of the Student Risk Screening Scale to the middle and high school levels and compared it with the SDQ and the Systematic Screening for Behavioral Disorders (SSBD; discussed later in the chapter). Lane et al. (2009) have also used the Student Risk Screening Scale and compared it with the SSBD at the kindergarten through third-grade levels in seven elementary schools. The ethnicity of students in these schools was rather homogeneous, with a 95% White population. The Student Risk Screening Scale scores were used as the predictor variable and SSBD risk classification was used as the outcome variable in a cross-sectional design. The Student Risk Screening Scale performed fairly well in identifying externalizing problems of children in the same manner as the SSBD but performed poorly for children with internalizing problems, suggesting the need for an alternate tool to identify internalizing difficulties.

The BASC–2 Behavioral and Emotional Screening System (BESS; Kamphaus & Reynolds, 2007) is a screening system with parent-, teacher-, and self-report forms to identify behavioral and emotional resources and risks among students in kindergarten through 12th grade. The BESS forms are essentially short forms of the Behavior Assessment System for Children—Second Edition (BASC–2; Reynolds & Kamphaus, 2004), include fewer than 30 items each, take less than 10 minutes to complete for an individual child, and require no informant training. The items contained in these scales include items assessing internalizing, externalizing, school problems, and adaptive skills (reverse scored). Items are rated on a 4-point Likert scale indicating frequency of occurrence. Summing the items and transforming the raw score distribution generates a total *T* score, with higher scores reflecting greater risk. Scores of 20 to 60 suggest a normal level of risk; scores of 61 to 70 suggest an elevated level of risk; and scores of 71 or higher suggest an extremely elevated level of risk.

The normative samples for the BESS system are representative of the general population of U.S. children, closely matched demographically to the 2001 U. S. population with regard to sex, race and ethnicity, and clinical or special education classification (Kamphaus & Reynolds, 2007). Split-half reliability (.90–.97), test–retest reliability (.79–.91), and interrater reliability estimates (.70–.90) were moderate to high (Kamphaus & Reynolds, 2007). The BESS has demonstrated acceptable convergent validity with other measures of behavioral and emotional adjustment, with moderate to high correlations between .60 and .82 with the Total Problems scale of the Achenbach System of Empirically Based Assessment (ASEBA) when comparing like informants and between .68 and .73 with the Conners' Global Index. Finally, the BESS predicted the full BASC–2 (Reynolds & Kamphaus, 2004) Behavioral Symptoms Index for the same informant with levels of sensitivity (.66–.82), specificity (.95–.97), positive predictive value (.72–.82), and negative predictive value (.94–.97) that were in the moderate to high range. Lane, Menzies, Oakes, and Kalberg (2012), however, noted that few independent validity studies of the BESS are available and suggested cautious use until more independent research becomes available.

Assessment for progress monitoring. Returning to the RtI model described previously (Reschly & Bergstrom, 2009), it is critical for school psychologists to conduct periodic assessments to monitor progress associated with interventions so that treatment plans can be modified or maintained as necessary. When monitoring student progress, the school psychologist provides frequent behavioral, social, and emotional assessments; tracks change over time using graphical representation of progress; and compares this progress to stated goals and benchmarks. This procedure assists in determining whether interventions are demonstrating evidence of effectiveness at the level of the individual student (Gresham, 1991, 2002; Reschly & Bergstrom). Progress monitoring in an RtI framework has its roots in practices such as applied behavior analysis to modify behavioral, social, and emotional outcomes in settings including home and school (Bandura, 1969; Witt, Elliott, & Gresham, 1988). The application of applied behavior analysis principles has allowed researchers to develop schoolwide interventions to target behavioral problems (e.g., Horner & Sugai, 2000; Sugai et al., 2000).

The methods for tracking changes in behavior related to treatment in an applied behavior analysis setting are comparable to tracking behavioral, social, and emotional change. School psychologists and behavioral therapists are often left to determine which instruments and procedures, often developed for other purposes, can be used effectively for progress monitoring. In addition, the threshold for determining adequate progress on a given measure might also be left to subjective judgments on the part of the practitioner. Therefore, those interested in the assessment of behavioral and emotional progress should work on selecting or developing reliable, valid, and brief tools that can be used to track and test for significant progress over time related to intervention efforts. Additionally, research is needed to examine the most effective frequency for monitoring assessments to best assess progress and capture the true pattern of change over variable periods of time.

Progress monitoring procedures are relatively well developed for the assessment of children experiencing academic problems, thanks to the

programmatic research of Fuchs and Fuchs (e.g., Fuchs, Fuchs, & Compton, 2004). The focus of much of the work to date has been on tracking progress in reading in the elementary grades (Wayman, Wallace, Wiley, Ticha, & Espin, 2007). Comparatively speaking, the development of measures and methods for behavioral, social, and emotional progress monitoring has lagged, although the pace of development may be quickening.

One measure that has been developed for the sole purpose of monitoring progress in particular domains is the BASC–2 Progress Monitor (Reynolds & Kamphaus, 2009). After intervention delivery, the effectiveness of each intervention in addressing the targeted problem or problems is assessed using teacher-, parent-, or self-rated progress monitoring scales completed at frequent intervals, typically every 2 weeks. The Progress Monitor rating forms are designed to assess specific problems, including externalizing and ADHD problems, internalizing problems, social withdrawal, and adaptive skills. Progress is assessed by using statistical procedures to test the hypothesis that changes in a child's behavior are likely the result of the intervention and not of chance variation. Results are presented in the form of graphical trend lines and statistical significance tests. Median values for the reliability estimates for the combined-sex norm group ranged from .79 to .95 across all age levels. The progress monitoring measures also have strong supportive evidence of reliability and validity, suggesting that scores do indeed assess change (Reynolds & Kamphaus, 2009). The relationships between similar content areas for the Progress Monitor and other measures are strong; for example, correlations with the Behavioral Symptoms Index of the BASC–2 ranged from .68 to .91.

Methods of Assessment

Once the purpose of the behavioral, social, or emotional assessment is defined, psychologists still have many options for the particular instrument, method, or procedure they select to use. One factor that might influence the specific tool chosen is the method of assessment. Depending on the setting of interest, data desired, and available informants, one method of assessment may be deemed superior to another. In this section of the chapter, three common methods of behavioral, social, and emotional assessment in school psychology are described: rating scales, interviews, and behavioral observations. Examples are provided after the description of each method of assessment.

Rating scales. Behavior rating scales are forms that combine behavioral statements, which allow a respondent to rate the student's behavioral, social, and emotional functioning in a standardized format. The method can be traced to the 1950s, when rating scales were developed for use by hospital staff to rate the adjustment of psychiatric patients (Frick et al., 2009). By reviewing a number of statements on a completed rating scale (often ranging across 3 to 5 points with response anchors such as *never* and *always*), a psychologist can gain summary information about the perceptions a third-party rater holds about the student's socioemotional and behavioral adjustment. Behavior rating scales can be used for a wide variety of purposes, including screening, diagnosis, progress monitoring, and more general information gathering as part of an assessment plan.

Behavior rating scales are a popular means of assessment (Merrell, 2008) and tend to be a quick, inexpensive way to gather information about students. Moreover, most rating scales do not require specific training in their administration and scoring, increasing their practicality even further. Rating scales allow psychologists and others to learn about infrequent behaviors that might be difficult to observe directly. Also, the practitioner often has a choice of informants to complete a rating scale, such that experts on the student's behavior in different contexts have the opportunity to provide information in the assessment.

Despite their several strengths, rating scales also have notable limitations. It is important to recognize that rating scales depend on the informants' perceptions of behavior and may therefore be subject to problems with bias and subjectivity (Merrell, 2008; O'Donnell & Frick, 2009). Also, rating scales may be constructed on the basis of different models; therefore, although many rating scales measure similar constructs, operationalization of those constructs might vary from one scale to another. When

selecting a particular rating scale, understanding how the scale was constructed and for what purposes is critical; for example, one rating scale on inattention might have been constructed on the basis of *DSM–IV–TR* diagnoses of ADHD, whereas another scale might have been based on a theoretical model of inattention as part of a larger domain of problems in the school setting. Below, three commonly used behavioral and socioemotional rating scales are described.

Achenbach System of Empirically Based Assessment. The ASEBA is a system of assessments including the parent-report Child Behavior Checklist, the teacher-report Teacher Report Form, and the self-report Youth Self-Report forms to rate child and adolescent behavior (Achenbach & Rescorla, 2000, 2001). The ASEBA has been used extensively in dozens of countries and has translations in 69 languages (Achenbach & Rescorla, 2001). The rating scales were first available in preschool- and school-age versions, with versions later developed for adult use (Achenbach & Rescorla, 2004). Each form includes 113 items, and the examiner's manual suggests that the scale requires 10 to 15 minutes to complete. Computer-scoring software is available, which converts raw scores into standardized scores based on appropriate developmental and gender-based norms. Internalizing, externalizing, social problems, and attention problems are included in the ASEBA, resulting in a Total Problems score and Internalizing and Externalizing composite scores. The norms for the ASEBA have been based on large, representative samples of children and adolescents in the United States (Achenbach & Rescorla, 2001). Support for the ASEBA scales has been gathered extensively in the United States and across the globe (Achenbach, Rescorla, & Ivanova, 2005). Finally, adequate reliability and validity information is provided in the manual, with test–retest reliabilities mainly ranging from .75 to .90 and correlations with the Conners–3 (Conners, 2008) in the .71 to .89 range (Achenbach & Rescorla, 2001). Because of the ASEBA's large research base and practicality, it became the first system of child behavioral rating scales to gain wide acceptance by psychologists and trainers of school psychologists in the 1990s. The creation of the ASEBA effectively moved the field of child behavioral and emotional assessment toward dominance by the rating scale methodology.

Behavior Assessment System for Children—2. The BASC–2 is a system of assessments including the parent-report Parent Rating Scales, the teacher-report Teacher Rating Scales, and the Self-Report of Personality forms to rate child and adolescent behavior (Reynolds & Kamphaus, 2004). The BASC–2 is available in both English and Spanish versions. Developmentally appropriate forms are available for preschool, childhood, and adolescence. Each form includes 100 to 200 items and requires 15 to 25 minutes to complete. Computer-scoring software is available that converts raw scores into standardized *T* scores and percentile ranks. Internalizing, externalizing, school problems, and adaptive skills are included in the BASC–2, resulting in a Behavioral Symptoms Index and individual scale scores. The inclusion of adaptive skills information is a particular strength of the BASC–2, such that student strengths and difficulties can be considered in tandem. The content of the items and the scales themselves vary across informant, so the informant used might depend on the information desired. The norms for the BASC–2 have been based on large, representative samples of children and adolescents in the United States (Reynolds & Kamphaus, 2004). Reliability, validity, and factor-analytic support for the BASC–2 are presented in the system's manual (Reynolds & Kamphaus, 2003).

Conners–3. The Conners–3 (Conners, 2008) is the most recent revision and expansion of this widely used behavior rating scale system. This form places particular emphasis on the assessment of externalizing problems in that it also provides *DSM–IV–TR* Symptoms scales for the identification of the externalizing disorders (i.e., three ADHD subtypes, oppositional defiant disorder, and conduct disorder) and a brief ADHD Index. The Long Form contains 110 items, the Short Form contains 45 items, and a 10-item Global Index form is also available. The Conners–3 takes 10 to 20 minutes to complete, depending on which form is used. Screening items for depression and anxiety are also included, as are items for assessing impairment in home, school, and social relationships. The Conners–3 also offers

critical items such as those on the BASC–2 (Frick et al., 2009), a couple of open-ended questions such as those on the ASEBA, and three validity scales: Positive Impression, Negative Impression, and an Inconsistency Index. Other features of the scale include a Spanish translation, hand-scoring and computer-scoring options, computation of norm-referenced scores separately by age groups, and linear derivation of *T* scores.

The Conners–3 parent form uses a four-choice item response format on which 0 = *not at all true* (*never, seldom*) and 4 = *very much true* (*very often, very frequently*). The parent rating form, designed for use with children ages 6 through 18, provides five scales: Hyperactivity/Impulsivity, Executive Functioning, Learning Problems, Aggression, and Peer Relations. The teacher form of the Conners–3 is very similar to the parent form in length and scale content and features the same response format. The Conners–3 also includes a self-report rating scale for ages 8 through 18 that is shorter overall, consisting of 59 items. All three rating scales include short forms of about 40 items.

Separate norms for boys and girls are provided, but a general national norm sample is not. The norming samples for the three forms vary somewhat in their match to U.S. Census Bureau estimates. Reliability estimates for most scales and subscales are good, typically higher than .80. Estimates for some scales are higher than .90, and others are in the .70s. Because of the large number of derived scores offered, any Conners–3 user would be well advised to study the individual scale reliabilities carefully before drawing score inferences. Evidence of factorial, criterion-related, and known-groups validity is provided, all of which require equally careful study before interpreting scores.

Overall, the Conners–3 is a significant and important improvement over its prior version. The Conners–3 is now a comprehensive system of multi-informant instruments like the ASEBA and BASC–2. Some concern exists regarding all items being negatively worded, which could introduce response sets or ill feelings on the part of the rater, and the limited assessment of adaptive competencies and internalizing problems, although specialized long forms are available for this purpose (Frick et al., 2009).

Interviews. Face-to-face interviewing has been a popular approach for gathering information regarding the behavioral, social, and emotional functioning of students both historically and in recent years (Watkins, Campbell, Nieberding, & Hallmark, 1995). Interviews allow the assessor to ask questions about a variety of areas directly; therefore, interviews can be used for purposes ranging from background data collection to diagnosis. Interviews vary in the amount of predetermined structure they provide the assessor. Unstructured interviews are more flexible and permit the interviewer to tailor questions and follow-up probes to fit the responses of the person being interviewed. This method is more dependent on the expertise of the interviewer and may be more subject to biases; therefore, the flexibility of unstructured interviews should be weighed against the potential for unreliability (McClellan & Werry, 2000). Structured interviews, however, provide the interviewer with a predetermined set of questions to be asked, often in a particular sequence, and with a specific rubric for scoring the responses. In structured interviews, follow-up questions are commonly contingent on previous responses; therefore, the interviews often take longer to administer when more areas of concern arise (O'Donnell & Frick, 2009). Because of the complexity of some structured interviews, computerized administrations have become more commonplace (Loney & Frick, 2003). An example of such a structured interview, the DISC–IV, is described next.

Diagnostic Interview Schedule for Children, fourth edition. As mentioned in a previous section, the DISC–IV is a structured interview designed to coincide with the major classifications of the *DSM–IV–TR* that are applicable to children ages 9 to 17 (Fisher et al., 1992; Shaffer et al., 2000). Child and parent versions are available; each includes between 200 and 300 items and takes approximately 1 hour for administration. The DISC–IV results in scores in 27 areas that correspond to *DSM–IV–TR* classifications (American Psychiatric Association, 2000). Because of its highly structured format, the DISC–IV requires little training time for administration and scoring. The Voice DISC is also available, which allows the interview to be conducted via computer terminal, completely eliminating the need for a

trained interviewer. However, the assessor would still need to be familiar with the appropriate interpretation of scores from the Voice DISC. Adequate reliability and validity information has been found in several studies (e.g., Costello, Edelbrock, Duncan, & Kalas, 1984; Edelbrock & Costello, 1988) as well as being provided in the manual (Shaffer et al., 2000).

CURRENT CHALLENGES AND FUTURE DIRECTIONS IN ASSESSMENT

The final section of this chapter moves from what is available in terms of behavioral, social, and emotional assessment to a critical discussion of three current challenges that are likely to be the focus of assessment research efforts in the upcoming years. In particular, this section provides information concerning choice of informant in assessment, multiple-gating systems of assessment, and diversity issues related to assessment as well as potential future directions to better address these challenges.

Choice of Informant

As stated previously, school psychologists often have several options regarding the informant who will provide the information, which may present a challenge when deciding which and how many informants to include in an assessment. Three common informants are parents, teachers, and the student him- or herself; however, alternate informants such as peers, teacher's aides, and others might be included in the assessment. Many popular systems of assessment offer forms for multiple informants, including the ASEBA, BASC–2, and Conners–3 systems reviewed earlier. Although the "more is always better" stance dominates contemporary thinking (Jensen et al., 1999; Power et al., 1998; Verhulst, Dekker, & van der Ende, 1997), the identification of particular informants and the number of informants desired depends highly on the setting or settings of the focal problem, each informant's knowledge regarding the problem, the type of problem being assessed, and developmental considerations.

Considerable research regarding choice of informants for behavioral, social, and emotional screening tests remains conflicting in the sense that all raters show evidence of validity under some conditions (VanDeventer & Kamphaus, in press). Although the belief that the inclusion of more informants provides the optimal amount of information is common, little empirical evidence has supported combining raters to make a classification decision (Johnston & Murray, 2003; McFall, 2005). For example, several studies (Biederman, Keenan, & Faraone, 1990; Lochman, 1995) have found that adding another informant added little variance to the identification process beyond that provided by the first informant. Jones, Dodge, Foster, and Nix (2002) concluded, similarly, that the effect of combining parent and teacher ratings was equal to or minimally better than that of the teacher-only rating. However, some evidence has also supported using combinations of informants. Goodman, Ford, Corbin, and Meltzer (2004) found that prediction was best when both caregiver and teacher ratings were combined. In addition, Kerr, Lunkenheimer, and Olson (2007) found that mothers', fathers', and teachers' ratings of externalizing problems in preschool each added significant incremental validity to the prediction of problems 3 years later. Therefore, no consensus exists on the number of informants and the type of informants that should be included in the assessment process, suggesting that this topic needs additional research (Johnston & Murray, 2003).

A lack of consistency often exists among raters, as evidenced by weak to moderate correlations at best (Achenbach, McConaughy, & Howell, 1987; Kerr et al., 2007), suggesting that perhaps multiple raters provide different yet valuable information. Agreement tends to be even lower when rating internalizing problems compared with externalizing behavior, perhaps because of the internal nature of these difficulties (Glaser, Krosnoble, & Forkner, 1997; Grietens et al., 2004). Recent research by Mattison, Carlson, Cantwell, and Asarnow (2007) has supported earlier findings, providing evidence that teachers rate externalizing and internalizing problems as accurately as parents in a study distinguishing between teacher and parent ratings of children diagnosed with depression, children diagnosed with ADHD, and children with no diagnosis.

In addition to parents and teachers being important sources of information, some evidence has

supported the use of child and adolescent self-report measures of externalizing and, even more strongly, internalizing symptomatology (Grills & Ollendick, 2003; Loeber, Green, & Lahey, 1990). Although self-reports among young children may not be developmentally appropriate because of the cognitive demands involved (e.g., Canino, Bird, Rubio-Stipec, & Bravo, 1995), self-reports from older children often function similarly to reports from adult informants (Achenbach, 2006). In fact, self-report assessments have been recommended as instruments of choice for middle and high school–aged students (Glover & Albers, 2007; Levitt et al., 2007). Overall, the best choice of informant given the areas of interest and student's developmental age and how to combine ratings from multiple informants are issues that deserve further attention (Renk, 2005).

Multiple Gating

Child psychological assessment is increasingly a multisession or multistage process, as exemplified by the recent introduction of RtI models that include progress monitoring as integral to the assessment process. An assessment plan that follows the multiple-gating approach includes assessments based on the levels of a public health or RtI model of service provision, reviewed previously. A multiple-gate identification procedure begins with first-gate universal screening of an entire population for behavioral, social, and emotional risk. At the second gate, those students identified by the screening instrument as being at risk for behavioral, social, and emotional problems are then assessed in a selected assessment procedure using a different assessment, such as an omnibus behavior rating scale (Kamphaus & Reynolds, 2007) or a screener completed by another informant (August, Realmuto, Crosby, & MacDonald, 1995). Students who are identified by the second-gate assessment as having behavioral, social, or emotional problems would then receive a more individualized or comprehensive indicated assessment as a third gate, which may be used to inform specific intervention or diagnostic decisions. This type of procedure has been shown to increase identification and diagnostic accuracy and serves to reduce costs through better identification of students in need (Hill, Lochman, Coie, & Greenberg, 2004; Lochman, 1995; Walker & Severson, 1990).

The number of stages and the amount of time and training required to implement each stage effectively highlight the practical limitations of multiple-gating approaches. School psychologists should consider their resources, in terms of time and monetary costs, before implementing a full multiple-gating system. Whether multiple screening gates are, in fact, superior to single-stage screening is still unknown, and the optimal number of gates to be used is undecided. In one study, the addition of an omnibus rating scale at the second gate significantly improved identification accuracy compared with the first-gate assessment alone (VanDeventer, 2007). However, other studies have found that brief screening instruments perform just as well as three-stage, multiple-gating systems (e.g., Lane et al., 2009). In addition, these questions require further research because many screening studies have estimated the effect of using multiple gates rather than implementing a real-world multiple-gate screening procedure (VanDeventer & Kamphaus, in press), likely because of the investment of time and monetary resources required to carry out a full multiple-gating assessment. On the basis of the evidence available, both universal screening at the first gate and use of an omnibus rating scale or diagnostic interview at the second would be good starting points for implementing preventive and selected interventions.

An example of a multiple-gating assessment procedure is the SSBD (Walker & Severson, 1992). The SSBD is a three-stage procedure including teacher nominations, teacher ratings, and classroom observations. The first gate requires teachers to nominate the top 10 students in internalizing and externalizing problems in their classroom. On the basis of the teacher's perceptions of the severity of the students' symptoms, the top three students in each category receive the second-gate assessment. At the second gate, teachers complete two behavior rating scales for each of these students that gather information about the behaviors of concern and their frequency. Any students that exceed specified cut points on these two instruments proceed to the third gate. At the third gate, those students who scored above

these cut points are observed across school contexts, including the classroom and the playground. Decisions such as further assessment, intervention, and referral to special education are then informed by these third-gate observations as well as by the other data collected through the entire SSBD process. Information on reliability and other psychometric findings is presented in the manual (Walker & Severson, 1992).

Assessment and Student Diversity

Cross-cultural researchers have long been concerned about whether respondents from different cultures interpret a measure in a conceptually similar manner, with many studies conducted on intelligence testing and ethnic group bias (Kim, Kim, & Kamphaus, 2010). The assessment of behavioral, social, and emotional functioning, however, is behind the field of intelligence testing in this regard (Dana, 1996; Merrell, 2008). For example, in the United States many assessments have been developed on the basis of predominantly European Caucasian, English-speaking samples (Padilla, 2001). The *Standards for Educational and Psychological Testing* published by American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999) defined standards of practice that require culturally appropriate assessment methods and interpretations of results. Although test developers and researchers have made some progress toward the goal of culturally appropriate assessment, this goal deserves a great deal more attention.

Padilla (2001) listed three ways in which behavioral, social, and emotional assessments might be biased against or toward a particular group. First, the content or construction of the assessment might be designed in a way that is advantageous to one group. Second, incidental features of the assessment such as formatting, the method of assessment, or characteristics of the person conducting the assessment might favor one group. Finally, the application or interpretation of the assessment might be manipulated in a way that provides an advantage to a specific group. The second and third ways in which bias may be introduced can often be addressed by the school psychologist conducting the assessment by considering how the delivery and interpretation of the assessment might influence responses depending on cultural factors. The first type of bias, however, requires careful development and examination of assessment measures before they are considered appropriate for use across diverse cultural, linguistic, gender, and socioeconomic groups.

When examining the psychometric properties of behavior rating scales and other forms of behavioral, social, and emotional assessment, developers should include analyses examining group equivalence (Tyson, 2004). If the measurement invariance of an assessment of behavioral, social, and emotional problems across linguistic, ethnic, gender, and socioeconomic groups is not verified, any mean-level group differences that are detected may not be meaningful because of potential differences in measurement properties that vary across these groups. Therefore, it is essential that the psychometric properties of such assessments be examined within and across the groups of interest before using these assessments to make decisions for individuals who may fall into groups that were underrepresented in the assessment's development. Differential item functioning is shown when characteristics of an individual item vary across members of subgroups who have similar mean levels of the latent trait and therefore the condition measurement invariance is not upheld across these subgroups (Bond & Fox, 2007). Interpretations should be made with caution when measures are variant, or not equivalent, across the groups of interest; in fact, revisions of the assessment to increase its appropriateness for different groups should be considered.

In terms of results across gender groups, some evidence has suggested that the measurement of young girls' and boys' behavior by teachers and parents is invariant. Konold, Walthall, and Pianta (2004) administered the Child Behavior Checklist for Ages 4 to 18 and the Teacher Response Form to teachers, mothers, and fathers whose children were ages 54 months and again when the children were in first grade. They found the factor structure for the scales to be invariant for both gender groups across informants and development. Measurement invariance by gender was also examined in an

investigation of the aggression subscales of the BASC–2 (Kim et al., 2010) using both confirmatory factor analysis and item response theory. Confirmatory factor analysis results showed that there was not enough evidence to support the measurement invariance of the aggression scales across gender at a scale level; item response theory results found that only a few items were significantly different across gender groups. In another study of measurement invariance of an aggression scale by gender, Tomada and Schneider (1997) reported that the measure used in that investigation was not fully invariant across gender. These two studies provided evidence for the importance of establishing measurement invariance across gender before drawing conclusions from an assessment, because the measures might not assess the same construct for boys and girls.

Ethnic differences in behavior rating scales, which are often used for screening purposes in schools, have consistently been documented. Although some have suggested that there are true mean-level differences across these groups (see Epstein, March, Conners, & Jackson, 1998), others have pointed to differential measurement functioning as the cause for observed differences. For example, Reid et al. (1998) tested the equivalence of an ADHD rating scale for Caucasian versus African American male students. Although mean ADHD scores were higher for the African American students, the psychometric functioning of the scale varied across groups, suggesting a lack of measurement equivalence. Therefore, these results provided evidence for the importance of evaluating the appropriateness and validity of the measurement tool being used within the population of interest.

Although the translation of behavioral, social, and emotional assessments into different languages is a first step to culturally appropriate assessment, variations in cultural experiences and acculturation may lead to continued differences in the interpretations of and performance on these assessments (Padilla, 2001; Sperber, Devellis, & Boehlecke, 1994). In a recent investigation examining the measurement invariance of the BESS Teacher form (Kamphaus & Reynolds, 2007) for 142 limited-English-proficient and 110 English-proficient students, the majority of screening items were found to be invariant across language proficiency groups on the basis of item response theory analyses (Dowdy, Dever, DiStefano, & Chin, 2011). The Dowdy et al. (2011) study provided some evidence to suggest that at least partial measurement invariance is likely to be found for teacher screeners across students of different language proficiency groups; however, these findings need to be replicated across various samples of culturally diverse students.

The studies reviewed here provide a mere glimpse of the research being conducted to examine the measurement equivalence of behavioral, social, and emotional assessments across diverse groups of students. Although great work is already being done in this area, it is clear that much more research is necessary to provide information regarding the functioning of such assessments when used with various subgroups within the U.S. population and across the globe. It is essential for school psychologists to understand and seek information about the appropriateness of various assessments, and the interpretation of such assessments, for the focal groups of interest to that psychologist. Future research on measurement invariance and related issues would provide a great service to the field and the increasingly diverse students it serves.

SUMMARY AND CONCLUSION

Child behavioral, social, and emotional assessment practices have changed dramatically over the past 2 to 3 decades, with a notable increase in use of behavioral assessment techniques such as rating scales (Shapiro & Heick, 2004). In contrast, intellectual assessment practices have remained largely the same in that the Wechsler Scales, their derivatives, and imitators still hold sway as they have done since World War II. Largely gone from the schools are the inkblot, storytelling (thematic), drawing, and related projective techniques that dominated the psychological assessment practices with children in and out of school settings between World War II and Achenbach's publication of his first Child Behavior Checklist in 1981 (Frick et al., 2009). In fact, one would be hard pressed to find a child's formal assessment file in school these days that did not include at least one or more ASEBA,

BASC–2, Conners, or similar rating scale or self-report form.

In addition, the topics included in this chapter are far different than was typical 20 years ago when the emphasis was on making diagnostic, classification, and special education eligibility decisions. As shown in this chapter, there is considerable momentum in the direction of prevention and early intervention assessment services such as screening and progress monitoring. Indeed, school psychologists' assessment practices have changed, and for the better.

However, practical and psychometric challenges are enduring. There will always be a need for "faster, better, cheaper" assessment tools and methods as more is demanded of schools, children, and the psychologists who serve them. New disorders and subtypes (the three ADHD subtypes are one example) are constantly being considered, as are new constructs such as life satisfaction. This expansion requires psychologists to expand their assessment batteries accordingly, putting additional pressures on testing time. Thus, there will be a continuing need to attend to practicalities such as reducing test length while simultaneously enhancing reliability and validity evidence.

The evidence base for these modern child assessment practices remains meager. Whether screening practices lead to significant improvements in child well-being long term is not yet known. The relationship between child behavioral, social, and emotional adjustment and academic outcomes is still debated. In addition, the comparative personnel costs of public health or RtI service delivery models are not yet fully understood. Progress in the area of child behavioral and emotional assessment is being made; however, the profession must prove that continuing efforts are successful.

References

Achenbach, T. M. (2005). Advancing assessment of children and adolescents: Commentary on evidence-based assessment of child and adolescent disorders. *Journal of Clinical Child and Adolescent Psychology, 34*, 541–547. doi:10.1207/s15374424jccp3403_9

Achenbach, T. M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science, 15*, 94–98. doi:10.1111/j.0963-7214.2006.00414.x

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213–232. doi:10.1037/0033-2909.101.2.213

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms and profiles*. Burlington: University of Vermont.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms and profiles*. Burlington: University of Vermont.

Achenbach, T. M., & Rescorla, L. A. (2004). The Achenbach System of Empirically Based Assessment (ASEBA) for ages 1.5 to 18 years. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment: Vol. 2. Instrument for children and adolescents* (pp. 179–213). Mahwah, NJ: Erlbaum.

Achenbach, T. M., Rescorla, L. A., & Ivanova, M. Y. (2005). International cross-cultural consistencies and variations in child and adolescent psychopathology. In C. L. Frisby & C. R. Reynolds (Eds.), *Comprehensive handbook of multicultural school psychology* (pp. 674–709). Hoboken, NJ: Wiley.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological assessment* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Atkins, M. S., Frazier, S. L., Adil, J. A., & Talbott, E. (2003). School-based mental health services in urban communities. In M. Weist, S. Evans, & N. Lever (Eds.), *Handbook of school mental health: Advancing practice and research* (pp. 165–178). New York, NY: Kluwer Academic/Plenum.

August, G. J., Realmuto, G. M., Crosby, R. D., & MacDonald, A. W., III. (1995). Community-based multiple-gate screening of children at risk for conduct disorder. *Journal of Abnormal Child Psychology, 23*, 521–544. doi:10.1007/BF01447212

Bandura, A. (1969). *Principles of behavior modification*. New York, NY: Holt, Rinehart, & Winston.

Belfer, M. L. (2008). Child and adolescent mental disorders: The magnitude of the problem across the globe. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 49*, 226–236. doi:10.1111/j.1469-7610.2007.01855.x

Biederman, J., Keenan, K., & Faraone, S. V. (1990). Parent-based diagnosis of attention-deficit disorder predicts a diagnosis based on teacher report. *Journal of the American Academy of Child and Adolescent Psychiatry, 29*, 698–701. doi:10.1097/00004583-199009000-00004

Blashfield, R. K. (1998). Diagnostic models and systems. In A. A. Bellack, M. Hersen, & C. R. Reynolds (Eds.), *Comprehensive clinical psychology: Vol. 4. Assessment* (pp. 57–80). New York, NY: Elsevier Science.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Bracken, B. A., & Keith, L. K. (2004). *Professional manual for the Clinical Assessment of Behavior*. Lutz, FL: Psychological Assessment Resources.

Bradley, R., Doolittle, J., & Bartolotta, R. (2008). Building on the data and adding to the discussion: The experiences and outcomes of students with emotional disturbance. *Journal of Behavioral Education, 17*, 4–23. doi:10.1007/s10864-007-9058-6

Canino, G., Bird, H. R., Rubio-Stipec, M., & Bravo, M. (1995). Child psychiatric epidemiology: What we have learned and what we need to learn. *International Journal of Methods in Psychiatric Research, 5*, 79–92.

Catalano, R. F., Haggerty, K. P., Oesterle, S., Fleming, C. B., & Hawkins, J. D. (2004). The importance of bonding to school for healthy development: Findings from the social development research group. *Journal of School Health, 74*, 252–261. doi:10.1111/j.1746-1561.2004.tb08281.x

Conners, C. K. (2008). *Conners (third edition)*. Toronto, Ontario, Canada: Multi-Health Systems.

Costello, E. J., Edelbrock, C. S., Duncan, M. K., & Kalas, R. (1984). *Testing of the NIMH Diagnostic Interview Schedule for Children (DISC) in a clinical population*. Pittsburgh, PA: University of Pittsburgh, Department of Psychiatry.

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G., & Angold, A. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry, 60*, 837–844. doi:10.1001/archpsyc.60.8.837

Dana, R. H. (1996). Culturally competent assessment practice in the United States. *Journal of Personality Assessment, 66*, 472–487. doi:10.1207/s15327752jpa6603_2

Dawber, T. R., & Kannel, W. B. (1966). The Framingham study: An epidemiological approach to coronary heart disease. *Circulation, 34*, 553–555. PMid:5921755.

DiStefano, C. A., & Kamphaus, R. W. (2007). Development and validation of a behavioral screening for preschool-age children. *Journal of Emotional and Behavioral Disorders, 15*, 93–102. doi:10.1177/10634266070150020401

Dowdy, E., Dever, B. V., DiStefano, C., & Chin, J. (2011). Screening for emotional and behavioral risk among students with Limited English Proficiency. *School Psychology Quarterly, 26*, 14–26. doi:10.1037/a0022072

Dowdy, E., Mays, K. L., Kamphaus, R. W., & Reynolds, C. R. (2009). Roles of diagnosis and classification in school psychology. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (4th ed., pp. 191–209). Hoboken, NJ: Wiley.

Drummond, T. (1994). *The Student Risk Screening Scale (SRSS)*. Grants Pass, OR: Josephine County Mental Health Program.

Durlak, J. A. (1997). *Successful prevention programs for children and adolescents*. New York, NY: Plenum Press.

Edelbrock, C., & Costello, A. J. (1988). Structured psychiatric interviews for children. In M. Rutter, A. H. Tuma, & I. S. Lann (Eds.), *Assessment and diagnosis in child psychopathology* (pp. 87–112). New York, NY: Guilford Press.

Epstein, J. N., March, J. S., Conners, C. K., & Jackson, D. L. (1998). Racial differences on the Conners Teacher Rating Scale. *Journal of Abnormal Child Psychology, 26*, 109–118. doi:10.1023/A:1022617821422

Fisher, P., Wicks, J., Shaffer, D., Piacentini, J., & Lapkin, J. (1992). *NIMH Diagnostic Interview Schedule for Children user's manual*. New York, NY: State Psychiatric Institute.

Flanagan, K. S., Bierman, K. L., & Kam, C. M. (2003). Identifying at-risk children at school entry: The usefulness of multibehavioral problem profiles. *Journal of Clinical Child and Adolescent Psychology, 32*, 396–407. doi:10.1207/S15374424JCCP3203_08

Frick, P. J., Barry, C., & Kamphaus, R. W. (2009). *Clinical assessment of child and adolescent personality and behavior* (2nd ed.). New York, NY: Springer.

Fuchs, D., Fuchs, L., & Compton, D. (2004). Identifying reading disabilities by responsiveness-to-instruction: Specifying measures and criteria. *Learning Disability Quarterly, 27*, 216–227. doi:10.2307/1593674

Glaser, B. A., Krosnoble, K. M., & Forkner, C. B. W. (1997). Parents and teachers as raters of children's problem behaviors. *Child and Family Behavior Therapy, 19*, 1–13. doi:10.1300/J019v19n04_01

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*, 117–135. doi:10.1016/j.jsp.2006.05.005

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 38*, 581–586. doi:10.1111/j.1469-7610.1997.tb01545.x

Goodman, R. (1999). The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 40*, 791–799. doi:10.1111/1469-7610.00494

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 1337–1345. doi:10.1097/00004583-200111000-00015

Goodman, R., Ford, T., Corbin, T., & Meltzer, H. (2004). Using the Strengths and Difficulties Questionnaire (SDQ) multi-informant algorithm to screen looked-after children for psychiatric disorders. *European Child and Adolescent Psychiatry, 13*, 1125–1131.

Goodman, R., & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: Is small beautiful? *Journal of Abnormal Child Psychology, 27*, 17–24. doi:10.1023/A:1022658222914

Gresham, F. M. (1991). Conceptualizing behavior disorders in terms of resistance to intervention. *School Psychology Review, 20*, 23–36.

Gresham, F. M. (2002). Responsiveness to intervention: An alternative approach to the identification of learning disabilities. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Learning disabilities: Research to practice* (pp. 467–519). Mahwah, NJ: Erlbaum.

Grietens, H., Onghena, P., Prinzie, P., Gadeyne, E., Van Assche, C., Ghesquiere, P., & Hellinckx, W. (2004). Comparison of mothers', fathers', and teachers' reports on problem behavior in 5- to 6-year-old children. *Journal of Psychopathology and Behavioral Assessment, 26*, 137–146.

Grills, A. E., & Ollendick, T. H. (2003). Multiple informant agreement and the Anxiety Disorders Interview Schedule for parents and children. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*, 30–40. doi:10.1097/00004583-200301000-00008

Gutman, L. M., Sameroff, A. J., & Cole, R. (2003). Academic growth curve trajectories from 1st grade to 12th grade: Effects of multiple social risk factors and preschool child factors. *Developmental Psychology, 39*, 777–790. doi:10.1037/0012-1649.39.4.777

Hill, L. G., Lochman, J. E., Coie, J. D., & Greenberg, M. T.; Conduct Problems Prevention Research Group. (2004). Effectiveness of early screening for externalizing problems: Issues of screening accuracy and utility. *Journal of Consulting and Clinical Psychology, 72*, 809–820. doi:10.1037/0022-006X.72.5.809

Horner, R. H., & Sugai, G. (2000). School-wide behavior support: An emerging initiative. *Journal of Positive Behavior Interventions, 2*, 231–233. doi:10.1177/109830070000200407

Hysing, M., Elgen, I., Gillberg, C., Lie, S. A., & Lundervold, A. J. (2007). Chronic physical illness and mental health in children: Results from a large-scale population study. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 48*, 785–792. doi:10.1111/j.1469-7610.2007.01755.x

Individuals With Disabilities Education Act of 1997, Pub. L. 105–117.

Individuals With Disabilities Education Improvement Act of 2004, Pub. L. No. 108–446, 20 U.S.C. § 1400 *et seq.*

Jamieson, K., & Romer, D. (2005). A call to action on adolescent mental health. In D. L. Evans, E. B. Foa, R. E. Gur, H. Hendin, C. P. O'Brien, M. E. P. Seligman, & B. T. Walsh (Eds.), *Treating and preventing adolescent mental health disorders: What we know and what we don't know: A research agenda for improving the mental health of our youth* (pp. 598–615). New York, NY: Oxford University Press. doi:10.1093/9780195173642.001.0001

Jensen, P. S., Rubio-Stipec, M., Canino, G., Bird, H. R., Dulcan, M. K., Schwab-Stone, M. E., & Lahey, B. B. (1999). Parent and child contributions to diagnosis of mental disorder: Are both informants always necessary? *Journal of the American Academy of Child and Adolescent Psychiatry, 38*, 1569–1579. doi:10.1097/00004583-199912000-00019

Johnston, C., & Murray, C. (2003). Incremental validity in the psychological assessment of children and adolescents. *Psychological Assessment, 15*, 496–507. doi:10.1037/1040-3590.15.4.496

Jones, D., Dodge, K. A., Foster, E. M., & Nix, R. (2002). Early identification of children at risk for costly mental health service use. *Prevention Science, 3*, 247–256. doi:10.1023/A:1020896607298

Kamphaus, R. W., Dowdy, E., Kim, S., & Chen, J. (in press). Diagnosis, classification, and screening systems. In C. R. Reynolds (Ed.), *Oxford handbook of psychological assessment of children and adolescents*. New York, NY: Oxford University Press.

Kamphaus, R. W., & Reynolds, C. R. (2007). *BASC–2 Behavioral and Emotional Screening System.* Minneapolis, MN: Pearson Assessment.

Kamphaus, R. W., Thorpe, J. S., Winsor, A. P., Kroncke, A. P., Dowdy, E. T., & VanDeventer, M. (2007). Development and predictive validity of a teacher screener for child behavioral and emotional problems at school. *Educational and Psychological Measurement, 67*, 342–356. doi:10.1177/0013164407067002 1001

Kauffman, J. M. (2000). *Characteristics of emotional and behavioral disorders of children and youth* (7th ed.). Columbus, OH: Merrill/Prentice-Hall.

Kerr, D. C. R., Lunkenheimer, E. S., & Olson, S. L. (2007). Assessment of child problem behavior by multiple informants: A longitudinal study from preschool to school entry. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 48*, 967–975. doi:10.1111/j.1469-7610.2007.01776.x

Kim, S., Kim, S. H., & Kamphaus, R. W. (2010). Is aggression the same for boys and girls? Assessing measurement invariance with confirmatory factor analysis and item response theory. *School Psychology Quarterly, 25*, 45–61. doi:10.1037/a0018768

Konold, T. R., Walthall, J. C., & Pianta, R. C. (2004). The behavior of child behavior ratings: Measurement structure of the Child Behavior Checklist across time, informants, and child gender. *Behavioral Disorders, 29*, 372–383.

Lane, K. L., Carter, E. W., Pierson, M. R., & Glaeser, B. C. (2006). Academic, social, and behavioral characteristics of high school students with emotional disturbances or learning disabilities. *Journal of Emotional and Behavioral Disorders, 14*, 108–117. doi:10.1177/10634266060140020101

Lane, K. L., Little, M. A., Casey, A. M., Lambert, W., Wehby, J., Weisenbach, J. L., & Phillips, A. (2009). A comparison of systematic screening tools for emotional and behavioral disorders. *Journal of Emotional and Behavioral Disorders, 17*, 93–105. doi:10.1177/1063426608326203

Lane, K. L., Menzies, H. M., Oakes, W. P., & Kalberg, J. R. (2012). *Systematic screenings of behavior to support instruction from preschool to high school.* New York, NY: Guilford Press.

Lane, K. L., Parks, R. J., Robertson Kalberg, J., & Carter, E. W. (2007). Systematic screening at the middle school level: Score reliability and validity of the Student Risk Screening Scale. *Journal of Emotional and Behavioral Disorders, 15*, 209–222. doi:10.1177/10634266070150040301

Lane, K. L., Robertson Kalberg, J., Parks, R. J., & Carter, E. W. (2008). Student Risk Screening Scale: Initial evidence for score reliability and validity at the high school level. *Journal of Emotional and Behavioral Disorders, 16*, 178–190. doi:10.1177/1063426608314218

Langenbucher, J., & Nathan, P. E. (2006). Diagnosis and classification. In M. Hersen & J. C. Thomas (Eds.), *Comprehensive handbook of personality and psychopathology* (Vol. 2, pp. 3–20). Hoboken, NJ: Wiley.

Levitt, J. M., Saka, N., Romanelli, L. H., & Hoagwood, K. (2007). Early identification of mental health problems in schools: The status of instrumentation. *Journal of School Psychology, 45*, 163–191. doi:10.1016/j.jsp.2006.11.005

Lochman, J. E.; Conduct Problems Prevention Research, G. (1995). Screening of child behavior problems for prevention programs at school entry. *Journal of Consulting and Clinical Psychology, 63*, 549–559. doi:10.1037/0022-006X.63.4.549

Loeber, R., Green, S. M., & Lahey, B. B. (1990). Mental health professions' perceptions of the utility of children, mothers, and teachers as informants on childhood psychopathology. *Journal of Clinical Child Psychology, 19*, 136–143. doi:10.1207/s15374424jccp1902_5

Loney, B. R., & Frick, P. J. (2003). Structured diagnostic interviewing. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Personality, behavior, and context* (2nd ed., pp. 235–247). New York, NY: Guilford Press.

Mattison, R. E., Carlson, G. A., Cantwell, D. P., & Asarnow, J. R. (2007). Teacher and parent ratings of children with depressive disorders. *Journal of Emotional and Behavioral Disorders, 15*, 184–192. doi:10.1177/10634266070150030501

McClellan, J. M., & Werry, J. S. (2000). Introduction. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 19–27. doi:10.1097/00004583-200001000-00013

McFall, R. M. (2005). Theory and utility—Key themes in evidence-based assessment: Comment on the special section. *Psychological Assessment, 17*, 312–323. doi:10.1037/1040-3590.17.3.312

McIntosh, K., Flannery, K., Sugai, G., Braun, D., & Cochrane, K. (2008). Relationships between academics and problem behavior in the transition from middle school to high school. *Journal of Positive Behavior Interventions, 10*, 243–255. doi:10.1177/1098300708318961

Merrell, K. W. (2008). *Behavioral, social, and emotional assessment of children and adolescents* (3rd ed.). New York, NY: Erlbaum.

Najman, J. M., Heron, M. A., Hayatbakhsh, M. R., Dingle, K., Jamrozik, K., Bor, W., . . . Williams, G. M. (2008). Screening in early childhood for risk of later mental health problems: A longitudinal study. *Journal of Psychiatric Research, 42*, 694–700. doi:10.1016/j.jpsychires.2007.08.002

O'Connell, M., Boat, T., & Warner, K. (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities.* Washington, DC: National Academies Press.

O'Donnell, C. W., & Frick, P. J. (2009). Assessment of personality and adjustment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (4th ed., pp. 287–306). Hoboken, NJ: Wiley.

Padilla, A. M. (2001). Issues in culturally appropriate assessment. In L. A. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment* (2nd ed., pp. 5–27). San Francisco, CA: Jossey-Bass.

Power, T. J., Andrews, T. J., Eiraldi, R. B., Doherty, B. J., Ikeda, M. J., DuPaul, G. J., & Landau, S. (1998). Evaluating attention-deficit/hyperactivity disorder

using multiple informants: The incremental utility of combining teacher with parent reports. *Psychological Assessment, 10*, 250–260. doi:10.1037/1040-3590.10.3.250

Rapport, M. D., Denney, C. B., Chung, K. M., & Hustace, K. (2001). Internalizing behavior problems and scholastic achievement in children: Cognitive and behavioral pathways as mediators of outcome. *Journal of Clinical Child Psychology, 30*, 536–551. doi:10.1207/S15374424JCCP3004_10

Reid, R., DuPaul, G. J., Power, T. J., Anastopoulos, A. D., Rogers-Adkinson, D., Noll, M., & Riccio, C. (1998). Assessing culturally different students for attention-deficit/hyperactivity disorder using behavior rating scales. *Journal of Abnormal Child Psychology*, *26*, 187–198. doi:10.1023/A:1022620217886

Renk, K. (2005). Cross-informant ratings of behavior of children and adolescents: The "gold standard." *Journal of Child and Family Studies, 14*, 457–468. doi:10.1007/s10826-005-7182-2

Reschly, D. J., & Bergstrom, M. K. (2009). Response to intervention. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (4th ed., pp. 434–460). Hoboken, NJ: Wiley.

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment System for Children* (2nd ed.). Circle Pines, MN: American Guidance Service.

Reynolds, C. R., & Kamphaus, R. W. (2009). *BASC–2 progress monitor*. Minneapolis, MN: NCS Pearson.

Ringel, J. S., & Sturm, R. (2001). National estimates of mental health utilization and expenditure for children in 1998. *Journal of Behavioral Health Services and Research, 28*, 319–333. doi:10.1007/BF02287247

Romer, D., & McIntosh, M. (2005). The roles and perspectives of school mental health professionals in promoting adolescent mental health. In D. L. Evans, E. B. Foa, R. E. Gur, H. Hendin, C. P. O'Brien, M. E. P. Seligman, & B. T. Walsh (Eds.), *Treating and preventing adolescent mental health disorders: What we know and what we don't know* (pp. 597–615). New York, NY: Oxford University Press. doi:10.1093/9780195173642.003.0032

Ruchkin, V., Jones, S., Vermeiren, R., & Schwab-Stone, M. (2008). The Strengths and Difficulties Questionnaire: The self-report version in American urban and suburban youth. *Psychological Assessment, 20*, 175–182. doi:10.1037/1040-3590.20.2.175

Saunders, S. M., & Wojcik, J. V. (2004). The reliability and validity of a brief self-report questionnaire to screen for mental health problems: The Health Dynamics Inventory. *Journal of Clinical Psychology in Medical Settings, 11*, 233–241. doi:10.1023/B:JOCS.0000037617.04463.e1

Schmitz, N., Kruse, J., Heckrath, C., Alberti, L., & Tress, W. (1999). Diagnosing mental disorders in primary care: The General Health Questionnaire (GHQ) and the Symptom Check List (SCL–90–R) as screening instruments. *Social Psychiatry and Psychiatric Epidemiology, 34*, 360–366. doi:10.1007/s001270050156

Scotti, J. R., & Morris, T. L. (2000). Diagnosis and classification. In M. Hersen & R. T. Ammerman (Eds.), *Advanced abnormal child psychology* (2nd ed., pp. 15–32). Mahwah, NJ: Erlbaum.

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M., & Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children, Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 28–38. doi:10.1097/00004583-200001000-00014

Shapiro, E. S., & Heick, P. F. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools, 41*, 551–561. doi:10.1002/pits.10176

Sperber, A. D., Devellis, R. F., & Boehlecke, B. (1994). Cross-cultural translation: Methodology and validation. *Journal of Cross-Cultural Psychology, 25*, 501–524. doi:10.1177/0022022194254006

Sugai, G., Horner, R. H., Dunlap, G., Hieneman, M., Lewis, T. J., Nelson, C. M., . . . Ruef, M. (2000). Applying positive behavioral support and functional behavioral assessment in the schools. *Journal of Positive Behavior Interventions, 2*, 131–143. doi:10.1177/109830070000200302

Tomada, G., & Schneider, B. H. (1997). Relational aggression, gender, and peer acceptance: Invariance across culture, stability over time, and concordance among informants. *Developmental Psychology, 33*, 601–609. doi:10.1037/0012-1649.33.4.601

Tyson, E. H. (2004). Ethnic difference in using behavior rating scales to assess the mental health of children: A conceptual and psychometric critique. *Child Psychiatry and Human Development, 34*, 167–201. doi:10.1023/B:CHUD.0000014996.26276.a5

U.S. Public Health Service. (2000). *Report of the Surgeon General's conference on children's mental health: A national action agenda,* Washington, DC: U.S. Department of Health and Human Services. Retrieved from http://www.surgeongeneral.gov/topics/cmh/childreport.htm

VanDeventer, M. C. (2007). *Child mental health screening at school: Disorders, gates, informants, and consequences*. Unpublished doctoral dissertation, University of Georgia, Athens.

VanDeventer, M. C., & Kamphaus, R. W. (in press). *Universal emotional and behavioral screening for*

children and adolescents: Prospects and pitfalls. New York, NY: Springer.

Verhulst, F. C., Dekker, M. C., & van der Ende, J. (1997). Parent, teacher, and self-reports as predictors of signs of disturbance in adolescents: Whose information carries the most weight? *Acta Psychiatrica Scandinavica, 96*, 75–81. doi:10.1111/j.1600-0447.1997.tb09909.x

Wagner, M., Kutash, K., Duchnowski, A., & Epstein, M. (2005). The Special Education Elementary Longitudinal Study and the National Longitudinal Transition Study: Study designs and implications for children and youth with emotional disturbance. *Journal of Emotional and Behavioral Disorders, 13*, 25–41. doi:10.1177/10634266050130010301

Walker, H. M., & Severson, H. (1990). *Systematic Screening for Behavior Disorders (SSBD)*. Longmont, CO: Sopris West.

Walker, H. M., & Severson, H. (1992). *Systematic screening for behavior disorders* (2nd ed.). Longmont, CO: Sopris West.

Watkins, C. E., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26*, 54–60. doi:10.1037/0735-7028.26.1.54

Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41*, 85–120. doi:10.1177/00224669070410020401

Werner, E. E. (1994). Overcoming the odds. *Journal of Developmental and Behavioral Pediatrics, 15*, 131–136. doi:10.1097/00004703-199404000-00012

Witt, J. C., Elliott, S. N., & Gresham, F. M. (Eds.). (1988). *The handbook of behavior therapy in education*. New York, NY: Guilford Press. doi:10.1007/978-1-4613-0905-5

CHAPTER 7

DYNAMIC ASSESSMENT

Carol Robinson-Zañartu and Jerry Carlson

For nearly a century, psychologists working in schools have used standardized intelligence tests to predict educational achievement (Lee & Slaughter-Defoe, 2004). Over time, sociopolitical and educational concerns have led to questions about the validity and usefulness of traditional testing and set the stage for considering the use of dynamic assessment (DA) in school settings.

In contrast to conventional standardized ability testing, DA is designed to examine cognitive change rather than to produce a cognitive measure of relative stability, to make use of intentional examiner–examinee interactions rather than to limit them, and to embed some type of intervention within the assessment rather than to maintain strict standardization without assists. In that context, three questions frame the potential of DA to enhance school psychological practice:

1. Rather than solely examining relatively stable traits for the prediction of achievement, might it be equally or more useful in school settings to assess conditions under which student change occurs, resulting in improved student approaches to learning and problem solving?
2. Because school learning relies on the relationship of teachers or other significant adults with students, might it be useful to assess the effect of imposing deliberate examiner–examinee interactions designed to support enhanced engagement with learning or response to problem solving rather than to deliberately limit the effect of those interactions?
3. Because school psychologists use assessments to inform their teacher recommendations, might it be useful for teachers to know how a student responded to help when he or she produced inadequate responses rather than to know how the student responded to noninterference with student responses to a standardized set of questions?

A fourth issue, cultural fairness, has been central to most forms of DA and treated differently across the range of DA models. The complex question of differentiating cultural difference from disability was the subject of much of Feuerstein's early research (Feuerstein, Rand, & Hoffman, 1979) and continues to be the subject of considerable interest (cf. Budoff, 1987; Green, McIntosh, Cook-Morales, & Robinson-Zañartu, 2005; Hessels, 2000; Lidz & Peña, 1996, 2009; Robinson-Zañartu & Aganza, 2000). A fifth issue is rooted in the emerging literature on the importance of strength-based practice in schools, especially for culturally diverse youths (Clauss-Ehlers, 2004; Edwards, Holtz, & Green, 2007). Might DA, drawing from cognitive strengths embedded within the students' cultural experiences, serve as a strength-based cognitive assessment that could help psychologists to support teachers to assist diverse students in becoming more engaged and successful learners?

In this chapter, the authors address these questions in light of the processes, commonalities, differences, and research relevant to major and emerging practice across a wide range of DA models, from

DOI: 10.1037/14049-007
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

successive cuing to clinical attempts to create permanent change in students' use of cognitive skills (Carlson, 1994, 1995; Feuerstein, Miller, Rand, & Jensen, 1981; Lidz & Elliott, 2000; Sternberg & Grigorenko, 2002). The contexts influencing school psychological assessment relevant to dynamic models are discussed as well as the premises, procedures, and practices involved in several of the major models. The authors begin with a review of the major historical roots of DA.

HISTORICAL ROOTS OF DYNAMIC ASSESSMENT IN SCHOOLS

Although the use of DA in schools per se emerged in school psychology in the second half of the 20th century, attempts to bring about change in cognitive functioning began centuries ago. For instance, Itard's work at the turn of the 19th century with Victor, the young boy who had lived alone in the woods for years, supported his contention that "an enriched environment could compensate for developmental delays" (Indiana University, 2007, para. 8). Close to a decade later, Alfred Binet set up special classrooms in an attempt to improve the cognitive functioning of some 50 French students who were not making progress in regular classrooms. Discovering that his instructional program had produced cognitive gains in some of the youths, Binet became an important early advocate of the notion that intelligence might be modifiable (Finger, 1994; Zazzo, 2000). Dweck (2006) discussed Binet's motivation in test design as his belief that his test might help researchers design programs to assist children to return to regular classrooms. He believed that education and practice could bring about fundamental changes in intelligence. Citing from Binet's (1909) book, *Modern Ideas About Children,* Dweck (2006) selected the following passage to summarize Binet's stance on the modifiability of intelligence:

> A few modern philosophers . . . assert that an individual's intelligence is a fixed quantity, a quantity which cannot be increased. We must protest and react against this brutal pessimism. . . . With practice, training, and above all, method, we manage to increase our attention, our memory, our judgment, and literally to become more intelligent than we were before. (p. 5)

Although a number of researchers throughout the 20th century contributed to the notion that ability could be modified through environmental changes (Lidz, 1987b), attempting to incorporate modifiability into the assessment process was somewhat unique. Early efforts at this emerged with attempts to test the ability to learn while observing learning in progress (e.g., Dearborn 1921; De Weerdt, 1927). Vygotsky's (1935, 1934/1962, 1934/1978) social–interactionist model of intellectual development involved his clinical analyses of children's changeable learning ability. His proposals to observe the results of deliberate stimulation of learning have been considered seminal in the DA movement. Promoted outside the Soviet Union by his colleague Luria, Vygotsky's work influenced some of the important early DA work in the West (e.g., Budoff, 1967; Budoff & Friedman, 1964).

Many school psychologists have identified the introduction of Feuerstein's learning potential assessment device in the book *The Dynamic Assessment of Retarded Performers* (Feuerstein et al., 1979) as the introduction of DA into school psychology (Lidz, 1987a). His attention to cultural factors in the development of what he called *learning potential* and differentiation of functional deficits from those based in deprivation of deliberate and rich transmission of culture owing to factors such as war and poverty, gave rise to considerable interest. As discussed in the following sections, interest in Feuerstein's work by school psychologists in the United States began at a time of sociopolitical controversy over intelligence testing in schools, specifically related to cultural variables. However, both Vygotsky's and Piaget's work must also be considered foundational to school-based DA.

Vygotsky's Social–Interactionist Model

Vygotsky's (1934/1962, 1934/1978) instrumentalist model of cognitive change stresses two major features: (a) Intellectual growth occurs through sociocultural mediation and the history of the child's

personal experiences, and (b) language serves as the primary tool for the development of mental processes and their internalization. Through language and attendant symbol systems, internalization of thought and patterns of action occurs as natural, social, and cultural processes. Whether by formal instruction or informal means, new, previously external codes become internalized codes or schemes. Children's development involves an active dialectical process in which new learning is based on prior knowledge and experience and the interactions of the child with key, that is, particularly relevant, others in the world in which he or she lives. Vygotsky's (1934/1978) theory "presupposes a specific social nature and process by which children grow into the intellectual life of those around them" (p. 88) and implies that assessment of learning should take into consideration not only the product of learning, but the cultural and social circumstances that mediate it. These circumstances include the interactions that children have in everyday life with other children and adults. (For an elaboration of these issues, see Nisbett, Peng, Choi, & Norenzayan 2001, and Rogoff, 1990.)

As children learn, they move through what Vygotsky (1934/1978) termed their *zone of proximal development,* described as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (p. 86). However, the upper boundaries of that zone of proximal development were not conceived of as static but rather as "constantly changing with the learner's increasing independent competence" (Bransford, Brown, & Cocking, 2000, p. 81). Vygotsky (as cited in Minick, 1987) saw that evaluation should not merely examine symptoms and present status but should assume interaction with a significant adult and thereby try to predict the child's course of development, or potential to learn:

> If we are limited to the definition and measurement of symptoms of development, we can never go beyond a purely empirical statement of what is already known about the child. The best we can do is to refine the statement of these symptoms and verify their measurement. We will never, however, be able to explain the phenomena we observe in the child's development, predict the subsequent course of development, or indicate what practical measures must be taken. (Minick, 1987, p. 136)

Vygotsky (as cited in Minick, 1987) recognized the significance of alternative evaluative procedures, which today are called *interactive* or *dynamic assessment.* Commenting on the theoretical significance of interactive approaches to work within the individual's zone of proximal development, he noted the key interactive role between child and mediator and its diagnostic potential:

> When we use this principle of cooperation in establishing the zone of proximal development, we gain the potential for directly studying that which most precisely determines the level of mental maturation that must be completed in the proximal or subsequent period of his age development. (Minick, 1987, p. 136)

Piaget's Model of Cognitive Development and Interactive Clinical Approach to Assessing Cognitive Structures

For Piaget, cognitive development involves development and refinement of mental structures (Flavell, 1963). Cognitive growth proceeds through a series of levels or stages. Similar to Vygotsky, emerging abilities are internalized transformations and develop through an individual's interactions with his or her environment and maturation. Although Piaget (1934/1962) did not stress social interaction as Vygotsky did, he did not exclude it. In fact, along with maturation, the child's actions on objects, and the mediational activity Piaget termed *equilibration,* social transmission was an essential feature of his model.

Piaget's approach to assessing cognitive growth involved an interactive, probing clinical method. For him, assessment was linked to the processes of thinking and development (Piaget, 1967). In that sense and through its interactive approach, it is mediational. Using this approach, Piaget examined

not only the hierarchical growth of a child's mental abilities but also sought to determine the true level of the child's cognitive development. Piaget's belief was that the assessor should interact with the child to determine the current level of performance and explore the potential for the highest level of performance, which was accomplished through verbal interactions and probing between the child and the assessor as well as modifications and elaborations of the materials used in the assessment. Piaget's clinical method (his terminology) was flexible in content and approach; its purpose was to ascertain the highest level that a child had achieved or could achieve. For Piaget, learning and elaborations during the assessment procedure were seen as inextricably linked, not as separate from the child's intellectual processes and structural development.

Feuerstein's Dynamic Assessment of Learning Potential

Feuerstein's approach to the assessment of cognitive abilities paralleled many of Vygotsky's basic assumptions, especially in his emphasis on the modifiability of human functioning and the social contextual factors of influence. Recognizing that "some individuals from different ethnocultural and low socioeconomic subgroups regularly perform below the levels of functioning characteristic of mainstream culture" (Feuerstein et al., 1979, p. 1), Feuerstein set out to determine specific factors in the processes of cultural transmission and mediation that affect and promote cognitive development and intellectual performance.

Feuerstein et al. (1979) held that "observed [cognitive] deficits . . . are determined by deficiencies in certain prerequisites of adequate cognitive functioning" (p. 37). Moreover, they held that these observed deficits could potentially be modified by the application of mediated learning experiences, discussed in greater detail later. In this context, DA would be used to determine specific cognitive targets of an individual's modifiability as well as the type and intensity of mediated intervention observed to produce change. As Feuerstein et al. described it, their learning potential assessment device was designed "to search for the modifiability of these [deficient cognitive functions] and concomitantly to look for strategies and modalities for the most efficient and economical way to overcome the barriers imposed by these differences" (p. 125).

Controversy Over Assessment: The Changing Sociopolitical Climate

In the United States, the divergent work in DA of Feuerstein, Budoff, and Campione and Brown converged in the 1970s (Lidz, 1987b), a time of sociopolitical change. Breakthrough litigation about the rights of students, school placements, and the tests and assessments that led to those placements emerged. In the decade after the landmark *Brown v. Board of Education of Topeka* (1954) decision, a period of civil protest emerged, including legal action and social policy to advance racial equality (McIntosh & Green, 2004). By the early 1970s, advocates had come forth to "secure educational rights of disenfranchised groups, including individuals with disabilities" (Green et al., 2005, p. 84).

In *Hobson v. Hansen* (1967/1969), the first major case raising questions about placement in special education, the court ruled that using test scores to group students into tracks was unconstitutional because it discriminated against African Americans and people who were poor. Linguistic bias in the intelligence tests used to place students in special education classes was also the basis of legal challenges. Although *Diana v. California State Board of Education* (1970) never actually came to court, it resulted in a significant mandate. Because intelligence tests (in English) were used for placement in classes for educable people with mental retardation, the consent decree allowed non-Anglo children to choose the language in which they would respond, banned the use of verbal sections of the tests, and "required state psychologists to develop an IQ test appropriate for Mexican-Americans and other non-English-speaking students" (ERIC Clearinghouse, 1985, para. 5).

Larry P. v. Riles (1972) expanded the ruling in the *Diana* case, citing wrongful placement and overrepresentation in classes for educable people with mental retardation. The court ruled that schools would be responsible for providing tests that did not discriminate on the basis of race. In 1975, Congress enacted the Education for All Handicapped Children Act, which codified the right to a free, appropriate

public education for children with physical and mental disabilities. It required public schools to evaluate those children and to create educational plans with parent input that would be as close as possible to the educational experience of students without disabilities.

During this period, school psychology had come under significant criticism for widespread use of intelligence tests that had contributed to the overrepresentation in special education of culturally diverse children. New ways to assess children were sought to better predict which children would require special education support and which should be able to profit instructionally by continuing schooling in a general education setting. On the basis of Feuerstein et al.'s (1979) research in differentiating cultural difference and disability, DA appeared to hold promise for addressing these issues.

WHAT'S DYNAMIC ABOUT DYNAMIC ASSESSMENT: THE PSYCHOMETRIC-TO-CLINICAL CONTINUUM OF INTERPRETATIONS

Although the various forms of DA come from different traditions, similarities run through all of them and help define what is dynamic about DA. Lidz and Elliott (2000) provided a concise summary of three defining characteristics of DA models: (a) Because the assumption is that cognition can be modified, dynamic models attempt to produce change rather than measure stability; (b) to produce that change, there is an interactive relationship between the assessor and the person being assessed; and (c) although varying across models, dynamic models hold the expectation that embedding an intervention in the assessment will lead to a link between the two and that the learner response to the intervention will carry meaning. In contrast to static assessment, for which change threatens validity, under dynamic paradigms, change can support validity claims. As described next, the nature of the interactions involved in DA vary considerably by model. Nonetheless, all involve some level of interaction or mediation in an attempt place the learner's growth in the context of interactive support.

From the time of Luria's introduction of Vygotsky's notion of zone of proximal development in Europe, DA advocates have differed in whether to take a psychometric or a clinical approach to determining, measuring, or creating change (Poehner, 2008). Psychometrically oriented psychologists have been concerned with standardization and the development of DA measures that could be researched in conventional ways. Thus, their questions to examinees follow standardized or quasi-standardized protocols. Their probes to determine student responsiveness, and therefore change, are based on particular procedures such as verbalization, problem-specific hints, or metacognitive cues.

Those psychologists favoring a clinical approach have not subscribed to a standardized approach because subject responses are used to guide examiner questions, examples, or probes and vary on the basis of each examinee's individual dynamics. Examiner responses are drawn from models of learning that include specific cognitive functions and a set of interactions believed to facilitate cognitive change across content areas. For example, using a clinical DA tool, an examiner's question might be designed to observe a student's approach to a new task, problem definition, and use of specific cognitive functions in doing so, for example, "Tell me what you see here and what you think you will have to do on this problem?" One student might respond to this question with an accurate and detailed explanation. A second might be vague and inaccurate. In the first case, the examiner might have observed the student using cognitive functions such as focus, multiple sources of information, and spatial orientation and problem formulation to gather sufficient information to form and communicate the goal of the problem in that context. In the second case, however, on the basis of the student's specific response, the examiner might hypothesize that he or she was not using one or more cognitive functions needed to gather information or to formulate the problem efficiently. Accordingly, the examiner would then attempt to work with the student to begin to understand and then to use such cognitive functions as attention, systematic search of the information available, or gathering and holding multiple sources of information. Next, they would

determine whether using that cognitive function or functions led to a different level of performance on the original question. If helpful change had occurred with the intervention, the examiner would assist that student to learn to use the function more regularly across items on the DA tool and then to transfer it from the DA to other contexts, including problem solving, life situations, and even schoolwork. Thus, the clinical models are concerned with being able to produce meaningful change on the basis of individual learner characteristics and responses.

Psychometric and clinical models of DA take a range of forms. Some consider content prompts to see whether student responses increase as a result of DA; others believe that DA involves attempts to help students access and efficiently use cognitive functions more permanently, creating what they call *structural cognitive change* (Carlson, 1994, 1995; Feuerstein et al., 1981; Guthke, 1992a, 1992b; Lidz & Elliott, 2000; Sternberg & Grigorenko, 2002). Reschly and Robinson-Zañartu (2000) described this range of models as a continuum, pointing out that "underlying theoretical assumptions, measurement, examiner-examinee interactions, goals for change, number and types of parameters targeted to intervention, and assumptions regarding transfer effect vary widely across these models" (p. 191). In this chapter, the authors present models at both ends of the continuum.

On one end of the continuum are models that adhere to psychometric standards of standardization in testing and intervention procedures. Lantolf and Poehner (2004) referred to this group of models as *interventionist,* describing the approach as remaining "closer to certain forms of static assessment and their concerns over the psychometric properties of their procedures" (p. 18). Their standardized procedures have yielded findings that can most easily be used in conventional forms of research, quantifying results to compare their groups with others and to predict performance on other measures, including achievement. The work of this group of models is rooted in the context of intellectual measurement; thus, they subscribe more closely to an individual mental abilities model. For them, DA is a way to enhance the validity of intelligence or ability measures, especially for groups presumed to have some disadvantage on traditional tests.

At the other end of the continuum are the more clinical or interactionist models, designed to produce change in clients' cognitive or knowledge structures. Advocates of these models describe learners as open systems and thus as capable of cognitive change with the intervention of a human being deliberately seeking to facilitate that change. They portray cognitive functions as "sensitive to specific investments in these . . . interactions, focusing on a dynamic interaction between context and cognition" (Reschly & Robinson-Zañartu, 2000, p. 191). As described earlier, the exact selection of cognitive targets of intervention occurs during rather than before the assessment process. Specific characteristics of the relationship between the assessor and the student facilitate the student's acquisition of previously inefficient cognitive functions; thus, these characteristics are a deliberate part of the interactive intervention. The aim is to help the examinee make efficient use of as full a range of cognitive functions as possible, producing structural cognitive change. The interventions and cognitive functions found to have changed during the assessment–intervention process become targets for longer term intervention, so that the newly efficient cognitive functions may then become habituated. Those who adhere to these models are concerned with the change produced in the individual examinee. Because the examinee is an active participant in the process, the nature of that change is sometimes described as having been coconstructed between the examiner and examinee.

DYNAMIC ASSESSMENT MODELS: PREMISES, PROCEDURES, AND PRACTICES

Dynamic assessment is based on models of cognitive development and change. Although approaches to dynamic assessment are varied, several assumptions are common. One is cognitive modifiability; a second is that the level of performance an individual may have on a test does not necessarily reflect his or her cognitive or learning potential; a third is that modifications in assessment methodologies can provide important information concerning approaches

that may be used to facilitate subsequent learning and change.

Four Psychometric Models

Psychometric models of dynamic assessment involve estimations of change in performance resulting from specific forms of intervention. The principal purpose is to reduce false negatives in assessment and provide more accurate information about an individual's cognitive competence than static, traditional psychometric models and testing approaches. An assumption of these models is that cognitive performance is affected by a variety of nontarget factors. These include metacognitive factors, problem-solving strategies, and motivation and orientation variables.

Budoff's Learning Potential Assessment. Concerned that for some low-achieving students, especially those from poor and minority groups, traditional mental ability assessment led to inappropriate special education placement, Budoff and his colleagues (Budoff, 1970, 1987; Budoff, Meskin, & Harrison, 1971) developed an alternative approach to assess cognitive functioning, the Learning Potential Assessment. As with several DA approaches, the Learning Potential Assessment involves a test–train–retest strategy. The initial testing is traditional in manner, using measures such as the Raven Matrices and Kohs Blocks. Training involves familiarization with the tasks, their demands, and the use of relevant learning strategies. Progressive simplification of the tasks is used to help the child understand the strategies needed in task solution but also to encourage him or her to gain the notion that he or she can perform at higher levels than usually expected. Taking into consideration that each child differs, standardization of the training was approximate, not absolute. Comparing pretest and posttest performance, Budoff (1987) distinguished "gainers" from "nongainers." Posttest performance was considered to be optimal performance, and the separation of gainers from nongainers was considered a demonstration that for some (nongainers), the initial, pretest score was an accurate index of the child's ability; for others (gainers), the posttest score was a more useful and valid index than the pretest score. The utility of the differentiation between gainers and nongainers was providing information concerning appropriate educational programs.

Guthke's learning tests. Based on Vygotsky's work as well as that of others working in the former Soviet Union and Germany (both the former Democratic Republic of Germany and the present Federal Republic of Germany), Guthke and his colleagues (Guthke, 1980, 1992a, 1992b; Guthke, Beckman, & Stein, 1995; Guthke & Lehwald, 1984) developed two approaches to assess learning potential. One approach was to use long-term tests (test–train–retest paradigm); the second was to use short-term tests that systematically involved feedback and interventions during one testing session. Guthke particularly emphasized the psychometric properties of the tests, notably their predictive utility and construct validity. Fundamental to Guthke's work was the conviction that clear, data-based approaches to assessment must take into consideration the basic components of intellectual functioning and how and under what circumstances they relate to individual differences in learning ability.

Campione and Brown's guided learning and transfer model. The Campione–Brown model, also influenced by Vgyotsky, is based on the premise that children begin to learn in situations of social interaction, usually with adults. Working with competent adults, children begin by observing and are guided by adult questions and directions, gradually taking more initiative. Over time, children move toward internalizing the self-regulation and self-interrogation roles of the adult so that they can carry out learning roles independently. Campione and Brown (1987) explained, "It is that transfer of control that we seek to capture in our assessment and instruction sessions" (p. 83). Their approach to DA (Brown & Campione, 1986; Campione & Brown, 1987; Palinscar, Brown, & Campione, 1991) focuses on the processes underlying successful performance and on specific domains rather than generalized ability. It uses a series of problem-specific standardized hints and questions designed to support the child's learning to solve the task. The hints can be detailed and lengthy, focusing on the tasks specifically, on more general metacognitive cues, or on both. Particular

attention is given to the rules or principles in problem solution because of their relationship to transfer. Campione and Brown (1987) provided the following examples using a rotation problem:

> Hint 1. This problem is called a turning problem. Think about why it might be called that. . . . Do you know how to solve the problem now or do you want another hint? Hint 2: This is row 1. Put picture 1 in the practice box. Touch the picture. Now try to make the picture look like the second picture. You did it. Now make it look like the last picture. Hint 4. You used the turning rule to make the last picture in rows 1 and 2. The last picture in row 3 is missing. Try to use the same rule to make the missing picture in row 3. (p. 110)

The type and number of hints about how to approach the task needed for a child to solve a problem, graduated in their detail, provide estimates of the child's learning potential. The metrics of analysis are how much aid is required for the child to reach a particular level of performance in the test–train–retest paradigm and the degree to which transfer can be demonstrated with tasks increasingly dissimilar from the initial posttest. The ability to transfer newly acquired skills to relatively novel situations is important because it is a good predictor of how responsive the child will be to instruction.

Carlson and Wiedl's testing the limits. Convinced that standardized assessment approaches can often underestimate an individual's intellectual ability (Scarr, 1981), Carlson and Wiedl (1979) developed a research program designed to assess which, how, and to what extent various nontarget variables such as impulsivity, anxiety, and motivation affect performance on mental ability tests. The goal of testing was to arrive at an accurate assessment of the target variable, that which is assumed to be assessed by the test, that is, the test's construct validity. To the extent that individual differences in performance affecting nontarget variables affect target-variable performance, a significant aspect of bias is introduced and the validity of the assessment is brought into question. The approach Carlson and Wiedl used and the methods they developed involve modifications in the test situation. Their approach differs from the test–train–retest paradigm of many DA approaches and avoids statistical problems related to the measurement of change. In a series of studies, Carlson, Wiedl, and their colleagues isolated a number of personal or noncognitive factors that negatively affect performance. These factors include anxiety and lack of motivation (Bethge, Carlson, & Wiedl, 1982), impulsive responding (Wiedl, 1980), poor ability to plan (Cormier, Carlson, & Das, 1990; Kar, Dash, Das, & Carlson, 1993), and lack of awareness and ability to spontaneously generate cognitive and metacognitive strategies to solve task problems (Carlson & Wiedl, 1979, 1980). The most effective methods of assessment shown to meliorate the effects of the performance-reducing factors involved active overt verbalization as the individual solved the tasks and elaborated feedback provided by the examiner, which involved providing the test taker with information about the correctness or incorrectness of his or her response. The performance of interest was on the following item, before any feedback. In this way, the test-taker's responsiveness to feedback was assessed. For children over a lower threshold of mental age 6, overt, active verbalization tended to be the most effective intervention. Several studies have shown the effectiveness of the approach with deaf children using American Sign Language (Carlson & Dillon, 1978) and with adult individuals with psychiatric disorders (Wiedl & Schoettke, 1995).

Three Clinical Models

Clinical models of dynamic assessment are designed to attempt to produce change in their students' or clients' cognition and to do so with deliberate interventions during the assessment. The client responses to intervention guide the next set of examiner questions, mediations, or prompts. The outcomes of these clinical models go beyond labeling a student as modifiable or making psychometrically based comparisons of their behaviors to others; rather, they attempt to design situations to foster the changes found. Three such models, used over time and across continents, are described here as strong representatives of this group.

Feuerstein's model of structural cognitive modifiability. Influenced by the work of Rey and Piaget, with whom he had studied, Feuerstein became dissatisfied with conventional cognitive assessment when working with children in Israel after World War II. He postulated that the harsh disconnect of many children from intentional intergenerational transmission of their home cultures while enduring the traumas of war had contributed significantly to low performance on these conventional measures. Believing that cognitive change was possible, he devised a series of methods of "mediating" the development of cognitive functions. Feuerstein et al. (1979) defined 10 facilitating interventions as characteristics of a mediated learning experience that would contribute to cognitive change. Those parameters describing the intense interaction between examiner and examinee were designed to help the examinee engage with the tasks and cognitive functions, find personal meaning in using the functions with efficiency, see their value beyond the immediate task, learn to self-regulate for efficiency, and gain a sense of competence in doing so. Moreover, mediated learning experience addressed supporting the examinee's gaining individuation and differentiation; goal-seeking, goal-setting, and goal achievement behaviors; seeking challenge; and using self-reflection regarding his or her own change to enhance "insight into his or her growing proficiency" (Jensen & Feuerstein, 1987, p. 389).

Feuerstein et al. (1979) identified 27 initial cognitive functions that he and colleagues had found in clinical settings to be open to modification with the use of mediated learning experience. They grouped the cognitive functions into three categories: those used mainly for input, for elaboration, and for output of information. Systematic exploration, verbal tools and concepts, and simultaneous use of two or more sources of information are examples of the input functions. Problem definition, relevant cue selection, spontaneous comparative behavior, summative behavior, inferential–hypothetical thinking, and planning behavior are examples of elaboration functions. Finally, examples of problematic cognitive functioning in the output of information include functions such as egocentric communication, blocking, impaired verbal tools, and impulsive responding. Feuerstein et al. recognized that two or more functions often operate simultaneously, often across those categories; thus, the list was not meant to be hierarchical or linear. Each individual was presumed to present a unique constellation of efficient and inefficient functions.

Combined with specific assessment instruments, Feuerstein et al.'s (1979) learning potential assessment device set forth principles and practices for DA. The following year, Feuerstein, Rand, Hoffman, and Miller (1980) published a companion program for the development of cognitive functions called *instrumental enrichment.* The theory that accompanies the learning potential assessment device and instrumental enrichment programs, structural cognitive modifiability, describes a complex set of variables and specific ways to help the learner re-form cognitive habits, undergo structural change at the cognitive level, and thereby enhance functioning (Lidz, 1991). Feuerstein's work has since permeated educational and therapeutic communities across five continents. A variety of clinical studies have demonstrated change in students' functional behaviors and cognitive skills (cf. Lidz & Elliott, 2000).

Jensen's mediated constructivism. Mogens Jensen worked with Feuerstein both in Israel and in the United States. With Singer, he researched the effects of Feuerstein's instrumental enrichment program (Jensen & Singer, 1987), demonstrating transfer of newly acquired or efficient cognitive functions to similar tasks (near transfer) and contributing to the validation of Feuerstein's three groupings of cognitive functions. Additionally, they found a fourth factor, cognitive control, which is similar to what is known as self-regulation. When their findings did not result in spontaneous use of newly acquired cognitive skills during instrumental enrichment in dissimilar situations such as the curriculum (far transfer), Jensen (2000) began the development of his MindLadder model.

Jensen's (2000) MindLadder model is based in his theory of mediated constructivism, which holds that students' active construction of meaningful information, intentionally mediated, should be coupled with "students' acquisition of content knowledge and behavioral skill—all within an active,

coherent and meaningful learning environment" (p. 191). The theory names 75 functions (45 intellective, 20 nonintellective, and 10 performance) and five mediating qualities, adapted from those described by Feuerstein, Rand, and Hoffman (1979). On the basis of this theory, Jensen developed a Web-based questionnaire for teachers to identify cognitive strengths and weaknesses of each student. In response to the functions identified, teachers would develop classroom lessons and infuse the functions identified into their curricula. To test the theory's treatment validity, he conducted a classroom-based research project, drawing on a sample of 347 fourth-, fifth-, and sixth-grade students, with support for teachers in the form of coaching. He found that on the Iowa Test of Basic Skills, the Mind-Ladder students outperformed control students in reading ($p < .007$), language ($p < .001$), social studies ($p < .01$), and the Iowa Test of Basic Skills composite ($p < .0001$; Jensen, 2003).

Tzuriel's Cognitive Modifiability Battery. Growing out of Tzuriel's work with Feuerstein, although equally influenced by Vygotsky (Tzuriel, 2000), Tzuriel's Cognitive Modifiability Battery targets work with young children from ages 5 to 7 and older children experiencing learning difficulties. The Cognitive Modifiability Battery uses DA tools designed for work with young children, both for diagnosis and for intervention (Tzuriel, 2000). Research studies using CMB have applied it either as a pre- and posttest measure or as an intervention; they have focused on changes in a variety of aspects of cognition, from spatial orientation to conceptual analogical thinking. Some of these studies have demonstrated near transfer, such as predicting better outcomes on cognitive education programs. In addition, some have also reached into diverse areas such as demonstration of closing gaps in gender differences (e.g., Tzuriel & Alfassi, 1994; Tzuriel & Egozi, 2010; Tzuriel & Klein, 1985).

DYNAMIC ASSESSMENT AND ISSUES OF VALIDITY

Accurate prediction of an individual's potential to function effectively in a variety of situations is useful. In the industrial sector, for example, accurate prediction can result in the selection of individuals who will need less training and are more likely to work efficiently, as described in several chapters of this handbook. It is cost effective. False positives can be monetarily expensive; false negatives—that is, not selecting a person with low scores but who might do well given the chance—are generally less so.

In education, the situation is different. Avoiding false negatives is essential for schools and school systems to provide fair, equal-opportunity, and effective education. Accordingly, the educational system must be responsive to individual and group differences that affect, or potentially affect, learning and the potential to learn. Advocates of DA have claimed that alternative, interactive approaches to mental ability testing can reduce false negatives and generally provide not only more accurate information about an individual's cognitive abilities than standard, traditional testing approaches but also evidence for effective teaching approaches. (For an historical overview, see Lidz, 1987b.)

To establish the general validity and usefulness of the approach, the question is how robust the findings are across different DA methodologies, dependent variables, and populations assessed. Two informative meta-analytic studies address this issue.

Swanson and Lussier (2001)

Framing validity issues in terms of usefulness, Swanson and Lussier (2001) conducted a study using meta-analytic techniques to determine the differences between DA and other approaches. Their goal was to present a synthesis of the DA literature concerning two general questions: First, does DA modify group differences in ability on different dependent variables and provide better estimates of ability? Second, are effect sizes related to DA artifacts of research design, treatment intensity, and type of instruction?

Swanson and Lussier's (2001) search for articles for analysis included an expanded review of the PsycINFO database (1964–1999) and two DA review articles: Laughon (1990), which included 62 articles, and Grigorenko and Sternberg (1998), which included 229 articles. Thirty articles met the criteria for inclusion in the final analysis; all

involved the test–train–retest paradigm, 23 from the United States, three from Germany, two from Israel, one from Canada, and one from India. The total sample in the groups involved 5,104 participants disproportionately representing five groups: those with learning disabilities, those who underachieved, those with hearing impairment, those who were educable with mental retardation, and those who were average achievers. Three DA models were represented: testing the limits, mediated training using coaching, and structured strategy training and feedback (scaffolding). Dependent variables were of two types: verbal (story recall, rhyming, phrase recall, Peabody Picture Vocabulary Test) and visual–spatial (visual matrix, Raven Matrices).

The results revealed an overall effect size of .70, indicating that DA resulted in substantial improvement on the dependent variables over static testing. The highest effect sizes were found for studies that involved verbal elaboration or mediation and feedback. Last, no ability group variables were shown to yield different effect sizes. That is, DA affected different ability groups similarly. Swanson and Lussier (2001) interpreted the latter as support for "the contention that changes in performance as a function of DA procedures reflect abilities *independent* [italics added] of measures of traditional classification and procedures" (p. 359).

Caffrey, Fuchs, and Fuchs (2008)

Focusing on the efficacy (validity) of DA for predicting future achievement, Caffrey, Fuchs, and Fuchs (2008) carried out a meta-analytic study, basing article selection on four criteria: (a) articles published in English; (b) participant samples enrolled in preschool through high school; (c) participants with high-incidence disabilities or at risk for school failure because of cultural or economic disadvantage, second-language learners, or normally achieving students; and (d) articles reporting data used to determine predictive validity.

Caffrey et al.'s (2008) search for articles included (a) ERIC, PsycINFO, and Exceptional Child Education Resources (ECER), using key phrases *dynamic assessment* or *interactive assessment* or *learning potential* or *mediated assessment;* (b) a review of a 1992 special issue of the *Journal of Special Education* on the topic; and (c) articles referred to in reviews by Grigorenko and Sternberg (1998) and Swanson and Lussier (2001). In a second ERIC, PsycINFO, and ECER search, Caffrey et al. expanded the terms to include *mediated learning* and *predictive validity*. A total of 24 studies met the criteria for inclusion in the analysis.

Caffrey et al. (2008) analyzed the data from the selected studies on four dimensions: first, the correlations between traditional testing and DA and achievement; second, the type of feedback, contingent or noncontingent, involved in the DA studies; third, the predictive validity of DA for various types of students; and fourth, the predictive relationships between DA and different achievement criteria, for example, teacher judgment, norm- or criterion-referenced tests, and independent performance. They reported results for each individual study separately, summarized as follows:

> When feedback is noncontingent, predictive validity is higher for DA approaches than for traditional assessment. It is also higher for students with disabilities than normal achieving but at-risk students. Finally, it is higher when the achievement criteria are assessed with criterion referenced tests as opposed to norm-referenced or teacher judgment. (Caffrey et al., 2008, p. 254)

CURRENT TRENDS IN SCHOOL PRACTICE AND DYNAMIC ASSESSMENT

Given the nature of DA and its link to intervention, its use in schools seems logical. DA holds the potential to assess conditions under which student change occurs, resulting in improved student approaches to learning and problem solving. It holds the potential to describe interactions that enhance student learning and to help school psychologists frame teacher recommendations based trial interventions embedded within the assessment.

What Impedes the Use of Dynamic Assessment in Schools

Sternberg (2000; Sternberg & Grigorenko, 2002) asserted that although it should seem obvious that

dynamic testing would be the assessment of choice, four things impede it use: first, inertia; second, its administration can be relatively complicated; third, it requires special training; and fourth, the psychometrics can be complicated and some are "subject to clinical interpretation" (Sternberg, 2000, p. xv). An additional potential barrier lies in the variety of DA models, so that both the definition and the use of DA in schools has not been clear.

School psychologists today participate both in eligibility decision making and in consultation and direct intervention work. Often, one or the other of these roles will dominate the psychologist's practice as well as his or her paradigm of practice. Each of these two paradigms has barriers that currently impede it from more widespread adoption (Robinson-Zañartu, 2008). The first, rooted in test administration and concerns for psychometric integrity, places diagnosis and special education eligibility determination as central. To this end, clinical DA can seem lengthy and complex and its outcomes only tangentially relevant to eligibility determinations. The psychometric models of DA have not had widespread visibility in school psychology and might greatly enhance this model of practice, even holding stronger predictive validity than static measures. The second paradigm places intervention and consultation at the forefront, which should be far more compatible with the clinical models. Bransford, Delclos, Vye, Burns, and Hasselbring (1987) concluded that, although the two models (psychometric or clinical) might both result in effective learning, the graduated prompting methods may be better suited for supporting classification but that the more clinical model, which used mediation, "seems to be associated with better transfer [as well as for] . . . discovering information about effective instructional strategies for individual children" (p. 487). However, this set of models requires more extensive training.

Curriculum-Linked Dynamic Assessment Methods

Linking DA directly to school achievement holds great intuitive appeal. It speaks more directly to the focus on achievement prevalent in school practice. Several methods have attempted to link the two; those of speech and language pathologists have been most widely used in schools to date.

Graduated prompts methods. Campione and Brown (1984, 1987) were among the earliest to link DA with school achievement, combining notions from Vygotsky with concerns for psychometric integrity. In their work (described earlier), they attempted to establish estimates of the person's learning potential and what they called *transfer efficiency.* Campione and Brown (1987) approached DA with the following assumptions:

> Assessment should evaluate as directly as possible the particular processes underlying successful performance . . . [and] should ideally be situated within a specific domain.. . . This in turn increases the likelihood that the processes can be specified in sufficient detail that instructional prescriptions can be designed. (p. 88)

Their early findings led them to believe that principled transfer was possible with structured interactive intervention and should involve principles for subsequent application to novel contexts. Their concern for transfer (and use of transfer tasks) was extremely important. The posttest was not only a parallel test but a transfer task. Their work with this assessment model led to Palincsar and Brown's (1985) work with reciprocal teaching, which operationalized notions of graduated prompts and the value of practice with overt verbalization while targeting the skills needed to perform reading comprehension.

Application of Cognitive Functions Scale. Responding to growing concerns that a closer relationship between assessment and instruction was needed in school psychology and special education, Lidz (2000) developed the Application of Cognitive Functions Scale for use with preschool children. This scale combines curricular and process skills (e.g., classification, perspective taking, auditory memory) and deliberately aligns with psychometric models of DA; thus, predetermined interventions are semiscripted. The interventions teach processes underlying the posttest areas but avoid test items and thus the practice effect. Several studies with

preschool children demonstrated gains from the pretest to posttest in at least some of the process areas after the intervention (e.g., Malowitsky, 2001; Shurin, 1998).

Curriculum-based dynamic assessment. Haywood and Lidz (2007) proposed a generic process in which curriculum-based measurement would be linked with DA via identification of the process components needed to accomplish the domain-specific tasks involved in the curriculum-based measurement probe. The dynamic assessor uses the curriculum-based measurement probe as a pretest, intervenes using DA of the processes from that analysis, then conducts a curriculum-based measurement posttest to determine to what extent "the processes demanded by the task . . . [are] developed and intact" (p. 178). They called this *curriculum-based DA*.

Dynamic assessment in speech–language pathology. In speech–language pathology, concern for more accurate identification of language impairments in culturally and linguistically diverse learners led to the use of DA. Lidz and Peña (2009) posited that because speech pathologists must go beyond assessment into interventions, those DA measures directly inform their intervention or curriculum (e.g., narrative skills): "The outcome of the DA assessment should be a specific plan of instruction that meets the needs of the individual learner" (p. 126). DA approaches in speech and language pathology have ranged from test–teach–retest to successive cuing. For instance, speech–language pathologists have used successive cuing approaches to determine ability such as articulation stimulability beyond that provided by static tests and to evaluate readiness for intervention. Lidz and Pena discussed examples of how test–teach–retest strategies have been helpful in successfully identifying or differentiating children with and without language impairments:

> Roseberry and Connell . . . found that children from culturally diverse backgrounds with and without language impairment learned an invented morpheme rule at different rates. This differential learning rate allowed the authors to classify the two groups with better sensitivity and specificity. Similarly, Jacobs found that the addition of a learning component to her computerized preschool language screening enhanced the information available to her linguistically diverse preschoolers from low socioeconomic backgrounds. (p. 123)

Peña et al. (2006; Peña, Iglesias, & Lidz, 2001) have used mediational approaches to differentiate difference and disorder and found children's metacognitive skill and flexibility to be highly predictive of language ability.

L2 and dynamic assessment. Poehner (2008) saw the heart of his Vygotskian approach to DA as the mediation between assessor and learner. He borrowed from Feuerstein et al.'s (1979) characteristics of mediated learning, emphasizing three: the mediation of intentionality (coupled with reciprocity, proposed by Lidz, 1991), of transcendence, and of meaning. The context of Poehner's work is second language acquisition; he reported specifically on working with advanced students in French. His focus was not on cognitive functions underlying the content but on helping learners "develop a new theoretical understanding of [a particular] feature of French that they could use to regulate their functioning in the language" (p. 112). After a content-specific pretest, 6 weeks of mediated intervention provided highly dialogic interactive support, which he called DA. During the posttest phase, the dialogic interactions continued, and the results were compared across time. He then inserted transcendence (transfer) tasks in the form of new and far more challenging problems.

DYNAMIC ASSESSMENT AND RESPONSIVENESS TO INTERVENTION MODELS

Commonalities across responsiveness-to-intervention (RtI) and DA models have led some investigators to discuss merging them and others to practice unique forms of such a merger. RtI models use a three-tiered approach to successively more intense levels of assessment and intervention, beginning with

monitoring at the whole-school level, then offering intervention to students who are falling behind. Their goals in general are to prevent and intervene with students with reading and behavioral difficulties and to provide valid identification of students having these difficulties as a result of disabilities (Bradley, Danielson, & Hallahan, 2002; Donovan & Cross, 2002). Lidz and Peña (2009) suggested that both DA and RtI are "ultimately concerned with promoting the competence of learners within educational settings" (p. 122), pointing out that they both focus on the outcomes of interventions as well as on documenting what produces change. Similarly, Grigorenko (2009) concluded, "Both approaches belong to one family of methodologies in psychology and education whose key feature is in blending assessment and intervention in one holistic activity" (p. 111).

Some advocates of the RtI movement have conducted research using DA to enhance RTI findings (e.g., D. Fuchs et al., 2007; L. S. Fuchs et al., 2008; Grigorenko, 2009; Lidz & Peña, 2009); some DA advocates have embraced the notions of RtI. Certainly there is overlap in the basic concept of using RtI as a means to support learner growth. In DA, "the response of the learner to the embedded interventions . . . is [its] essence and core" (Lidz & Peña, 2009, p. 122). However, the context, scope, and goals of each can differ considerably. The three-tiered RtI models begin with attention to the whole school's effective instruction and to screening at that level for difficulties in either academics or behavior. With the exception of Jensen's (2003) MindLadder program, which attempts to have the whole school screen for cognitive efficiencies, in most DA, the context begins with students having difficulty, which might occur at either Tier 2 or Tier 3 of RtI. Usually, RtI leads to interventions with (unspecified) evidence-based instructional methods, determining that responsiveness to those methods alone should determine whether a student needs special education support. L. S. Fuchs et al. (2008) suggested that DA might well be used in an RtI framework to help identify students who will ultimately prove unresponsive to Tier 1 prevention. Predicting who will later fail so as to provide earlier support might be helpful; however, it does not address the potential of DA to support more effective instruction. Some psychometric forms of DA hold similar goals; however, clinical forms of DA also hold the goal of supporting cognitive change in the way in which students approach, transform, and communicate information and in helping with the design of ongoing interventions to help the student sustain that change. Two examples illustrate these differences.

D. Fuchs et al. (2007) suggested that rather than having students go directly to special education if a conventional RtI intervention was not successful, DA could be used to try additional RtI-type interventions. L. S. Fuchs et al. (2008) defined DA as "helping students learn a task and indexing responsiveness to that instruction" (p. 829), and in this form of DA they intervene with academically grounded interactions. Their purpose is prediction of future learning, and they use the learning of content unrelated to current curriculum (e.g., algebra learning for third graders). Their research studies have found DA to be a helpful differentiator in both reading and math.

Founded in a similar belief that DA could provide useful additional interventions before classifying students for special education, Robinson-Zañartu and colleagues (Green et al., 2005; Robinson-Zañartu, 2009; Robinson-Zañartu, Fiz, Genet, Mercado, & Ornelas, 2010) designed and piloted what she calls *response to mediated intervention,* which begins with the teacher referral question, from which baseline data emerge (e.g., reading fluency, time on task). DA uses a clinical model to identify cognitive functions (called *thinking skills*) that when mediated produce change in performance, first on DA tasks to minimize blocking and then transferred into the domain of concern, such as reading or math. On the basis of the trial interventions embedded in the DA, a 6- to 8-week formal intervention is designed and carried out, using progress monitoring or pretest–posttest data to assess effectiveness. All interventions include the mediation of self-regulation, because it is both one of the facilitators of change in the model and has considerable research support for effect on academic change (e.g., Bail, Zhang, & Tachiyama, 2008; Paris & Paris, 2001; Pelco & Reed-Victor, 2007; Shimabukuro, Prater, Jenkins, &

Edelen-Smith, 1999). Response to mediated intervention results inform classroom intervention and instruction.

Well over a decade ago, Bransford et al. (1987) suggested that clinical models of DA appeared especially useful in determining appropriate instructional methodologies. The nature of the intervention in response to mediated intervention, the mediation of specific or grouped cognitive functions, becomes the subject of recommendations for instruction, or longer term intervention, with examples of their infusion into that content area of concern. Results from 24 initial case studies are extremely promising. Response to mediated intervention is designed to exist side by side with the behavioral and academic RtI models, augmenting both Tier 2 and Tier 3 interventions.

SUMMARY AND CONCLUSION

DA has a rich tradition, built on the belief that static assessments often do not tell a sufficient story about student ability and student change. Examiners using DA build on an interactive relationship; intervene by trying out instruction or mediation of cognition, thinking, or learning skills (depending on the theory); and determine student responses to that trial. DA has been of particular interest to professionals working with students from culturally and linguistically diverse backgrounds (Lidz & Peña, 2009). Two general traditions have emerged, one emphasizing the role of DA in measuring change for the purpose of predicting future responsiveness to instruction or intervention and the second involving DA to produce change and detail the nature of the processes that would continue to support enhanced change in the form of problem solving and academic success.

For school psychologists, assessment must, in part, have utility for teachers; that is, it must have a demonstrable relationship to curriculum and instruction. DA methods have the potential to help answer questions about who can profit from enhanced instruction and what kinds of interactions between teachers and students will support students' change in this process. Several DA models have emerged for working with or beside the current school reform movement using RtI. Finally, schools and psychologists in schools are beginning to be called on to address "21st-century skills," which include critical thinking and problem solving as core values (New Commission on the Skills of the American Workforce, 2007; Trilling & Fadel, 2009). Because the heart of some DA models is the examination of these skills, or elements of these skills, DA may play an increasingly central role in identifying and enhancing those skills in schools.

References

Bail, F. T., Zhang, S., & Tachiyama, G. T. (2008). Effects of a self-regulated learning course on the academic performance and graduation rate of college students in an academic support program. *Journal of College Reading and Learning, 39*, 54–73.

Bethge, H. J., Carlson, J., & Wiedl, K. H. (1982). The effects of dynamic assessment procedures on Raven Matrices performance, visual search behavior, test anxiety, and test orientation. *Intelligence, 6*, 89–97. doi:10.1016/0160-2896(82)90022-8

Bradley, R., Danielson, L., & Hallahan, D. P. (2002). *Identification of learning disabilities: Research to practice*. Mahwah, NJ: Erlbaum.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academies Press.

Bransford, J. D., Delclos, V. R., Vye, N. J., Burns, M. S., & Hasselbring, T. S. (1987). State of the art and future directions. In C. S. Lidz (Ed.), *Dynamic assessment: An interactive approach to evaluating learning potential* (pp. 479–496). New York, NY: Guilford Press.

Brown, A. L., & Campione, J. (1986). Psychological theory and the study of learning disabilities. *American Psychologist, 41*, 1059–1068. doi:10.1037/0003-066X.41.10.1059

Budoff, M. (1967). Learning potential among institutionalized young adult retardates. *American Journal of Mental Deficiency, 72*, 404–411.

Budoff, M. (1970). Learning potential: Assessing ability to reason in the educable mentally retarded. *Acta Paedopsychiatrica, 37*, 293–309.

Budoff, M. (1987). Measures for assessing learning potential. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 52–81). New York, NY: Guilford Press.

Budoff, M., & Friedman, M. (1964). Learning potential as an assessment approach to the adolescent mentally retarded. *Journal of Consulting Psychology, 28*, 434–439. doi:10.1037/h0040631

Budoff, M., Meskin, J., & Harrison, R. H. (1971). Educational test of the learning-potential hypothesis. *American Journal of Mental Deficiency, 76*, 159–169.

Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008). The predictive validity of dynamic assessment: A review. *Journal of Special Education, 41*, 254–270. doi:10.1177/0022466907310366

Campione, J. C., & Brown, A. L. (1984). Learning ability and transfer propensity as sources of individual differences in intelligence. In P. H. Brooks, R. D. Sperber, & C. McCauley (Eds.), *Learning and cognition in the mentally retarded* (pp. 265–294). Baltimore, MD: University Park Press.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 82–115). New York, NY: Guilford Press.

Carlson, J., & Wiedl, K. H. (1980). Applications of a dynamic testing approach in intelligence assessment: Empirical results and empirical formulations. *Zeitschrift für Differentielle Psychologie, 1*, 303–318.

Carlson, J. S. (1994). Dynamic assessment of mental abilities. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (Vol. 1, pp. 368–372). New York, NY: MacMillan.

Carlson, J. S. (Ed.). (1995). *European contributions to dynamic assessment: Advances in cognition and educational practice.* Greenwich, CT: JAI Press.

Carlson, J. S., & Dillon, R. (1978). Measuring intellectual capabilities of hearing-impaired children: Effects of testing-the-limits procedures. *Volta Review, 80*, 216–223.

Carlson, J. S., & Wiedl, K. H. (1979). Towards a differential testing approach: Testing-the-limits employing the Raven Matrices. *Intelligence, 3*, 323–344. doi:10.1016/0160-2896(79)90002-3

Clauss-Ehlers, C. S. (2004). A framework for school-based mental health promotion with bicultural Latino children: Building on strengths to promote resilience. *International Journal of Mental Health Promotion, 6*(2), 26–33. doi:10.1080/14623730.2004.9721928

Cormier, P., Carlson, J., & Das, J. P. (1990). Planning ability and cognitive performance: The compensatory effects of dynamic assessment. *Learning and Individual Differences, 2*, 437–449. doi:10.1016/1041-6080(90)90004-Z

Dearborn, W. F. (1921). Intelligence and its measurement. *Journal of Educational Psychology, 12*, 210–212. doi:10.1037/h0065003

De Weerdt, E. H. (1927). A study of the improvability of fifth grade children in certain mental functions. *Journal of Educational Psychology, 18*, 547–557. doi:10.1037/h0073097

Diana v. State Board of Education, CA 70 RFT (N.D. Cal., Feb. 3, 1970, 1973).

Donovan, M. S., & Cross, C. T. (2002). *Minority students in special and gifted education.* Washington, DC: National Academies Press.

Dweck, C. (2006). *Mindset: The new psychology of success.* New York, NY: Random House.

Education for All Handicapped Children Act of 1975, Pub. L. 94–142, 20 U.S.C. § 1400 *et seq.*

Edwards, L. M., Holtz, C. A., & Green, M. B. (2007). Promoting strengths among culturally diverse youth in schools. *School Psychology Forum: Research in Practice, 2*(1), 39–49.

ERIC Clearinghouse on Tests Measurement and Evaluation. (1985). *Legal issues in testing.* Retrieved from http://www.ericdigests.org/pre-927/legal.htm

Feuerstein, R., Miller, R., Rand, Y., & Jensen, M. R. (1981). Can evolving techniques better measure cognitive change? *Journal of Special Education, 15*, 201–219. doi:10.1177/002246698101500209

Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device.* Baltimore, MD: University Park Press.

Feuerstein, R., Rand, Y., Hoffman, M. B., & Miller, R. (1980). *Instrumental enrichment*. Baltimore, MD: University Park Press.

Finger, S. (1994). *Origins of neuroscience: A history of explorations into brain function.* New York, NY: Oxford University Press.

Flavell, J. (1963). *The developmental psychology of Jean Piaget*. Princeton, NJ: Van Nostrand.

Fuchs, D., Fuchs, L. S., Compton, D. L., Bouton, B., Caffrey, E., & Hill, L. (2007). Dynamic assessment as responsiveness to intervention: A scripted protocol to identify young at-risk readers. *Teaching Exceptional Children, 39*, 58–63.

Fuchs, L. S., Compton, D. L., Fuchs, D., Hollenbeck, K. N., Craddock, C. F., & Hamlett, C. L. (2008). Dynamic assessment of algebraic learning in predicting third graders' development of mathematical problem solving. *Journal of Educational Psychology, 100*, 829–850. doi:10.1037/a0012657

Green, T. D., McIntosh, A. S., Cook-Morales, V. J., & Robinson-Zañartu, C. (2005). From old schools to tomorrow's schools: Psychoeducational assessment of African American students. *Remedial and Special Education, 26*, 82–92. doi:10.1177/07419325050260020301

Grigorenko, E. L. (2009). Dynamic assessment and response to intervention: Two sides of one coin. *Journal of Learning Disabilities, 42*, 111–132. doi:10.1177/0022219408326207

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin, 124*, 75–111. doi:10.1037/0033-2909.124.1.75

Guthke, J. (1980). *Ist Intelligenz messbar?* [Can intelligence be measured?]. Berlin, Germany: Deutscher Verlag der Wissenschaften.

Guthke, J. (1992a). Learning tests: The concept, main research findings, problems and trends. In J. Carlson (Ed.), *Advances in cognition and educational practice* (Vol. 1, pp. 213–235). Greenwich, CT: JAI Press.

Guthke, J. (1992b). The learning test concept: Origins, state of the art and trends. In H. C. Haywood & D. Tzuriel (Eds.), *Interactive assessment* (pp. 64–94). New York, NY: Springer.

Guthke, J., Beckman, J. F., & Stein, H. (1995). Recent research evidence on the validity of learning tests. In J. Carlson (Ed.), *Advances in cognition and educational practice* (Vol. 3, pp. 117–143). Greenwich, CT: JAI Press.

Guthke, J., & Lehwald, G. (1984). On component analysis of the intellectual learning ability in learning tests. *Zeitschrift für Psychologie mit Zeitschrift für Angewandte Psychologie, 192*, 3–17.

Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications*. New York, NY: Cambridge University Press.

Hessels, M. G. P. (2000). The Learning Potential Test for Ethnic Minorities (LEM): A tool for standardized assessment of children in kindergarten and the first years of primary school. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 109–132). Oxford, England: JAI Press/Ablex.

Indiana University. (2007). *Human intelligence: Jean-Marc Gaspard Itard.* Retrieved from http://www.indiana.edu/~intell/itard.shtml

Hobson v. Hansen, 269 F. Supp. 401, 514 (D.D.C. 1967), *aff'd, sub nom*, Smuck v. Hobson, 408 F.2d 175 (D.C. Cir. 1969).

Jensen, M. R. (2000). The MindLadder model: Using dynamic assessment to help students learn to assemble and use knowledge. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 187–227). New York, NY: JAI Press.

Jensen, M. R. (2003). Mediating knowledge construction: Towards a dynamic model of assessment and learning: Part II. Applied programs and research. *Educational and Child Psychology, 20*, 118–142.

Jensen, M. R., & Feuerstein, R. (1987). The learning potential assessment device: From philosophy to practice. In C. S. Lidz (Ed.), *Dynamic assessment: An interactive approach to evaluating learning potential* (pp. 379–402). New York, NY: Guilford Press.

Jensen, M. R., & Singer, J. L. (1987). *Structural cognitive modifiability in low functioning adolescents: An evaluation of instrumental enrichment*. Hartford: Report to the State of Connecticut Department of Education, Bureau of Special Education and Pupil Personnel Services.

Kar, B. C., Dash, U. N., Das, J. P., & Carlson, J. (1993). Two experiments on the dynamic assessment of planning. *Learning and Individual Differences, 5*, 13–29. doi:10.1016/1041-6080(93)90023-L

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment: Bringing the past into the future. *Journal of Applied Linguistics, 1*, 49–72. doi:10.1558/japl.1.1.49.55872

Larry P. v. Riles, 343 F. Supp. 1306 (D.C.N.D. Cal., 1972), *aff'd.,* 502 F. 2d 963 (9th Cir. 1974), *further proceedings*, 495 F. Supp 926 (D.C.N.D. Cal., 1979), *aff'd.,* 502 F. 2d 693 (9th Cir. 1984).

Laughon, P. (1990). The dynamic assessment of intelligence: A review of three approaches. *School Psychology Review, 19*, 459–470.

Lee, C. D., & Slaughter-Defoe, D. T. (2004). Historical and sociocultural influences on African American education. In J. Banks & C. A. M. Banks (Eds.), *Handbook of research on multicultural education* (pp. 348–371). San Francisco, CA: Jossey-Bass.

Lidz, C. S. (Ed.). (1987a). *Dynamic assessment: An interactive approach to evaluating learning potential.* New York, NY: Guilford Press.

Lidz, C. S. (1987b). Historical perspectives. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach* (pp. 3–32). New York, NY: Guilford Press.

Lidz, C. S. (1991). *Practitioner's guide to dynamic assessment*. New York, NY: Guilford Press.

Lidz, C. S. (2000). The Application of Cognitive Functions Scale (ACFS): An example of curriculum-based dynamic assessment. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 407–439). New York, NY: Elsevier.

Lidz, C. S., & Elliott, J. G. (Eds.). (2000). *Advances in Cognition and Educational Practice: Vol. 6. Dynamic assessment: Prevailing models and applications* (J. S. Carlson, Series Ed.). New York, NY: Elsevier.

Lidz, C. S., & Peña, E. D. (1996). Dynamic assessment: The model, its relevance as a nonbiased approach, and its application to Latino American preschool children. *Language, Speech, and Hearing Services in Schools, 27*, 367–372.

Lidz, C. S., & Peña, E. D. (2009). Response to intervention and dynamic assessment: Do we just appear to be speaking the same language? *Seminars in Speech and Language, 30*, 121–133. doi:10.1055/s-0029-1215719

Malowitsky, M. (2001). *Investigation of the effectiveness of the mediation portion of two subscales of the application of cognitive function scale, a dynamic assessment procedure for young children* (master's thesis). Available from ERIC (TM033288).

McIntosh, A. S., & Green, T. D. (2004). Fifty years down the road: Have we lost our way? *Journal of School Public Relations, 25*, 1–22.

Minick, N. (1987). Implications of Vygotsky's theories for dynamic assessment. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 116–140). New York, NY: Guilford Press.

New Commission on the Skills of the American Workforce. (2007). *Tough choice or tough times* (Executive summary). Washington, DC: National Center on Education and the Economy. Retrieved from http://www.skillscommission.org/wp-content/uploads/2010/05/ToughChoices_EXECSUM.pdf

Palincsar, A. S., & Brown, A. L. (1985). Reciprocal teaching: Activities to promote reading with your mind. In T. L. Harris & E. J. Cooper (Eds.), *Reading, thinking and concept development: Strategies for the classroom* (pp. 147–158). New York, NY: College Board.

Palinscar, A. S., Brown, A. L., & Campione, J. (1991). Dynamic assessment. In H. L. Swanson (Ed.), *Handbook on the assessment of learning difficulties* (pp. 75–94). Austin, TX: Pro-Ed.

Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist, 36*, 89–101. doi:10.1207/S15326985EP3602_4

Pelco, L. E., & Reed-Victor, E. (2007). Self-regulation and learning-related social skills: Intervention ideas for elementary school students. *Preventing School Failure, 51*, 36–42. doi:10.3200/PSFL.51.3.36-42

Peña, E. D., Gillam, R., Malek, M., Ruiz-Felter, R., Resendiz, M., Fiestas, C., & Sabel, T. (2006). Dynamic assessment of children from culturally diverse backgrounds: Applications to narrative assessment. *Journal of Speech, Language, and Hearing Research, 49*, 1037–1057. doi:10.1044/1092-4388(2006/074)

Peña, E. D., Iglesias, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology, 10*, 138–154. doi:10.1044/1058-0360(2001/014)

Piaget, J. (1952). *Origins of intelligence* (M. Cook, Trans.). New York, NY: International Universities Press. doi:10.1037/11494–000.

Piaget, J. (1967). *Six psychological studies*. New York, NY: Random House.

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. New York, NY: Springer.

Reschly, D. J., & Robinson-Zañartu, C. A. (2000). Aptitude tests in educational classification and placement. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (3rd ed., pp. 183–201). Oxford, England: Pergamon.

Robinson-Zañartu, C. (2008, November). *Barriers to school psychologists' adopting dynamic assessment: Two issues of relevance*. Paper presented at the meeting of the International Association for Cognitive Education and Psychology, Lake Louise, Alberta, Canada.

Robinson-Zañartu, C. (2009, July). *Dynamic cognitive assessment for useful school-based intervention planning*. Workshop presented at the meeting of the International Association for Cognitive Education and Psychology, Osnabrück, Germany.

Robinson-Zañartu, C., Fiz, F., Genet, M., Mercado, P., & Ornelas, V. J. (2010, February). *Dynamic cognitive assessment for academic intervention: Design, response, report*. Symposium conducted at the meeting of the North American Regional Conference of the International Association for Cognitive Education and Psychology, San Diego, CA.

Robinson-Zañartu, C. A., & Aganza, J. S. (2000). Dynamic assessment and sociocultural context: Assessing the whole child. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 443–487). Oxford, England: JAI/Ablex.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. New York, NY: Oxford University Press.

Scarr, S. (1981). Testing for children: Assessment of the many determinants of intellectual competence. *American Psychologist, 36*, 1159–1166. doi:10.1037/0003-066X.36.10.1159

Shimabukuro, S. M., Prater, M. A., Jenkins, A., & Edelen-Smith, P. (1999). The effects of self-monitoring of academic performance on students with learning disabilities and ADD/ADHD. *Education and Treatment of Children*, *22*, 397–414.

Shurin, R. (1998). *Validity and reliability of the Application of Cognitive Functions Scale with preschool children with disabilities* (Master's thesis). Available from ERIC (TM030312).

Sternberg, R. J. (2000). Prologue. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. xiii–xv). New York, NY: Elsevier Science.

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing*. New York, NY: Cambridge University Press.

Swanson, H. L., & Lussier, C. (2001). A selective synthesis of the experimental literature on dynamic assessment. *Review of Educational Research, 71*, 321–363. doi:10.3102/00346543071002321

Trilling, B., & Fadell, C. (2009). *Twenty-first century skills: Learning for life in our times.* San Francisco, CA: Jossey-Bass.

Tzuriel, D. (2000). Dynamic assessment of young children: Educational and intervention perspectives. *Educational Psychology Review, 12*, 385–435. doi:10.1023/A:1009032414088

Tzuriel, D., & Alfassi, M. (1994). Cognitive and motivational modifiability as a function of instrumental enrichment (IE) program. *Special Services in the Schools, 8*, 91–128. doi:10.1300/J008v08n02_06

Tzuriel, D., & Egozi, G. (2010). Gender differences in spatial ability of young children: The effects of training and processing strategies. *Child Development, 81*, 1417–1430. doi:10.1111/j.1467-8624.2010.01482.x

Tzuriel, D., & Klein, P. (1985). Analogical thinking modifiability in disadvantages, regular, special education and mentally retarded children. *Journal of Abnormal Child Psychology, 13*, 539–552. doi:10.1007/BF00923140

Vygotsky, L. S. (1935). *Umstvennoe razvitie detei v protsesse obucheniia* [The mental development of children in the process of instruction]. Moscow, Soviet Union: Gosudarstvennoe Uchebnopedagogichskoe Izdatel'stvo.

Vygotsky, L. S. (1962). *Thought and language.* Cambridge, MA: MIT Press. doi:10.1037/11193–000 (Original work published 1934)

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press. (Original work published 1934)

Wiedl, K. H. (1980). Kompensatorische Interventionen im Rahman Intelligenz-diagnostischer Untersuchungen bei kognitiv impulsiven Kindern [Compensatory interventions during intellectual assessment of cognitively impulsive children]. *Zeitschrift für Klinische Psychologie, 9*, 219–231.

Wiedl, K. H., & Schoettke, H. (1995). Dynamic assessment of selective attention in schizophrenic subjects: The analysis of intra-individual variability of performance. In J. Carlson (Ed.), *European contributions to dynamic assessment: Vol. 3. Advances in cognition and educational practice* (pp. 185–208). Greenwich, CT: JAI Press.

CHAPTER 8

CURRICULAR ASSESSMENT

Tanya L. Eckert, Adrea J. Truckenmiller, Jennifer L. Rymanowski, Jennifer L. Koehler, Elizabeth A. Koenig, and Bridget O. Hier

Many children in U.S. public schools experience difficulty acquiring basic skills in the content areas of reading, writing, and mathematics. As an example, the most recent national educational assessments have indicated that among fourth-grade students, 68% were not reading at the proficient level (Lee, Grigg, & Donahue, 2007), 72% could not write at the proficient level (Persky, Daane, & Jin, 2003), and 62% were at the basic level in computation (Lee, Grigg, & Dion, 2007). Given that many children experience difficulty acquiring academic skills, it is important for psychologists to include curricular assessments as part of a comprehensive psychological evaluation. In this chapter, we review one academic assessment model, curriculum-based measurement (CBM), that can be used in the context of conducting curricular assessment. This assessment model is used to assess children's basic academic skill development and achievement in reading, writing, and mathematics. We begin by discussing the theoretical concepts related to CBM and then compare CBM with more traditional assessment approaches, such as norm-referenced assessment. Next, we demonstrate how children's academic skills can be assessed systematically by using CBM, which includes universal screenings. Finally, we conclude by discussing how CBM can be used to generate time-series graphs to depict children's academic progress over time and inform instruction in the classroom.

CONDITION OF EDUCATION AND IMPLICATIONS FOR ASSESSMENT PRACTICES

As illustrated previously, many children enrolled in U.S. public schools experience significant difficulty in learning to read, compute, and write. These trends in children's academic achievement become even more pronounced when specific demographic factors are taken into account, such as students' ethnicity or race and eligibility for free or reduced-price lunch. For example, in the area of reading, more than 75% of Black, Hispanic, and American Indian/Alaska Native children could not read at the proficient level (Lee, Grigg, & Dion, 2007). This finding was observed among fourth- and eighth-grade Black children (86% and 88%, respectively), Hispanic children (83% and 86%, respectively), and American Indian/Alaska Native children (80% and 81%, respectively) with considerably lower percentages below the proficient level among fourth- and eighth-grade White children (58% and 62%, respectively) and Asian/Pacific Islander children (55% and 61%, respectively). Among children eligible for free or reduced-price lunch, 83% of fourth- and 85% of eighth-grade children could not read at the proficient level, whereas 56% of fourth- and 61% of eighth-grade children ineligible for free or reduced-price lunch could not read at the proficient level.

The greatest achievement disparities were observed among children who were identified as English language learners or children who were eligible for special education services (Lee, Grigg, &

DOI: 10.1037/14049-008
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education, K. F. Geisinger (Editor-in-Chief)

Dion, 2007). Ninety-three percent of fourth- and 96% of eighth-grade children identified as English language learners had reading scores that fell below the proficient level compared with 66% of fourth and 69% of eighth-grade children not identified as English language learners. Among children eligible for special education services, 87% of fourth- and 93% of eighth-grade children had reading scores that fell below the proficient level compared with children who were not eligible for special education services (66% and 69%, respectively). An identical pattern for ethnicity, free or reduced-price lunch status, special education status, and English language learner status was found in the areas of mathematics and writing. However, in the area of writing, achievement level disparities as a function of primary language spoken or eligibility for special education services were not reported.

Recent longitudinal studies of adult literacy have further substantiated these findings. For example, Baer, Kutner, and Sabatini (2009) reported that as of 2003, only 13% of adults ages 16 or older demonstrated literacy in three domains: prose (e.g., search, read, and comprehend a paragraph), document (e.g., search, read, and comprehend a prescription label), or quantitative (e.g., use numbers embedded in print material). These findings suggest that approximately 11 million adults are nonliterate in English, with 7 million adults unable to answer simple test questions and 4 million more adults unable to participate in academic testing because of significant language barriers.

CURRENT CURRICULAR ASSESSMENT PRACTICES IN EDUCATION

Because many children and youth enrolled in U.S. public schools experience significant difficulty in learning to read, write, and compute, additional attention focused on improving the academic competencies of these at-risk children and youths is critical (Eckert, Truckenmiller, Rheinheimer, Perry, & Koehler, 2008). As early as 1983, efforts by the U.S. Department of Education (e.g., *A Nation at Risk;* National Commission on Excellence in Education, 1983) highlighted the need for educational reform to improve the academic achievement of children and youths. This call for reform, coupled with advances in learning and educational processes (Bransford, Brown, Cocking, Donovan, & Pellegrino, 2000; Kilpatrick, Swafford, & Findell, 2001; Snow, Burns, & Griffin, 1998) as well as numerous federal and policy reform initiatives (e.g., No Child Left Behind Act of 2001; Individuals With Disabilities Education Improvement Act of 2004), resulted in dramatic changes in educational practices.

One educational practice specifically targeted at the federal level was curricular assessment. During much of the past 50 years, curricular assessment practices focused on achievement testing to evaluate schools or individual students in an attempt to make screening, classification, and placement decisions (Deno, 2005). However, dissatisfaction with traditional assessment practices has mounted (Elliott & Fuchs, 1997; L. S. Fuchs, Fuchs, Hamlett, Phillips, & Bentz, 1994; L. S. Fuchs, Fuchs, & Speece, 2002), and increased attention has been directed toward studying the relationship among curriculum, instruction, and assessment (Webb, Horton, & O'Neal, 2002). For example, in a paper commissioned by the National Center on Education and the Economy, Pellegrino (2006) argued,

> Thus, the dollars we now spend on assessment should be reinvested in more targeted and efficacious assessment approaches tied to important curricular goals. These assessments should be meaningful to the individuals assessed and have real value in determining their readiness to move on in the educational system. (p. 2)

Consequently, the Individuals With Disabilities Education Improvement Act of 2004 provided the legal basis for a number of provisions that provide schools the opportunity to directly link curriculum, instruction, and assessment (D. J. Reschly, 2008).

One methodology for aligning curriculum, instruction, and assessment is response to intervention (RtI; Niebling, Roach, & Rahn-Blakeslee, 2008). RtI is a model of education service delivery that provides multitiered, research-based instruction and intervention matched to the needs of students (Batsche, Castillo, Dixon, & Forde, 2008). Key

characteristics of this model include (a) providing all students with research-based instruction; (b) implementing small-group, high-intensity interventions for students demonstrating inadequate progress; and (c) implementing individualized, high-intensity interventions for students continuing to demonstrate inadequate progress (D. Fuchs, Mock, Morgan, & Young, 2003; McMaster, Fuchs, Fuchs, & Compton, 2005; Mellard, Byrd, Johnson, Tollefson, & Boesche, 2004). It has been hypothesized that by aligning curriculum, instruction, and assessment, the number of students experiencing academic difficulties will be reduced (Al Otaiba & Torgensen, 2007), overall student achievement will be enhanced (Ervin, Schaughency, Goodman, McGlinchey, & Matthews, 2006), and special education decision making will be improved (Speece & Case, 2001).

Inherent in this methodology are three overlapping tiers, which "collectively represent a continuum of interventions that increase in intensity based on the corresponding responsiveness" of students (Sugai, 2007, p. 114). The first tier of RtI is referred to as the *universal* or *primary* level because all students receive a core set of classroom instruction designed to foster academic skill development. The second tier of RtI is referred to as the *targeted* or *secondary* level and is made up of more intensive interventions for those students who are not adequately responding to classroom instruction. The third tier is referred to as the *indicated* or *tertiary* level and is characterized by individualized and specialized interventions for those students who do not adequately respond to the universal and targeted tiers of instruction and intervention (Sugai, 2007; Walker et al., 1996). To determine students' academic performance and students' responsiveness to instruction, curricular assessments are a fundamental characteristic of this model. One form of curricular assessment, CBM (Deno, 1985), has been used to systematically assess students' performance in the basic academic skills of reading, writing, and mathematics. CBM can be used to conduct systematic assessments of students' academic performance levels in an effort to identify those students who are not benefiting from classroom instruction (i.e., universal screening) as well as monitor the academic performance of those students who are receiving more intensive interventions (i.e., formative assessments).

THEORETICAL CONCEPTUALIZATION OF CURRICULUM-BASED MEASUREMENT

As previously discussed, CBM is a method of assessment that is intended to link students' academic skills directly to instruction and intervention (Deno, 1985). Initially designed to evaluate students' basic academic skills and assist with instructional decision making, CBM involves the use of repeated, brief measures as a means of assessing student performance over time (Deno, Marston, & Mirkin, 1982; Deno, Marston, & Tindal, 1986; L. S. Fuchs, Deno, & Mirkin, 1984). CBM is conceptualized as a direct method of assessment and uses "direct observation and recording of a student's performance in the local curriculum as a basis for gathering information to make instructional decisions" (Deno, 1987, p. 41).

Expanding on the work of Deno (1984, 1987) and colleagues, Shinn (1989, 1998) created a systematic assessment process using CBM assessment results to inform educational decision making. Shinn's assessment model includes five sequential steps: (a) problem identification, (b) problem certification, (c) exploring solutions, (d) evaluating solutions, and (e) problem solution. These CBM procedures allow for screening, monitoring, and evaluating students' academic performance under a variety of conditions. For example, in the first step, problem identification, CBM is used to evaluate the student's basic academic skills and identify whether a problem exists (i.e., the student is reading at a slower rate than same-grade peers, suggesting a reading problem that warrants further investigation). The second step, problem certification, evaluates the severity of the problem on the basis of the referred student's instructional level or percentile rank compared with that of peers (i.e., the student's reading skills and instructional needs are significant and require intervention). The next step, exploring solutions, requires developing intervention activities to improve the student's deficient skills. Then, the effectiveness of the intervention activities is evaluated in the fourth step, evaluating solutions, which

requires administration of repeated CBM probes. The final stage, problem solution, necessitates periodic administration of CBM probes to evaluate the student's sustained progress and determine whether the intervention activities continue to be needed. In summary, the theoretical conceptualization has allowed school psychologists to use CBM as part of a problem-solving process for individual students as well as part of a schoolwide approach to promote early identification of academic skills deficits in reading, mathematics, and written language.

CBM results can be used to inform a variety of educational decisions, including prereferral classroom decisions (e.g., provision of special help), entitlement decisions (e.g., exceptionality), postentitlement decisions (e.g., instructional planning), and accountability decisions (e.g., program evaluation; Salvia, Ysseldyke, & Bolt, 2007). Over the past 2 decades, numerous research studies (see Deno, 2003; Deno, Fuchs, Marston, & Shin, 2001; Good & Jefferson, 1998; and Shinn, 2008, for overviews) have provided evidence supporting the CBM's reliability and validity for making educational decisions consistent with each type of educational outcome defined by *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Furthermore, several studies have provided evidence supporting the sensitivity of CBM in detecting student improvement over time (Deno et al., 2001; L. S. Fuchs & Fuchs, 1986, 1988, 1999, 2004) and estimating students' performance on state-mandated tests (Hintze & Silberglitt, 2005; Silberglitt & Hintze, 2005). A summary of the psychometric evidence in the content areas of reading, mathematics, and written expression are described in the following sections.

CURRICULUM-BASED MEASUREMENT IN READING

CBM has been evaluated as a measure for several different reading behaviors (i.e., reading words in isolation, reading words in context, oral reading fluency, cloze comprehension, word meaning). Of these behaviors, oral reading fluency has consistently been shown to be the most valid and reliable reading outcome measure among elementary-school-aged children (Deno, Mirkin, & Chiang, 1982; L. S. Fuchs, Fuchs, & Maxwell, 1988; Marston & Magnusson, 1985; Shinn, Good, Knutson, Tilly, & Collins, 1992). For example, in one of the earliest studies to examine CBM reading approaches, Deno, Mirkin, and Chiang (1982) correlated students' oral reading fluency, isolated word-list reading, contextual word-list reading, cloze comprehension scores, and word meaning scores with standardized measures of reading. Validity coefficients reported in this study ranged from .73 to .93, with oral reading fluency resulting in the highest coefficients. Most recently, A. L. Reschly, Busch, Betts, Deno, and Long (2009) conducted a meta-analysis of the correlational association between CBM reading measures and other standardized measures of reading achievement for students in Grades 1 to 6. Results of their study indicated a moderately high correlation (weighted average $r = .67$; $N = 289$ coefficients).

To assess oral reading fluency with CBM, students are given three short reading passages per grade level (see Shapiro, 2004). Standardized CBM reading passages are available from a number of sources, including Pearson Education Incorporated (http://www.aimsweb.com) and the University of Oregon Center on Teaching and Learning (Dynamic Indicators of Basic Early Literacy Skills; Good & Kaminski, 2009). During the assessment, the assessor instructs the student to read each passage aloud using standardized directions (see Shapiro, 2004). Students read each passage for 60 seconds. The assessor begins timing the student using a stopwatch as soon as the student begins reading the first word of the passage. As the student reads aloud, the assessor follows along on a separate copy and scores for errors. An error is recorded if a student mispronounces a word, substitutes another word, omits a word, or does not read a word within 3 seconds. If the student is unable to read a word within 3 seconds, the assessor provides the word. The assessor does not provide any additional instructional feedback while the student is reading the passage. After 60 seconds, the assessor instructs the student to stop reading, and the next passage is administered. The

number of words read correctly by the student during each 1-minute sample is calculated by subtracting the number of errors from the total number of words read. If the student skips an entire line of text, then the number of words contained within the line is subtracted from the total number of words read, and one error is recorded.

Studies of the psychometric properties of CBM in reading have shown test–retest reliability coefficients to be high (*r*s = .92–.97; Tindal, Germann, & Deno, 1983; Tindal, Marston, & Deno, 1983) and parallel form reliability coefficients to also be high (*r*s = .89–.94; Tindal, Germann, & Deno, 1983; Tindal, Marston, & Deno, 1983). The differential predictive validity of CBM in reading has been examined across ethnic groups (Hintze, Callahan, Matthews, Williams, & Tobin, 2002), and the results of a series of hierarchical multiple regression analyses indicated that CBM in reading was not a biased indicator of reading comprehension as a function of grade or ethnicity. Criterion-related validity coefficients comparing oral reading fluency with individually administered reading achievement tests (Marston, 1989) as well as with state-level, high-stakes achievement tests (Hintze & Silberglitt, 2005; McGlinchey & Hixson, 2004; Silberglitt & Hintze, 2005) have been moderate to high (*r*s = .60–.80). Moreover, oral reading fluency has been shown to differentiate general education students, special education students, and remedial education students, with remedial education students performing one grade level below their general education peers and special education students performing two grade levels below their general education peers (Marston & Magnusson, 1985; Shinn & Marston, 1985). The accuracy of using oral reading fluency to classify students as qualifying for special education services has been examined (Shinn, Tindal, Spira, & Marston, 1987). The results of discriminant analyses indicated that 97% of the cases were correctly classified using oral reading fluency scores.

CURRICULUM-BASED MEASUREMENT IN MATHEMATICS

Similar to the procedures described previously for CBM in reading, CBM in mathematics incorporates brief assessments of students' basic mathematics skills. However, unlike most norm-referenced standardized measures in mathematics, multiple assessment forms are available so that formative assessment (i.e., progress monitoring) of student performance can be conducted over the course of a school year. Several different mathematics behaviors have been examined as possible CBM outcome measures, including early number skills, computation (single and mixed operations), concepts and applications, and word-problem solving. More specifically, CBM early number skills target prekindergarten, kindergarten, and first-grade students and include tasks such as counting aloud, identifying numbers, naming missing numbers from a sequence, selecting which of two numbers represents the higher quantity, and circling, drawing, and writing numbers or circles (Chard et al., 2005; Clarke & Shinn, 2004; Vanderheyden et al., 2004; Vanderheyden, Witt, Naquin, & Noell, 2001). For older students (Grades 1–8), CBM outcome measures incorporate mathematical computational tasks (e.g., addition, subtraction, division, multiplication) as well as mathematical concepts and applications, such as reading charts and graphs and solving word problems (Fuchs, Hamlett, & Fuchs, 1998).

Because the CBM early numeracy tasks and the CBM mathematical concepts and applications tasks were developed within the past 10 years, fewer studies have examined the psychometric properties of these tasks. As a result, these measures are often conceptualized as emerging assessment methods (Chard et al., 2005) in CBM mathematics and have primarily been used for screening purposes. Of all the various outcomes developed, computational fluency has been the most frequently used CBM outcome measure (Shapiro, 2004). The computational fluency measures represent skills broadly defined to reflect proficiency in mathematics. These measures are not necessarily representative of the student's mathematics curriculum but rather are characterized as "robust indicators" of overall proficiency in mathematics on the basis of the relative strength of their correlations with general mathematics proficiency criteria or measures (Foegen, Jiban, & Deno, 2007).

According to administration guidelines outlined by Shapiro (2004), to assess a student's computation

fluency using CBM, three mixed-operation probes per grade level are administered and one single-operation probe per skill can be administered as well. Standardized CBM mixed-skill probes are available from a number of sources, including Pearson Education (http://www.aimsweb.com), Sopris West (Dynamic Indicators of Basic Early Literacy Skills; Good & Kaminski, 2009), or Pro-Ed (Monitoring Basic Skills Progress; L. S. Fuchs et al., 1998). The assessor instructs the student to work each problem using standardized directions. As soon as the student begins working the first problem, the assessor begins timing the student with a stopwatch. Students are permitted between 2 and 6 minutes to complete the probe, depending on skill (e.g., addition or multiplication) or grade. The number of digits in the answer computed correctly by the student is calculated by subtracting the number of errors from the total number of digits computed. Responses scored as correct include (a) individual digits, (b) place holder numbers, and (c) digits below the answer line. The following responses are scored as incorrect: (a) incorrect digits, (b) digits that were correct but appear in the wrong place value, and (c) omitted digits. To assess student performance using concepts and application probes, one probe is administered to a group of students who are required to work on the problems for 6 to 8 minutes (Foegen et al., 2007). Scoring requires the assessor to count the total number of blanks on the probe (many problems are multifaceted) and provide 1 point credit for each correctly completed blank. The number of correct answers is divided by the total answers possible to yield the percentage of correct points.

In recent years, more attention has been paid to investigating the psychometric properties of CBM in mathematics. A synthesis of CBM mathematics measures conducted by Foegen et al. (2007) reported the high test–retest and alternate-form reliability ($r > .80$) across outcome measures for computational fluency. Criterion-related validity coefficients, comparing CBM computational fluency measures with standardized mathematics achievement tests, ranged between .55 and .93, with the majority of coefficients falling between .60 and .80. However, empirical work by Hintze, Christ, and Keller (2002) has indicated that variability is greater for CBM mixed-operation probes than for single-operation probes. Most recently, Lembke, Foegen, Whittaker, and Hampton (2008) provided psychometric support for the CBM early number skills measures by modeling weekly growth rates for kindergarten and first-grade students. The results of their work indicated that number identification was found to have the highest estimated weekly growth rate, followed by quantity discrimination and missing numbers.

CURRICULUM-BASED MEASUREMENT IN WRITTEN LANGUAGE

Within the content area of written expression, several different writing behaviors (i.e., total words written, words spelled correctly, and correct writing sequences) have been evaluated (Espin et al., 2000). To assess a student's writing fluency using CBM, the student is asked to write three short stories. The student is provided with three story starters, which contain short sentence fragments that provide an idea to the student for writing a narrative story (e.g., "I was on my way home from school and . . ."). Similar to the conceptualization of CBM computational fluency measures, the writing behaviors assessed on the story starters represent skills broadly defined to reflect proficiency in written expression on the basis of the relative strength of their correlations with measures of written expression. However, unlike norm-referenced standardized measures in written expression, multiple forms are available so that formative assessment of students' writing skills can be assessed over the course of a school year. Examples of story stems are available from a number of sources, including Pearson Education (http://www.aimsweb.com) and Sopris West (Dynamic Indicators of Basic Early Literacy Skills; Good & Kaminski, 2009) or in the published empirical literature (McMaster & Campbell, 2006).

During assessment, the student is given 1 minute to engage in quiet story planning and 3 minutes to write the story (Shapiro, 2004). As soon as the student begins story writing, the assessor begins timing. The assessor prompts the student to continue writing for the entire 3 minutes. After 3 minutes, the assessor instructs the student to stop writing, and

the next story stem is administered. The total number of words written is calculated by counting every grouping of letters that is separated by a space regardless of spelling, whereas the total number of words spelled correctly is calculated by counting every word that is correctly spelled. The total number of correct writing sequences is calculated by analyzing each sentence for correct punctuation, capitalization, spelling, and syntax.

Two comprehensive reviews (McMaster & Espin, 2007; Powell-Smith & Shinn, 2004) provided a synthesis of the psychometric properties of CBM writing assessments. A wide range of alternate-form reliability coefficients have been reported for the total number of words written (*r*s = .56–.95, *Mdn* = .70), the total number of words spelled correctly (*r*s = .53–.95, *Mdn* = .72), and the total number of correct writing sequences (*r*s = .46–.80, *Mdn* = .75); the median coefficients for these outcomes are in the moderate range. A similar pattern of results was reported for the test–retest reliability evaluations of the total number of words written (*r*s = .42–.91, *Mdn* = .65) and the total number of words spelled correctly (*r*s = .46–.81, *Mdn* = .67). Students' performance was compared with standardized, norm-referenced measures of writing achievement, as well as holistic measures of students' writing, and a wide range of validity coefficients have been reported for the total number of words written (*r*s = .13–.84, *Mdn* = .47), the total number of words spelled correctly (*r*s = .17–.84, *Mdn* = .51), and the total number of correct writing sequences (*r*s = .29–.65, *Mdn* = .54). The median coefficients for the criterion-related validity studies are in the low to moderate range. Given the higher median reliability and validity correlations reported for the total number of correct writing sequences, this metric has been recommended for assessing students' writing fluency (Gansle, Noell, Vanderheyden, Naquin, & Slider, 2002; Hubbard, 1996).

APPLICATIONS OF CURRICULUM-BASED MEASUREMENT IN SCHOOL SETTINGS

Children's academic skills can be assessed systematically by using CBM as part of a comprehensive psychological assessment, and CBM results can be used to inform a variety of decisions that are made in school settings. Typically, applications of CBM in school settings include schoolwide assessments (i.e., universal screenings) of all elementary-aged students as well as more targeted, formative assessments (i.e., progress monitoring) of individual students experiencing academic skills problems.

Universal Screenings

CBM can be used as part of universal screening assessments that occur in school settings to systematically examine children's performance within classrooms, grade levels, and school buildings or at the school district level. Universal screenings provide important information regarding whether additional instructional supports or supplemental procedures are needed in the core curriculum. For example, if the analysis of universal screening data suggests that more than 20% of the students in the general education classroom are not making acceptable progress in relation to school benchmarks, then improvements need to be made to the core curriculum or delivery of the core curriculum (National Association of State Directors of Special Education, 2006). In addition, the results of universal screening data can be used to identify students who need additional instruction or intervention beyond the core curriculum, supplemental procedures used in the general education classrooms, or both (Ikeda, Neessen, & Witt, 2008). These students may initially receive small-group, supplemental research-based interventions (Tier 2) in addition to the core curriculum provided in the general education classroom. Students who do not demonstrate a sufficient rate of improvement in response to Tier 2 interventions are then provided with intensive, individualized, research-based interventions (Tier 3) in addition to the core curriculum provided in the general education classroom. As previously discussed, universal screening assessments are commonly adopted in schools using RtI practices to identify elementary-school-aged children at risk for academic failure.

To further illustrate how CBM is used as part of universal screening, the performance of students in a fictitious fifth-grade classroom is illustrated in Figure 8.1. The universal screening results, which represent the students' oral reading fluency using CBM,

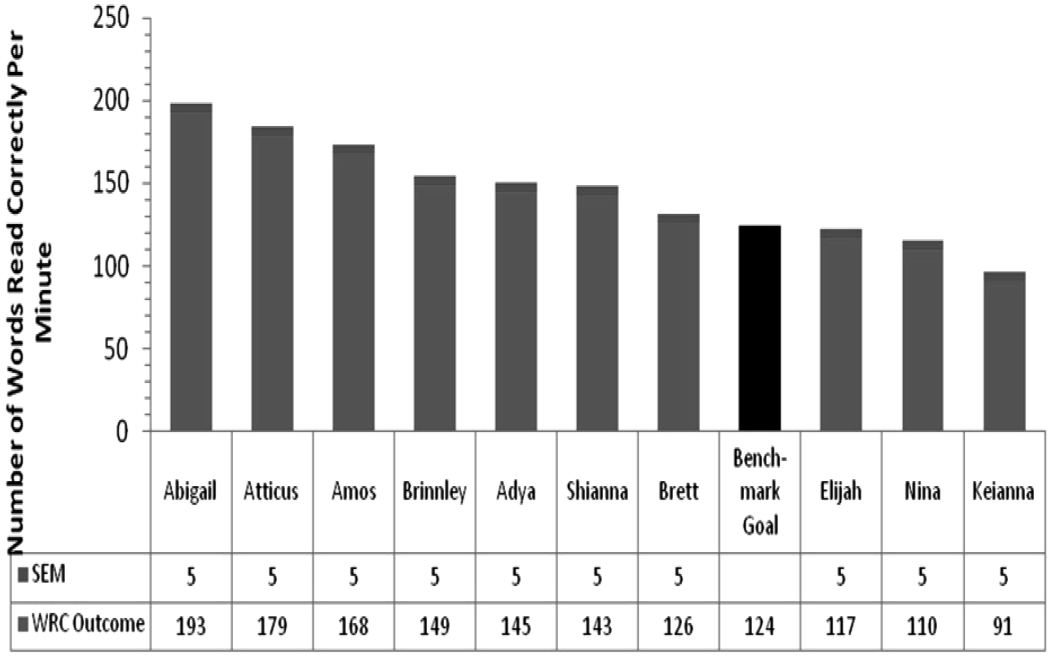


FIGURE 8.1. Universal screening results for a hypothetical fourth-grade classroom's performance using curriculum-based measurement for reading. The students' reading performance (i.e., number of words read correctly [WRC] outcome) is reported in relation to the benchmark goal for fourth-grade students. The standard error of measurement associated with this benchmark passage is included.

are sorted and display the rank ordering of students. The results of this universal screening assessment indicated that seven students (Abigail, Atticus, Amos, Brinnley, Adya, Shianna, and Brett) exceeded the spring benchmark goal of 124 words read correctly per minute, are evidencing growth, and appear to be benefitting from the core curriculum. However, the remaining three students in the classroom fell below the spring benchmark, suggesting that these students are not benefitting from the core curriculum. In addition, given that Brett's oral reading fluency closely approximates the spring benchmark goal and could potentially fall below the spring benchmark goal if his score was adjusted on the basis of the standard error of measurement (±5 words read correctly per minute), all four students (Brett, Elijah, Nina, and Keianna) are ideal candidates for CBM progress monitoring, which is described in the next section of this chapter. Furthermore, all four students' reading progress should be closely monitored over an 8- or 10-week period and supplemental procedures (Tier 2) should be used in an effort to improve their reading progress.

Typically, the recommendation is that CBM probes be administered twice weekly over an 8- or 10-week period, and the resulting data points are graphed using simple line graphs. This recommendation is based on numerous research studies validating progress monitoring schedules (L. S. Fuchs et al., 1984; L. S. Fuchs, Fuchs, & Hamlett, 1989; L. S. Fuchs, Fuchs, Hamlett, & Allinder, 1991; L. S. Fuchs, Fuchs, Hamlett & Ferguson, 1992; L. S. Fuchs, Fuchs, Hamlett, & Stecker, 1991; Stecker, & Fuchs, 2000; Whinnery & Stecker, 1992).

Progress Monitoring

CBM results can also be used to evaluate students' progress over time. Specifically, CBM probes in reading, mathematics, or written expression are repeatedly administered to assess whether students are making gains in their basic academic skills (Marston & Magnusson, 1985; Shapiro, 2004). Calendar days or weeks are labeled on the horizontal axis (*x*-axis) and the student's fluency (e.g., number of words read correctly per minute) is labeled on the vertical axis (*y*-axis). The student's data are

compared with an aim line, which can be derived by the classroom teacher on the basis of estimated performance (e.g., classroom normative data), predicted performance (e.g., average learning rates based on normative data), or mandated performance (e.g., state or national criteria). This approach allows educators to compare the student's expected reading performance (i.e., aim line) with the student's actual reading performance (i.e., CBM data points). An alternative approach involves calculating a trend line on the basis of the student's actual reading performance (i.e., CBM data points), which provides a statistical summary of the student's actual performance over time. This approach permits the computation of a slope value, which serves as an index of the student's progress over time. A more meaningful index of student progress, weekly gain, can be computed by multiplying the slope value by 7 (Parker, Tindal, & Stein, 1992; Shinn, Good, & Stein, 1989); a value of 7 has conventionally been used to reflect calendar days (Hintze, Daly, & Shapiro, 1998).

Figure 8.2 provides an example of CBM progress monitoring results in the content area of writing for a hypothetical second-grade student (i.e., Theo). In this case illustration, a goal of 26 words correct per 3 minutes at the end of the 10-week period was established and is indicated on the graph. The line represents Theo's goal line and was selected on the basis of normative data collected by the school district. In the winter of second grade, 26 words written per 3 minutes is the average number of words written by students ranked at the 50th percentile. In addition, an aim line was constructed on the basis of Theo's initial baseline performance and his expected performance at the end of the intervention. For Theo to meet the long-term goal of writing 26 total words in 3 minutes by the end of a 10-week period, Theo needs to improve his writing fluency by two words each week (26 words written per 3 minutes [goal] − six words written per 3 minutes [baseline] = 20 / 10 weeks = two words written per 3 minutes).

As data were collected over time, the anticipation was that the intervention would improve Theo's writing fluency, and his performance would match the aim line that appears on the graph. Interestingly, Theo's response to the intervention was immediate and strong. He displayed an increasing trend in his writing fluency, which stabilized over the course of the intervention. As a result, the data displayed in

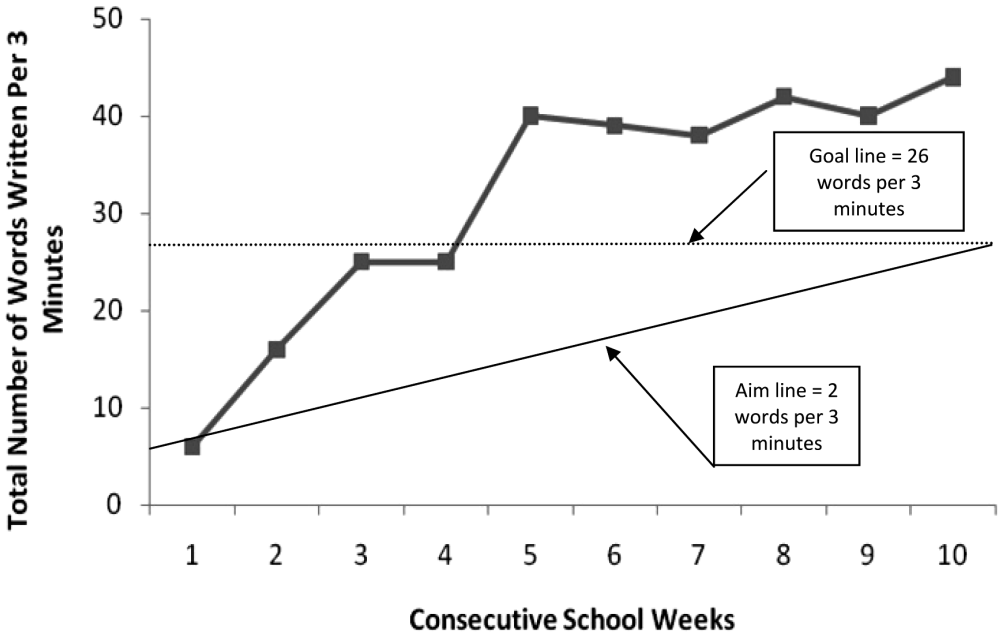


FIGURE 8.2. A hypothetical student's CBM writing progress monitoring results over an 8-week period. The student's writing performance over time is represented by the data path containing circles, and the expected weekly growth is represented by the aim line.

Figure 8.2 indicate that Theo is making greater progress than would be expected given his baseline writing performance.

To further assess Theo's long-term progress monitoring performance, an ordinary least squares trend estimation was conducted, which provides a mathematical estimation of Theo's writing progress over time. Theo's estimate was high and positive (i.e., slope estimate = 0.75). This estimate can be translated into an evaluation of Theo's weekly improvement in writing, which suggests that Theo's writing fluency improved by 5.25 words each week. These results further support the conclusion that Theo's improvements in writing are exceeding the slope of improvement (i.e., two words per week) that was needed to attain the school normative goal of 26 words written per 3 minutes. In conclusion, over the course of 10 weeks of intervention, Theo made considerable progress in his writing fluency.

CONCLUSION

It is important that psychologists include curricular assessments as part of a comprehensive psychological assessment. One academic assessment model, CBM, is frequently used by school psychologists in the context of conducting curricular assessment in the content areas of reading, writing, and mathematics. For example, Shapiro, Angello, and Eckert (2004) reported in a national survey of school psychologists that 54% of respondents reported using CBM in school-based practice. CBM can be used as part of universal screening assessments to systematically examine children's performance within classrooms, grade levels, and school buildings or at the school district level. Universal screenings provide important information regarding the students' academic performance and afford psychologists the opportunity to identify elementary-school-aged children at risk for academic failure. In addition, CBM can also be used to evaluate students' progress over time by repeatedly assessing students' academic skills over time and presenting the findings using simple line graphs. These measurement tools provide psychologists with an assessment of students' academic skills at the classroom level as well as at the student level. Moreover, the resulting assessment data can be used to inform instructional changes for students who are experiencing academic difficulties.

References

Al Otaiba, S., & Torgensen, J. (2007). Effects from intensive standardized kindergarten and first-grade interventions for the prevention of reading difficulties. In S. E. Jimerson, M. K. Burns, & A. M. Vanderheyden (Eds.), *Handbook of response to intervention* (pp. 212–222). New York, NY: Springer.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Baer, J., Kutner, M., & Sabatini, J. (2009). *Basic reading skills and the literacy of America's least literate: Results from the 2003 National Assessment of Adult Literacy (NAAL) Supplemental Studies* (NCES 2009–481). Washington, DC: National Center for Education Statistics.

Batsche, G. M., Castillo, J. M., Dixon, D. N., & Forde, S. (2008). Best practices in linking assessment to intervention. In A. Thomas & A. J. Grimes (Eds.), *Best practices in school psychology V* (Vol. 2, pp. 177–194). Bethesda, MD: National Association of School Psychologists.

Bransford, J. D., Brown, A. L., Cocking, R. R., Donovan, S., & Pellegrino, J. W. (Eds.). (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academies Press.

Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention, 30*(2), 3–14. doi:10.1177/073724770503000202

Clarke, B., & Shinn, M. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33*, 234–248.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.

Deno, S. L. (1987). Curriculum-based measurement. *Teaching Exceptional Children, 20*, 41.

Deno, S. L. (2003). Developments in curriculum-based measurement. *Journal of Special Education, 37*, 184–192. doi:10.1177/00224669030370030801

Deno, S. L. (2005). Problem-solving assessment. In R. Brown-Chidsey (Ed.), *Assessment for intervention* (pp. 10–40). New York, NY: Guilford Press.

Deno, S. L., Fuchs, L. S., Martson, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*, 507–524.

Deno, S. L., Marston, D., & Mirkin, P. K. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children, 48*, 368–371.

Deno, S. L., Marston, D., & Tindal, G. (1986). Direct and frequent curriculum-based measurement: An alternative for educational decision making. *Special Services in the Schools, 2*, 5–27. doi:10.1300/J008v02n02_02

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36–45.

Eckert, T. L., Truckenmiller, A. J., Rheinheimer, J. L., Perry, L. J., & Koehler, J. L. (2008). Improving children's academic performance: Benefits and barriers associated with fluency-based interventions. In D. H. Molina (Ed.), *School psychology: 21st century issues and challenges* (pp. 327–343). Hauppauge, NY: Nova Sciences.

Elliott, S. N., & Fuchs, L. S. (1997). The utility of curriculum-based measurement and performance assessment as alternatives to traditional intelligence and achievement tests. *School Psychology Review, 26*, 224–233.

Ervin, R. A., Schaughency, E., Goodman, S. D., McGlinchey, M. T., & Matthews, A. (2006). Merging research and practice agendas to address reading and behavior school-wide. *School Psychology Review, 35*, 198–223.

Espin, C. A., Skare, S., Shin, J., Deno, S. L., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *Journal of Special Education, 34*, 140–153. doi:10.1177/002246690003400303

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *Journal of Special Education, 41*, 121–139. doi:10.1177/00224669070410020101

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research and Practice, 18*, 172–186. doi:10.1111/1540-5826.00073

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measures and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21*, 449–460.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*, 199–208.

Fuchs, L. S., & Fuchs, D. (1988). Curriculum-based measurement: A methodology for evaluating and improving student programs. *Diagnostique, 14*, 3–13.

Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review, 28*, 659–671.

Fuchs, L. S., & Fuchs, D. (2004). *What is scientifically based research on progress monitoring?* Washington, DC: National Center on Progress Monitoring.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Effects of instrumental use of curriculum-based measurement to enhance instructional programs. *Remedial and Special Education, 10*, 43–52. doi:10.1177/074193258901000209

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Allinder, R. (1991). The contribution of skills analysis to curriculum-based measurement in spelling. *Exceptional Children, 57*, 443–452.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement using a reading maze task. *Exceptional Children, 58*, 436–450.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Phillips, N. B., & Bentz, J. (1994). Classwide curriculum-based measurement: Helping general educators meet the challenge of student diversity. *Exceptional Children, 60*, 518–537.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal, 28*, 617–641.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*, 20–28. doi:10.1177/074193258800900206

Fuchs, L. S., Fuchs, D., & Speece, D. L. (2002). Treatment validity as a unifying concept for the identification of learning disabilities. *Learning Disability Quarterly, 25*, 33–45. doi:10.2307/1511189

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1998). *Monitoring basic skills progress: Basic math computation* (2nd ed.) [Computer software, manual, and blackline masters]. Austin, TX: Pro-Ed.

Gansle, K. A., Noell, G. H., Vanderheyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review, 31*, 477–497.

Good, R. H., & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 61–88). New York, NY: Guilford Press.

Good, R. H., & Kaminski, R. A. (2009). *Dynamic Indicators of Basic Early Literacy Skills*. Longmont, CO: Sopris West.

Hintze, J. M., Callahan, J. E., III, Matthews, W. J., Williams, S. A. S., & Tobin, K. G. (2002). Oral reading fluency and prediction of reading comprehension in African American and Caucasian elementary school children. *School Psychology Review, 31*, 540–553.

Hintze, J. M., Christ, T. J., & Keller, L. A. (2002). The generalizability of CBM survey-level mathematics assessments: Just how many samples do we need? *School Psychology Review, 31*, 514–528.

Hintze, J. M., Daly, E. J., & Shapiro, E. S. (1998). An investigation of the effects of passage difficulty level on outcomes of oral reading fluency progress monitoring. *School Psychology Review, 27*, 433–445.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34*, 372–386.

Hubbard, D. D. (1996). *Technical adequacy of formative monitoring systems: A comparison of three curriculum-based indices of written expression*. Unpublished doctoral dissertation, University of Oregon, Eugene.

Ikeda, M. J., Neessen, E., & Witt, J. C. (2008). Best practices in universal screening. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 103–114). Bethesda, MD: National Association of School Psychologists.

Individuals With Disabilities Education Improvement Act of 2004, 20 U.S. C. § 1400 *et seq*.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.

Lee, J., Grigg, W., & Dion, G. (2007). *The nation's report card: Mathematics 2007* (NCES 2007-494). Washington, DC: National Center for Education Statistics.

Lee, J., Grigg, W., & Donahue, P. (2007). *The nation's report card: Reading 2007* (NCES 2007-496). Washington, DC: National Center for Education Statistics.

Lembke, E. S., Foegen, A., Whittaker, T. A., & Hampton, D. (2008). Establishing technically adequate measures of progress in early numeracy. *Assessment for Effective Intervention*, *33,* 206–214. doi:10.1177/1534508407313479

Marston, D. B. (1989). A curriculum-based measurement approach to assessing academic performance: What is it and why do it? In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York, NY: Guilford Press.

Marston, D. B., & Magnusson, D. (1985). Implementing curriculum-based measurement in special and regular education settings. *Exceptional Children, 52*, 266–276.

McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*, 193–203.

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing. *Journal of Special Education, 41*, 68–84. doi:10.1177/00224669070410020301

McMaster, K. L., & Campbell, H. (2006). *Advances in monitoring progress in writing*. Paper presented at the Progress Monitoring Conference, Minneapolis, MN.

McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children, 71*, 445–463.

Mellard, D. F., Byrd, S. E., Johnson, E., Tollefson, J. M., & Boesche, L. (2004). Foundations and research on identifying model responsiveness-to-intervention sites. *Learning Disability Quarterly, 27*, 243–256. doi:10.2307/1593676

National Association of State Directors of Special Education. (2006). *Response to intervention: Policy considerations and implementation*. Alexandria, VA: Author.

Niebling, B. C., Roach, R. T., & Rahn-Blakeslee, A. (2008). Best practices in curriculum, instruction, and assessment alignment. In A. Thomas & T. Grimes (Eds.), *Best practices in school psychology V* (pp. 1059–1072). Bethesda, MD: National Association of School Psychologists.

No Child Left Behind Act of 2001, Pub. L. 107–110, U.S.C. 115 Stat. 1425 (2002).

Parker, R., Tindal, G., & Stein, S. (1992). Estimating trend in progress monitoring data: A comparison of simple line-fitting methods. *School Psychology Review, 21*, 300–312.

Pellegrino, J. W. (2006). *Rethinking and redesigning curriculum, instruction and assessment: What contemporary research and theory suggest*. Retrieved from http://www.activelearner.ca/activelearning_media/Pellegrino-Redesigning.pdf

Persky, H. R., Daane, M. C., & Jin, Y. (2003). *The nation's report card: Writing 2002* (NCES 2003-529). Washington, DC: National Center for Education Statistics.

Powell-Smith, K. A., & Shinn, M. R. (2004). *Administration and scoring of written expression*

curriculum-based measurement (WE-CBM) for use in general outcome measurement. Retrieved from http://www.slideserve.com/cashlin/administration-and-scoring-of-written-expression-curriculum-based-measurement-we-cbm

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*, 427–469. doi:10.1016/j.jsp.2009.07.001

Reschly, D. J. (2008). School psychology paradigm shift and beyond. In A. Thomas & A. J. Grimes (Eds.), *Best practices in school psychology V* (Vol. 1, pp. 3–15). Bethesda, MD: National Association of School Psychologists.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2007). *Assessment in special and inclusive education* (10th ed.). Boston, MA: Houghton Mifflin.

Shapiro, E. S., Angello, L. M., & Eckert, T. L. (2004). Has curriculum-based assessment become a staple of school psychology practice? An update and extension of knowledge, use and attitudes from 1990 to 2000. *School Psychology Review, 33*, 249–257.

Shinn, M. R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children.* New York, NY: Guilford Press.

Shinn, M. R. (Ed.). (1998). *Advanced applications of curriculum-based measurement.* New York, NY: Guilford Press.

Shinn, M. R. (2008). Best practices in using curriculum-based measurement in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 243–262). Bethesda, MD: National Association of School Psychologists.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*, 459–479.

Shinn, M. R., Good, R. H., & Stein, S. (1989). Summarizing trends in student achievement: A comparison of methods. *School Psychology Review, 18*, 356–370.

Shinn, M. R., & Marston, D. (1985). Differentiating mildly handicapped, low-achieving, and regular education students: A curriculum-based approach. *Remedial and Special Education, 6*, 31–38. doi:10.1177/074193258500600207

Shinn, M. R., Tindal, G., Spira, D., & Marston, D. (1987). Practice of learning disabilities as social policy. *Learning Disability Quarterly, 10*, 17–28. doi:10.2307/1510751

Silberglitt, B., & Hintze, J. M. (2005). Formative assessment using CBM-R cut score to track progress toward success on state mandated achievement tests. *Journal of Psychoeducational Assessment, 23*, 304–325. doi:10.1177/073428290502300402

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academies Press.

Speece, D. L., & Case, L. (2001). Classification in context: An alternative to identifying early reading disability. *Journal of Educational Psychology, 93*, 735–749. doi:10.1037/0022-0663.93.4.735

Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research and Practice, 15*, 128–134. doi:10.1207/SLDRP1503_2

Sugai, G. (2007). Promoting behavioral competence in schools: A commentary on exemplary practices. *Psychology in the Schools, 44*, 113–118. doi:10.1002/pits.20210

Tindal, G., Germann, G., & Deno, S. L. (1983). *Descriptive research on the Pine County norms: A compilation of findings* (Research Report No. 132). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Tindal, G., Marston, D., & Deno, S. L. (1983). *The reliability of direct and repeated measurement* (Research Report No. 109). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Vanderheyden, A. M., Broussard, C., Fabre, M., Stanley, J., Legendre, J., & Creppell, R. (2004). Development and validation of curriculum-based measures of math performance for preschool children. *Journal of Early Intervention*, *27*, 27–41. doi:10.1177/105381510402700103

Vanderheyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review, 30*, 363–382.

Walker, H. M., Horner, R. H., Sugai, G., Bullis, M., Sprague, J. R., Bricker, D., & Kaufman, M. J. (1996). Integrated approaches to preventing antisocial behavior patterns among school-age children and youth. *Journal of Emotional and Behavioral Disorders, 4*, 194–209. doi:10.1177/106342669600400401

Webb, N. L., Horton, M., & O'Neal, S. (2002, April). *An analysis of the alignment between language arts standards and assessments for four states.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Whinnery, K. W., & Stecker, P. M. (1992). Individual progress monitoring to enhance instructional programs in mathematics. *Preventing School Failure, 36*(2), 26–29. doi:10.1080/1045988X.1992.9944266

CHAPTER 9

ADAPTIVE BEHAVIOR: ITS HISTORY, CONCEPTS, ASSESSMENT, AND APPLICATIONS

Thomas Oakland and Matthew Daley

Adaptive behavior refers to an individual's independent display of behaviors associated with meeting his or her daily personal and social needs, including behaviors expected in domestic and social environments (Nihira, Leland, & Lambert, 1993). The behaviors that make up the construct of adaptive behavior have a pervasive effect on people's quality of life, including the ability to function independently at school, work, and home and in the community. This chapter describes adaptive behavior, traces its history, and summarizes professional and legal standards that govern its assessment and use.[1] Considerable attention is given to theories of adaptive skill development in children from birth through age 5 and to developmental data derived from three standardized tests. Research on the impact of intellectual and other disabling conditions on adaptive behavior is summarized. General guidelines for the assessment of adaptive behavior are discussed, followed by a review of three comprehensive measures of adaptive behavior.

GENERAL DESCRIPTION AND HISTORY OF ADAPTIVE BEHAVIOR

Think for a moment of the adaptive skills you may have displayed at one time or another today. After arising, you may have bathed, dressed, eaten, taken vitamins or other medications, and planned your day. You may then have communicated and socialized with others; used your previous school-acquired knowledge; worked at home or elsewhere; cared for your home, family members, and office; and used community resources (e.g., taxis, buses, post office). These combined adaptive skills are both time-tested and universal indicators of how well you take personal responsibility for your welfare and engage your environment.

These and other adaptive behaviors have historically been used to judge people's applied intelligence or ability to adapt to their environment. The ancient Greek civilization may have been the first to formally consider diminished adaptive behavior to reflect mental retardation[2]—a tradition that continues and has been formalized as part of the diagnostic

Sara S. Sparrow, PhD, professor emerita of psychology and chief psychologist at Yale University's Child Study Center from 1977 to 2002, was to write this chapter. Sadly, she passed away on June 10, 2010, after a long illness. Sparrow was the author of more than 100 articles and chapters on psychological assessments and developmental disabilities and was senior author of the Vineland Adaptive Behavior Scales. Her research focused on the assessment of adaptive behavior, child neuropsychology, and developmental disabilities across a wide range of diagnostic groups of children and also across cultures. Psychology has lost a very able scholar who contributed much to the understanding of children and youth, including those with autism spectrum disorders, intellectual disability, and emotional disorders and gifted children. She was active for decades in the training of mental health professionals at the doctoral and postdoctoral levels.

[1]See Oakland and Harrison (2008) for a more complete discussion of the history of adaptive behavior, especially its association with the concept and assessment of mental retardation (e.g., intellectual disability).

[2]The term *mental retardation* has been changed to *intellectual disability* by some agencies and in some policy statements. The American Association on Mental Retardation announced on November, 2, 2006, that it had changed its name to the American Association on Intellectual and Developmental Disabilities. The term *mental retardation* is used in this chapter, when appropriate, to reflect its historic use.

DOI: 10.1037/14049-009
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

criteria in the United States and elsewhere. Esquirol's introduction in the 1800s of the term *idiot,* used clinically until the 20th century to describe people with mental retardation, is derived from a Greek word that signifies people who do not engage in the public life of the community. People whose self-care and community engagement were similar to others their age were thought to be normal; those individuals whose self-care and engagement were considerably lower were thought to display mental retardation.

The humanitarian and education-focused efforts that emerged during and after the Enlightenment period emphasized the need for more formal and objective methods to distinguish diminished functioning from normal development as well as to differentiate various levels of mental retardation. Binet and Simon developed one of the first widely accepted, comprehensive, objective, and standardized methods to assess children's intelligence. They recognized intelligence as a personal quality needed for people to competently engage in important daily life activities. In *Development of Intelligence of Children*, Binet and Simon (1912) wrote,

> An individual is normal if he is able to conduct his affairs of life without having need of the supervision of others, if he is able to do work sufficiently remunerative to support his own personal needs, and finally if his intelligence does not unfit him for the social environment of his parents. (p. 88)

Parallels between Binet's views as to the importance of adaptive behavior and those expressed 18 centuries earlier by the Greeks are obvious. Moreover, psychologists continue to use similar standards in the assessment of adaptive behavior skills in the 21st century.

The development and use of intelligence tests during the first 3 decades of the 20th century led to casting aside the long-held importance of adaptive behavior in favor of an exclusive reliance on data from intelligence tests to diagnose mental retardation. The American Association on Mental Retardation (AAMR) had a long history of leadership in defining mental retardation and adaptive behavior. However, although its early definitions of mental retardation emphasized the incurability of cognitive deficiencies, the association did not address adaptive behavior. The association's fifth definition of mental retardation (Heber, 1959) broadened the condition of mental retardation to include subaverage general intellectual functioning that originates during the developmental period and is associated with impairment in one or more of the following additional areas: maturation, learning, and social adjustment. A couple of years later, AAMR (2001) offered its first approved definition of *adaptive behavior* (Heber, 1961):

> the effectiveness with which the individual copes with the natural and social demands of the environment. It has two major facets: (a) the degree to which the individual is able to function and maintain himself or herself independently, and (b) the degree to which he or she meets satisfactorily the culturally imposed demands of personal and social responsibility. (p. 21)

This definition launched the joint use of measures of intelligence and adaptive behavior when diagnosing mental retardation and established a practice that continues today.

CURRENT DEFINITIONS OF *ADAPTIVE BEHAVIOR*

AAMR had a tradition of redefining adaptive behavior over the years. For example, the 1992 definition highlighted the importance of adaptive skills, not merely the broader construct of adaptive behavior. *Adaptive skills* were defined as "an array of competencies that reflect both the ability to fit into a given niche as well as the ability to change one's behavior to suit the demands of the situation" (AAMR, 2002, p. 22). As part of this more focused definition, AAMR (2002) identified 10 adaptive skills as critical to a person's adaptation: communication, self-care, home living, social skills, community use, self-direction, health and safety, functional academics, leisure, and work.

The following is the 2002 AAMR definition of adaptive behavior and skills:

> Adaptive behavior is the collection of conceptual, social, and practical skills that have been learned by people in order to function in their everyday lives. Limitations in adaptive behavior affect both daily life and the ability to respond to life changes and environmental demands, and should be considered in light of four other dimensions: Intellectual Abilities; Participation, Interaction, and Social Roles; Health; and Context. Significant limitations in adaptive behavior can be established only through the use of standardized measures normed on the general population, including people with disabilities and people without disabilities, and are defined as performance that is at least two SDs below the M of (a) one of the following three types of adaptive behavior: conceptual, social, or practical, or (b) an overall score on a standardized measure of conceptual, social, and practical skills. (AAMR, 2002, p. 23)

Conceptual skills include receptive and expressive language, reading and writing, and self-direction. Social skills include responsibility, obeying rules and laws, naiveté, and competence in interpersonal interactions. Practical skills include personal and instrumental self-care activities such as toileting, taking medication, dressing, preparing meals, eating, using the telephone, managing money, and using transportation as well as occupational skills and maintaining a safe environment (AAMR, 2002).

In 2010, AAIDD altered its definition of adaptive behavior to emphasize conceptual, social, and practical domains:

> Adaptive behavior is the collection of conceptual, social, and practical skills that have been learned and are performed by people in their everyday lives. For the diagnosis of intellectual disability, significant limitations in adaptive behavior should be established through the use of standardized measures normed on the general population. On these standardized measures, significant limitations in adaptive behavior are operationally defined as performance that is approximately 2 standard deviations below the mean of either (a) one of three types of adaptive behavior, conceptual, social, or practical, or (b) an overall score on a standardized measure of conceptual, social, and practical skills. The assessment instrument's standard error of measurement must be considered when interpreting the individual's score. (AAIDD, 2010, p. 41)

PROFESSIONAL AND LEGAL STANDARDS FOR THE USE OF MEASURES OF ADAPTIVE BEHAVIOR

The use of measures of adaptive behavior is often governed by professional and legal standards. Professional standards include those from the American Association on Intellectual and Developmental Disabilities (AAIDD), the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), and the World Health Organization's (2001) *International Classification of Functioning, Disability and Health* (ICF). Legal standards range from those that govern test use with young students to those that determine whether inmates on death row may be executed. Professional and legal standards are reviewed next (Oakland & Harrison, 2008).

Professional Standards

Psychologists and other professionals frequently administer measures of adaptive behavior when evaluating clients for possible mental retardation or intellectual disability. Their work is typically guided by definitions promulgated by the AAMR–AAIDD and the *DSM*.

AAMR–AAIDD's definition of *mental retardation.* The AAMR's 10th and AAIDD's current definition of

mental retardation includes both intellectual abilities and adaptive functioning as part of the diagnostic criteria. The AAMR (2002) defined *mental retardation* as "a disability characterized by significant limitations both in intellectual functioning and in adaptive behavior as expressed in conceptual, social, and practical adaptive skills. This disability originates before age 18" (p. 1).

Diagnostic and Statistical Manual of Mental Disorders. The *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text revision; *DSM–IV–TR;* American Psychiatric Association, 2000) provides the most authoritative guide on mental health diagnoses. It defines *mental retardation* as the display of significantly subaverage intelligence (i.e., an IQ of approximately 70 or lower) along with concurrent deficits or impairments in present adaptive functioning, with an onset occurring before age 18. *Adaptive functioning* is defined as the person's effectiveness in meeting the standards expected for his or her age by his or her cultural group in at least two of the following skills: communication, self-care, home living, social and interpersonal skills, use of community resources, self-direction, safety, functional academic skills, leisure, and work (American Psychiatric Association, 2000, p. 49). In contrast to the importance of intelligence, the *DSM–IV–TR* states that

> impairments in adaptive functioning, rather than a low IQ, are usually the presenting symptoms in individuals with Mental Retardation. *Adaptive functioning* refers to how effectively individuals cope with common life demands and how well they meet the standards of personal independence expected of someone in their particular age group, sociocultural background, and community setting. (American Psychiatric Association, 2000, p. 42)

The proposed fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders*. In its fifth edition of the *DSM,* the *DSM–5,* the American Psychiatric Association is proposing retention of some current features in the diagnosis of mental retardation and changes to other features. Mental retardation, as considered in the *DSM–5,* will retain the current features of intellectual deficits of 2 or more standard deviations below the mean and change the standards for both defining adaptive behavior and using adaptive behavior data. The proposed *DSM–5* emphasizes deficits in two of the three domains of adaptive functioning (i.e., conceptual, social, and practical) at 2 or more standard deviations below the mean. Under the proposed revisions, a qualifying score in either intelligence or adaptive behavior will generally translate into performance in the lowest 3% of a person's age or cultural group or standard scores of 70 or below. In addition, intelligence and adaptive behavior should be measured with an individualized, standardized, culturally appropriate, psychometrically sound measure (American Psychiatric Association, 2010).

Empirical support for the construct of adaptive behavior. As noted previously, the construct of adaptive behavior has undergone change seemingly every decade. The basis for this construct and its evolving nature appears to have rested more on theory than on a stable empirical foundation. Staying current with this ever-changing definition is difficult for those who develop and publish adaptive behavior scales as well as for those who use them.

Among measures of adaptive behavior, the theoretical structure of the Adaptive Behavior Assessment System—Second Edition (ABAS–II; Harrison & Oakland, 2003) best reflects the AAMR–AAIDD and *DSM–IV–TR* theory by incorporating a General Adaptive Composite (GAC), three domains, and 10 skill areas. Thus, data derived from the ABAS–II provide an empirical opportunity to test the structure of the current prevailing theory. Confirmatory factor analyses with the ABAS–II have verified that a one-factor model provides a good fit to the observed data from the standardization sample for both younger and older children (Aricak & Oakland, 2010; Harrison & Oakland, 2003; Oakland & Algina, 2011; Wei, Oakland, & Algina, 2008). Support for a one-factor model is consistent with research by McGrew and Bruininks (1989) as well as that by Harries, Guscia, Kirby, Nettelbeck, and Taplin (2005), who each found that most adaptive behavior instruments measure a general or global factor. Although the one-factor model provides the most parsimonious fit to the data, a three-factor

model also produces a close fit to the data and thus provides support for combining adaptive skills into the three proposed domains (Harrison & Oakland, 2003).

In support of a three-factor model of adaptive behavior, Greenspan and Driscoll (1997) suggested that adaptive behavior is better defined as a subset of personal competence, more specifically conceptual, practical, and social intelligences. Although similar to AAMR's 2002 definition of adaptive behavior, Greenspan (2006; Greenspan & Driscoll, 1997) proposed that adaptive behavior should be viewed as a tripartite model of intelligence used to identify a disorder on the basis of a demonstration of deficits in conceptual, practical, and social intelligences. By conceptualizing adaptive behavior in this manner, Greenspan (2006; Greenspan & Driscoll, 1997) proposed a focus on patterns of deficits in all three areas of intelligence rather than on the statistical rigidity of meeting a performance criterion of deficits greater than 2 standard deviations on a standardized measure of adaptive behavior.

Understanding adaptive behavior in light of the ICF. The World Health Organization's (2001) ICF provides a framework for describing adaptive behavior by distinguishing a person's activities (i.e., their potential engagement in functional life activities) and participation (i.e., their actual participation in life activities). Examples of activities include such behaviors as the ability to write, talk, calculate, and display other adaptive skills. Corresponding examples of participation include writing letters, talking with others, and determining the purchase price of food items when needed. Although a person may be able to perform an activity—that is, to display a skill—he or she may not display it reliably through participation when needed.

Measures of adaptive behavior that are consistent with the ICF should distinguish between the ability to perform an adaptive skill and the reliability and competence of its actual performance. For example, a person may be able to independently write letters, talk to others, and determine the purchase price of food items when needed yet require assistance in performing these skills in a reliable and competent manner. Thus, although the skills are within a person's repertoire, they are not displayed independently when needed.

The distinction between activities and performance is incorporated in how ABAS–II items are scored. An item rating of 1 is assigned to behaviors the person has the ability to perform but never or almost never does when needed or without being reminded. Higher ratings are assigned to behaviors the person has the ability to perform and does display sometimes (a rating of 2) or always or almost always (a rating of 3) when needed. For example, when rating a child's ability to button his or her clothing, the item is rated as 1 point if the child has the ability to button his or her clothing yet does not do so when needed, 2 points if the child displays this desired behavior sometimes when needed, or 3 points if the child displays this desired behavior always or almost always when needed. Using ICF terminology, a rating of 1 signifies activities, whereas ratings of 2 and 3 signify performance.

Legal Standards

Legal standards that emerge through legislation and related administrative decisions as well as case law help define required uses of adaptive behavior. Two legal standards, one that impacts the use of adaptive behavior measures with students and another that impacts adults sentenced to death, deserve attention.

For students. The Individuals With Disabilities Education Improvement Act of 2004 (IDEIA) governs the provision of early intervention, special education, and related services by state and local educational agencies for children ages 2 through 21 (United States Code Service, 2007).

Part C of IDEIA addresses assistance for infants and toddlers with disabilities by authorizing states to develop and maintain early childhood intervention programs (Appling & Jones, 2005). Consistent with the focus of Part C, the use of adaptive behavior measures is especially pertinent for young children. For example, eligibility for services may be based on a diagnosis of developmental delay, which requires evidence of diminished adaptive behavior. The assessment of adaptive behavior provides data that can be helpful in establishing the presence and

degree of impairment, which is essential for determining eligibility for services as well as identifying interventions and monitoring treatment effectiveness.

Part B of IDEIA addresses the provision of assistance to students, ages 3 through 21, who display special needs (U.S. Department of Education, 2006). School-based child study teams determine whether special needs students who meet eligibility criteria for one or more of the 13 disabilities authorized under IDEIA (e.g., children who display a specific learning disability, autism, or mental retardation or intellectual disability) are eligible for special education and related services (Apling & Jones, 2005).

Information gathered from comprehensive measures of adaptive behavior provides data that can assist in diagnosis, intervention, and progress monitoring. This information is directly relevant when a child study team is attempting to determine the student's diagnosis. For example, evidence of a deficit in adaptive behavior is needed for a diagnosis of a developmental or intellectual disability. Additionally, information on a student's adaptive behavior may also be helpful when attempting to understand the effect of disabilities on children's functional daily living skills, to inform educational programming efforts, to help monitor interventions, and to provide a baseline needed to determine progress in light of data from subsequent reevaluations (Harman, Smith-Bonahue, & Oakland, 2010).

The contributing characteristics of adaptive behavior scales are typified in a study by school district personnel that examined the adaptive behavior profiles of students referred for special education (Ditterline, Banner, Oakland, & Becton, 2008). Some referred students qualified for services because of specific learning disabilities, emotional disturbance, specific learning disabilities in combination with emotional disturbance, and autism. However, all referred students displayed deficits in adaptive behavior and skills. Students with more severe disability diagnoses (e.g., autism) or multiple diagnoses (e.g., coexisting specific learning disability and emotional disturbance) commonly displayed more severe adaptive behavior and skill deficits. Thus, students who are referred for special education services are likely to display deficits in adaptive behaviors, thus warranting the assessment of such behaviors even when this information is not needed for diagnostic purposes.

Under IDEIA, local educational agencies are required to use multiple assessment methods and sources of information to document disabilities. These proposed data have three primary purposes: to assist in determining whether children have disabilities, to inform and guide the content of an educational plan, and to provide baseline data for evaluating the efficacy of the intervention (Council for Exceptional Children, 2004). Local education agencies are encouraged under IDEIA legislation to assess functional skills such as adaptive behavior to help determine meaningful interventions that can have a direct and functional influence on important practical life skills.

Atkins v. Virginia. Information on adaptive behavior can also be important when determining whether people sentenced to death for a capital crime live or are executed. In 2002, the U.S. Supreme Court, in *Atkins v. Virginia,* considered whether people with mental retardation could be executed for capital crimes (Olley & Cox, 2008). The court's ruling highlighted three essential conditions. First, it prohibited the execution of people with mental retardation; it affirmed the use of the *DSM–IV–TR* definition for the diagnosis of mental retardation (see previous mention); and it left to states the responsibility for determining the procedures for establishing mental retardation in capital cases. The use of the *DSM–IV–TR* definition was uniformly applauded nationally by psychologists because it provided a professionally developed and widely used standard for diagnosing mental retardation. However, the provision that allowed states to determine the procedures for establishing mental retardation has led to different standards from state to state. For example, criteria for determining the diagnostic cut for identifying significant deficits or impairments in adaptive functioning differ considerably between states. Some states use a standard score of 70, plus or minus the standard error of measurement (e.g., typically about 3 points), other states use a cut score equal to 70 and lower, and others use a standard score less than 70 (i.e., ≤ 69).

As noted previously, the *DSM–IV–TR* has defined mental retardation as the display of significantly

subaverage intelligence (i.e., approximately 70 or lower) along with concurrent deficits or impairments in present adaptive functioning, with the onset being before age 18. Most psychologists, psychiatrists, and other mental health specialists have interpreted "along with concurrent deficits or impairments in present adaptive functioning" to refer to deficits in adaptive functioning before age 18. However, courts differ in how they interpret the term *present adaptive functioning.* Some courts have interpreted the phrase consistent with the beliefs of most mental health specialists. Other courts have interpreted the term literally to mean either at the time of the capital offense (e.g., which could be beyond age 18) or currently—when the prisoner is incarcerated and thus in a setting in which adaptive behaviors cannot be displayed independently. Thus, the U.S. Supreme Court decision in *Atkins* was initially thought to provide considerable consistency nationally in decisions as to whether death row prisoners could be executed through the court's application of the *DSM–IV–TR* definition of mental retardation. However, states' ability to determine the procedures for establishing mental retardation has resulted in different standards among the states, including the age at which information on adaptive behavior becomes relevant to determining mental retardation.

AGE DIFFERENCES IN CURRENT ADAPTIVE BEHAVIOR THEORY

Significant age differences have been found in adaptive behavior and skills throughout the life span of individuals. Developmental data from measures of adaptive behavior across the age span show steep growth early and often, topping out and plateauing between the ages of 20 and the 60s, followed by a decline in adaptive skills in the 70s. The decline in the 70s primarily depends on individual characteristics as well as psychopathology, autism, Down syndrome, intellectual disability, or medical diagnoses such as dementia, Alzheimer's disease, and Parkinson's disease, among many others. Individuals with these conditions typically exhibit a rapid and profound decline in adaptive skills, including their independent functioning, communication, motor skills, and various personal responsibilities (De Ridder, Schreurs, & Bensing, 1998; Matson, Rivet, Fodstad, Dempsey, & Boisjoli, 2009; Prasher, Chung, & Haque, 1998; Zigman, Schupf, Urv, Zigman, & Silverman, 2002). In the general population, older adults in their late 80s, compared with their younger counterparts in their late 70s and early 80s, generally display a greater need for assistance with motor, leisure, and self-care adaptive behaviors such as walking, shopping, bathing, and toileting (Rubel, Reinsch, Tobis, & Hurrell, 1995).

Adults with intellectual disability often use community resources such as work training programs and housing assistance. The goal of such programs is to help participants develop the skills needed for independent functioning. Various scholars have examined the nature and efficacy of these programs (Emerson et al., 2000; Stancliffe, Hayden, Larson, & Lakin, 2002). For example, researchers conducting a study of 272 adults with intellectual disability found that general adaptive behavior accounted for 40% and 43% of variance in participants' work and residential independence, respectively (Woolf, Woolf, & Oakland, 2010). A 25-year follow-up study of 91 people with severe intellectual disability found their levels of social impairment best predicted the later presence of negative outcomes in their independent functioning, residential placement, employment, and quality of life (Beadle-Brown, Murphy, & Wing, 2005).

Adaptive Behavior: Three Theories of Early Development

The rate of children's growth and development is more rapid from birth through age 5 than at any other time in their lives. Knowledge and understanding of the developmental process during this critical period are enhanced through theories and research on infants (from birth–age 1), toddlers (ages 1–3), and young children (ages 3–5). This body of scholarship helps professionals and parents chart normal development, nurture its development, and identify when a child's growth deviates from normal standards.

This section describes the early development of adaptive skills exhibited by children from birth through age 5. A theoretical framework for understanding early

child growth and development, including the development of adaptive skills, is provided first, followed by a description of the development of the nine adaptive skill areas developed in early childhood: communication, self-care, home living, social skills, community use, self-direction, health and safety, functional academics, and leisure.

Adaptive skills can be considered the functional expression of young children's ability to meet their daily needs and manage the demands of their environment. Conceptually, adaptive skills can be viewed as the interaction among developmental ability, family and cultural expectations, and environmental opportunities (Harman et al., 2010). A conceptual understanding of the development of adaptive skills in infants and young children from birth through age 5 can be aided by knowledge of several contemporary theories of child development. The three theories discussed next provide an understanding of the development of adaptive skills in young children: Piaget's (1951) theory of cognitive development, Vygotsky's (1982) zone of proximal development theory, and Gesell's (1952) maturation theory.

Piaget's theory of cognitive development. Piaget's (1951) theory of cognitive development focuses on child development beginning in infancy. He proposed that children develop *schemas* (i.e., an internal representation of the world) that they later adapt to the environment through the use of *assimilation* (i.e., incorporating a new object or concept into an existing schema) and *accommodation* (i.e., altering an existent schema on the basis of a new object or concept). Piaget's theory proposes that infants and young children progress through stages of development, including the sensorimotor stage (birth through age 2) and the preoperational stage (ages 2–6; Piaget, 1951).

During the sensorimotor stage of development, infants begin to adapt to their environment by coordinating their sensory perceptions and simple motor behaviors (e.g., reaching for an object, nursing at their mother's breast). Piaget (1952) identified six sensorimotor substages that further delineate the hierarchical nature of child development: *reflex schemas* (i.e., involuntary grasping and sucking), *primary circular reactions* (i.e., repetition of actions that infants find pleasurable), *secondary circular reactions* (i.e., awareness of the relationship between their own actions and the environment), coordination of secondary circular reactions (i.e., combining schemas to achieve a desired effect), *tertiary circular reactions* (i.e., experimenting to determine consequences of actions), and the beginning of *symbolic representation* (i.e., associating words and images to objects). Similar to the development of adaptive skills, an infant's gradual sensory and motor development during these stages builds on previously acquired skills and develops from simple perceptual orientations to increasingly conceptual complexity (Piaget, 1952).

According to Piaget, infants who have acquired sensorimotor skills begin to transition their cognitions from actions in their environment to a more internalized state (i.e., from percepts to concepts; Piaget & Inhelder, 1969). This internalized process is not fully operational initially and is thus referred to as *preoperational* (i.e., the preoperational stage). During this stage, from about age 2 until age 6, toddlers and young children begin to cognitively represent objects through the use of symbols, words, and gestures. Given their new ability to understand and express themselves, toddlers are typically egocentric until approximately age 4, thus viewing themselves as the center of the world and assuming that others share a similar viewpoint. As toddlers mature, they continue to take in information about the world and assimilate it to fit their ideas while using little to no accommodation (Cole & Cole, 1996).

Piaget's work in childhood development provides a framework for understanding the developmental sequence through which infants and later toddlers progress as they acquire adaptive behaviors. He, as did Vygotsky, described qualitative changes that occur initially through infants' actively acquiring and retaining new objects and concepts that become integrated into schemas. Through the use of assimilation and accommodation, infants alter existing schemas, thus leading to the emergence of symbolic development, including language, at about age 2 (Piaget, 1963).

For example, Piaget's (1971) sensorimotor stages enable one to see how young infants develop and

display motor skills that later become integrated into broader cognitive abilities. For example, infants first clench their fists, then move their hands to their mouth, later grasp and then manipulate small objects, which later leads to their ability to use a pencil to write and thus express their thoughts and feelings.

Piaget's theory provides the underlying explanation of how young children use their developing sense of internalization and egocentric view of the world to develop adaptive behaviors such as using and understanding pronouns such as *you* and *me,* recognizing the ownership of belongings, and beginning to develop patience with others, although not necessarily understanding their reasoning (Piaget, 1924).

Vygotsky's zone of proximal development theory. Vygotsky characterized human development as the ability to acquire mental tools. These tools are analogous to industrial tools in that they are developed through social activity rather than organically or biologically. Vygotsky viewed biological components of early childhood development as being shaped by social and historical development (Hedegaard, 2005). He proposed that learning is conducted through a social mechanism that he labeled the *zone of proximal development*.

The zone of proximal development represents a range of behaviors within reach of engagement, yet just beyond a child's current abilities. Thus, the zone of proximal development is represented by the difference between what a child can do independently and what he or she can do with the support of a caretaker, behaviors that later can be displayed independently when needed. These next-to-develop behaviors are acquired and sustained more proficiently when guided by caretakers than when acquired on one's own (Vygotsky, 1934/1987).

An understanding of how Vygotsky (1978) characterized child development through social mechanisms, including the zone of proximal development, assists one in understanding the development of adaptive behavior in young children. In short, infants typically learn best by engaging socially with their peers as well as caregivers as they actively work to acquire skills that further extend their current development. The development of play and vocal language skills exemplifies this process.

As infants grow into young children, they begin to develop adaptive behaviors related to social relationships through play with other children. Infants typically engage in play activities by themselves. After some time, toddlers observe other children playing yet will not engage in play with them (e.g., an initial form of parallel play). As toddlers continue to develop, they engage in associate play with others that includes aspects of both solitary and occasional cooperative play. Finally, young children develop to the point that they will engage in cooperative play with other children (DeVries, 2008). One can observe infants and young children developing skills related to play through the use of the zone of proximal development, which provides a foundational understanding of one way in which young children learn adaptive behaviors.

The development of vocal language also exemplifies the process by which young children learn adaptive behaviors through the zone of proximal development (Vygotsky, 1934/1987). Infants' receptive language develops first, followed by expressive language. Their receptive language development requires an environment in which others who interact with the infant use language socially. Infants rely on and increasingly incorporate language patterns expressed by those around them. Infants deprived of a language environment are delayed in their language development. Expressive language skills emerge from receptive language first through the use of nonverbal (e.g., hand and facial) expressions and later by imitation of peer and caregiver sounds. Infants' first words are inarticulate, and they gradually display more refined articulation together with terminology (e.g., words) and concepts needed to engage socially. Throughout this process, infants develop adaptive language behaviors with the assistance of peers and a caregiver who serve as models and who reinforce and provide other forms of support that help infants to reach their next stage of development (i.e., their zone of proximal development) to further extend and promote their language development.

Gesell's maturation theory of child development. Gesell's (1952) maturation theory of early childhood

development proposes that infants develop adaptive behaviors in an orderly and sequenced fashion. He viewed growth as analogous to *development,* which is defined as an organizational process that is both unitary and integrative. Gesell considered growth as unitary—that is, the body and mind do not differ. Gesell also viewed growth as integrative—that is, the development of the body and mind are expressed through changes of form and patterning. Current skills become more refined and integrated during later stages of development (Gesell, 2007).

Gesell proposed that virtually all behavior has a motor origin combined with other productive aspects (e.g., vision, speech, mental imagery, conceptual thought). He even proposed that emotion has a motor origin. Gesell proposed several foundational concepts that form the basis for his theory. These concepts include maturation, general to specific, cephalo–caudal growth, proximal–distal growth, readiness, and regression (Daly, 2004).

Maturation refers to the rate at which infants and young children develop and exhibit new behaviors. Gesell believed that most children exhibit and develop behaviors and skills at roughly the same ages. The belief that behavior evolves from general to specific reflects how infants' behaviors at first are unorganized and later become more deliberate as their environment becomes more demanding and their skills thus require greater self-control. Gesell (1930) proposed that the rate at which skills develop is governed by two conditions: heredity and time. Both are known to affect young children's rate of development. The rate at which children acquire skills is governed by the family's history as transmitted through genetics. Moreover, the steps through which infants pass when acquiring skills are set and common to all. Thus, little can be done to speed up the normal developmental process given its reliance on both genetics and time. These two qualities account for the largest amount of variance associated with maturational development.

The concept of cephalo–caudal development recognizes that physical control begins with the head and face, leading to the trunk, and finally extending to the extremities. For example, infants begin life in a prone position, then crawl and later walk, thus culminating in control of their extremities. The concept of proximal–distal growth refers to the development of skills from the body's midline to its extremities (e.g., infants develop the use of arms before developing use of fingers). Cephalo–caudal and proximal–distal growth patterns occur in all children, albeit at somewhat different rates.

Gesell characterized readiness as the time period when infants can be expected to exhibit a specific behavior, such as crawling or walking (Gesell, Halverson, & Amatruda, 1940). Infants develop behaviors sequentially and can display them only if they are ready maturationally—that is, when they have mastered prerequisite skills. He proposed that readiness is progressive and unalterable, unless interrupted or halted by disease or trauma. Although adaptive behaviors are acquired at different times for different children, their behaviors emerge in a predictable fashion, and only when the young child is neurologically ready to acquire and retain them.

The concept of regression refers to when a young child has displayed a given behavior and later seemingly loses the ability to perform the behavior, having regressed to an earlier step (e.g., an infant who has been toilet trained reverts to self-soiling). Gesell described behavioral regression as a process that allows young children to gain control, stability, and integrity before continuing to master new skills (Gesell et al., 1940). As infants develop new skills and progress through transition and regression periods, their behavior destabilizes as they temporarily exhibit previously developed skills, which is generally then followed by the exhibition of a new skill (Sadurní, Perez Burriel, & Plooij, 2010). Researchers have identified several distinct periods in infancy during which infants typically exhibit periods of regression before developing new skills (Van De Rijt-Plooij & Plooij, 1992, 1993). Among children who are developing normally and have not been exposed to trauma, these periods of regression should typically be viewed as a positive sign of development because they precede new developmental milestones.

Gesell's (1952) maturation theory of early childhood development provides a framework for understanding how and when adaptive behaviors typically develop in infants and young children. His theory is particularly relevant to understanding motor skill

development. The concepts of cephalo–caudal and proximal–distal growth patterns provide an understanding of how infants gain control of their motor skills and when they should develop. Gesell's theory also explains the continuous process through which infants exhibit new adaptive behaviors only after they have reached a point of maturation and readiness for that skill. The rate at which infants and toddlers generally develop adaptive behaviors can be determined by normative data collected from the general population. In addition, Gesell's theory emphasizes regression and its importance to both caregivers and professionals. Although regression may initially be a cause of concern, regression is typically exhibited temporarily and before the development of new adaptive skills. Thus, regression can be considered a typical part of the developmental process.

DEVELOPMENT OF 10 ADAPTIVE SKILLS BETWEEN BIRTH AND AGE 5

An understanding of the theoretical basis of how infants and young children develop adaptive behaviors enhances one's understanding of children's adaptive behavior between birth and age 5. This understanding is further accomplished by identifying the development of adaptive behaviors using data from three nationally standardized and commonly used measures that assess development of young children: the ABAS–II (Harrison & Oakland, 2003), the Bayley Scales of Infant and Toddler Development—Third Edition (Bayley, 2006), and the Scales of Independent Behavior—Revised (SIB–R; Bruininks, Woodcock, Weatherman, & Hill, 1996). The items from these scales are arranged sequentially and thus reflect a pattern of normal development across the age span. Data from these measures enable psychologists to determine when behaviors are typically displayed by a cross-section of the U.S. population of young children. Information from the Vineland Adaptive Behavior Scales—Second Edition (VABS–II; Sparrow, Cicchetti, & Balla, 2005), another prominent measure used in the assessment of adaptive behavior, was also considered. However, its data were not used in this analysis of adaptive skill development because of the difficulty in determining the relationship between v-scale scores (VABS–II subdomain scores) and age-related adaptive skill development.

Item data from these instruments were reviewed in light of the model of adaptive behavior promulgated by the *DSM–IV–TR,* one that highlights the following 10 skill areas: communication, self-care, home living, social–interpersonal skills, use of community resources, self-direction, safety, functional academic skills, leisure, and work. The following review focuses on adaptive skill development in young children. Thus, the skill of motor development was substituted for work.

Communication

Adaptive communication can be divided into two types of behavior: receptive and expressive communication. *Receptive communication* refers to the ability to both recognize and understand verbal stimuli. *Expressive communication* refers to the way in which one communicates by gesturing, speaking, writing, or signing (e.g., interacting with others through verbal and nonverbal actions). Developmentally, receptive communication skills are acquired before and provide the foundation for the acquisition of expressive communication skills (i.e., children generally hear and understand sound before producing meaningful vocalizations). During this period (i.e., birth through age 5), communication skills range in difficulty from acknowledging a caregiver to using complex sentences to express desires and opinions.

From birth to the 1st year of life. Infants' orientation responses to caregivers' voices often constitute some of their first signs of communication. The emergence of communication skills may first be seen in accepting attention and displaying calming behaviors in response to the caregiver's voice. Receptive language skills continue to develop as infants begin to discriminate between sounds that then become recognizable auditions (e.g., recognizing their own name). As an infant's skills continue to develop, infants will disengage from an activity if verbally prompted by a caregiver. Before their 1st birthday, infants' receptive communication skills enable them to attend to and respond differently to two or more words from a caregiver. Infants also modify their behaviors in response to spoken

requests from caregivers (e.g., "No-no"). Infants typically display an attention span of as long as 10 seconds when interacting with others.

Expressive communication skills also begin to develop during the 1st half-year of life. Infants begin vocalizations by producing undifferentiated sounds that later develop into social vocalizing of differentially louder or softer sound volumes, combined with appropriate social smiling. These communication skills enable infants to capture the attention of others and serve to reinforce caretakers' behaviors. Infants typically display at least two consonant and vowel sounds before age 6 months. Before their 1st year, infants verbally communicate distress through vocal expressions of fluctuating volume. Infants utter one-word expressions such as "mama" or "dada." Infants also communicate using nonverbal methods (e.g., signaling yes or no with facial gestures in response to short yes-or-no questions).

From the 1st to the 2nd year of life. As infants grow into toddlers at age 1, receptive communication skills enable toddlers to follow simple yes-or-no commands and to listen intently to caregivers speaking for up to 1 minute. Toddlers are able to follow simple one-part directions when prompted by a caregiver (e.g., "Pick up the toy"). Before age 2, toddlers follow more complex language features, including such concepts as understanding the meaning of *over* or *under* (e.g., "Climb over the chair"). Toddlers also typically identify and point to 20 or more familiar objects (e.g., clothing items or body parts).

Expressive communication skills enable toddlers to draw a caregiver's attention to an object, use two or more words appropriately, and combine words with a matching gesture (e.g., say "ball" and simultaneously point to the ball). A toddler's ability to communicate is also displayed by his or her ability to repeat three or more discrete words when prompted by a caregiver (e.g., *dog, cat,* and *house*). Before age 2, a toddler's expressive communication skills are developing rapidly. For example, toddlers use at least eight words correctly and begin using short multiple-word sentences. Toddlers also begin singing short songs, understanding and using transitional word tenses (e.g., from present to past tense), and asking questions.

From the 2nd to the 3rd year of life. Communication skills continue to develop rapidly as toddlers continue to grow at age 2. Receptively, toddlers are able to follow increasingly complicated verbal instructions (e.g., "Put your toys away") and begin to understand pronouns (e.g., *me, you*), whole–part relationships (e.g., door of the house or tail of the cat), and preposition series (e.g., to follow directions). Communication skills also enable toddlers to comprehend and use the possessive (e.g., *mine, yours*), nouns (e.g., *dad, mom*), and gerunds (e.g., *walking, running, jumping*).

As expressive communication skills continue to develop, toddlers speak in three- to six-word sentences and vocally express their favorite activities. Toddlers use plural words (e.g., *toys*) and develop new word combinations (e.g., noun and verb; noun and adjective; or noun, verb, and adjective). Toddlers also request others to engage in activities with them.

From the 3rd to the 4th year of life. As toddlers grow into young children at age 3, receptive language skills enable them to understand and follow multipart directions in the correct order (e.g., "Put your toys away and then come sit at the table"). Young children also focus on discussions of one topic for several minutes (e.g., to discuss what they did that day). Young children also read their name from a list of several names (e.g., their name above a coat hook).

Expressively, young children answer specific questions (e.g., what and where) and end conversations appropriately. Young children communicate using more complex sentences to express their desires and opinions, using words such as *because* (e.g., "I like the cartoon because it's funny").

From the 4th to the 5th year of life. Although child development occurs in a uniform fashion, the amount of growth during each interval is not always equal. Young children at this age continue to display previously developed communication skills that are often displayed more fluently and in a more integrated fashion.

Community Use

Community use skills include those needed to function and act appropriately within a community.

Community use skills do not develop during early infancy and instead begin to emerge at 18 months of age. A toddler's first community is his or her home, and it then expands to encompass their yard, neighborhood, and finally the larger community. Community use skills may also be reflected in a toddler's use of time, money, punctuality, work skills, and home–community orientation. During the following 3.5-year period, these skills range from locating objects in the home that are kept in the same place to identifying the location of service rooms such as public restrooms.

From age 18 months to the 2nd year of life. During this period, toddlers ages 18 months to 2 years are able to differentiate the inside and outside of their home. They also inform parents or caregivers when someone arrives at their home. With supervision, toddlers are able to walk safely on the sidewalk rather than in the street and identify their own house in a neighborhood. Toddlers begin to show respect for others' belongings (e.g., throw trash in the trash can). Toddlers also knock on a door or ring a doorbell before entering a home and behave more appropriately in public places (e.g., a house of religion or movie theater).

From the 2nd to the 3rd year of life. At age 2, toddlers request caregivers to take them to their favorite places (e.g., to go to a park or other community location). Inside their home, toddlers are able to go to a designated room when requested by a caregiver. Toddlers identify and recognize community buildings within categories (e.g., hospitals, fire stations, and police stations). Toddlers associate the purchase of items from stores as having financial value as they begin to develop an understanding of currency.

From the 3rd to the 4th year of life. As toddlers grow into young children at age 3, they identify specific locations to which their family needs to go for desired items (e.g., where to buy groceries or clothing items). Young children also identify the title and purpose of common jobs in the community (e.g., policemen, firefighters, doctors). Young children display awareness of their need to look both ways before crossing a street. Currency skills continue to develop as young children further develop their awareness of the value of money by retaining theirs in a special location.

From the 4th to the 5th year of life. At age 4, young children request to be taken to community locations (e.g., a specific restaurant, movie theater, library). Young children also understand that people in the community have work roles and know from whom to request items or services if needed. Before age 5, young children identify the location of service rooms such as public restrooms and identify coin currency.

Functional Preacademic Skills

Functional preacademic skills include those that help form the foundation for later reading, writing, and mathematics—skills needed for independent functioning at school and elsewhere (e.g., letter recognition, counting numbers, drawing simple shapes). Because of the complex nature of functional preacademic skills, their earliest development often starts at age 18 months. During the following 3.5-year period, functional preacademic skills range in difficulty from identifying pictures in a book to developing the writing and comprehension skills necessary to print numbers 1 through 10.

From 18 months to the 2nd year of life. During this period, toddlers are able to identify pictures within text (e.g., a horse, a dog). Toddlers also hold a writing utensil on a sheet of paper. Before age 2, toddlers begin to understand the counting system and state their age in numbers and count three or more objects (e.g., blocks, books, pencils). Toddlers also begin to replicate objects in their drawings, albeit crudely (e.g., drawing objects in their environment).

From the 2nd to the 3rd year of life. At age 2, toddlers are able to identify six or more colors. They also memorize and recite short songs or nursery rhymes (e.g., "Mary Had a Little Lamb"). Toddlers also recognize and differentiate between numbers as well as shapes. Their reading skills begin to emerge as they read their first name. Toddlers' counting skills also continue to develop (e.g., they count to at least five).

From the 3rd to the 4th year of life. As toddlers grow into young children at age 3 they are able to count to 10 without the aid of objects or counting their

fingers. Their ability to draw nonabstract pictures increases as they attempt to draw human figures that have identifiable anatomical parts (e.g., a body with a head, both arms, legs). Young children later use rote memory to count to at least 20. Young children are also able to identify most letters of the alphabet.

From the 4th to the 5th year of life. At age 4, young children begin to develop writing skills, including the ability to write several letters of their name. Reading skills continue to develop as young children identify common signs and comprehend their meaning (e.g., an exit or stop sign). Young children have also memorized the days of the week in order and developed the comprehension and writing skills necessary to print numbers 1 through 10.

Home Living

Home living skills include those needed for basic care of a home or living setting. These skills include cleaning and arranging the rooms in one's home, helping caregivers with household tasks, and taking care of one's personal possessions. Home living skills typically emerge during the 1st year of life. During the first 5 years, home living skills range in difficulty from removing food from a bag or box to developing a better understanding of household rules and acting accordingly.

From birth to the 1st year of life. During this period, home living skills include infants being able to remove food from a bag or box, manipulate electronic entertainment devices (e.g., turn the television on or off), and express concern to caregivers if they break something by accident. Infants also begin to comprehend where their personal belongings are stored.

From the 1st to the 2nd year of life. As infants grow into toddlers at age 1, they are able to turn light switches on and off (they may use a chair or stool), assist a caregiver in cleaning up messes, and retrieve their own snack from a cabinet. Toddlers begin to offer assistance to adults with such tasks as cooking or cleaning. They also develop a better sense of household rules and how to dispose of trash properly.

From the 2nd to the 3rd year of life. At age 2, toddlers are able to clean up after themselves (e.g., put their cup and plate in the sink after using them) and identify where their clothing is kept. Toddlers also become aware of and engage in putting items in places where they belong (e.g., putting dirty clothing items in a basket).

From the 3rd to the 4th year of life. As toddlers grow into young children at age 3, they know where many household items belong, and both obtain and replace them. Young children are also able to store their toys in an organized fashion and assist a caregiver without being requested to help.

From the 4th to the 5th year of life. At age 4, young children develop a better understanding of household rules and act accordingly (e.g., removing dirty or wet clothing items before entering the house). During the last half of this period, young children's previous home living skills are displayed more fluently, in an integrated fashion, and with less prompting or assistance.

Health and Safety

Health and safety skills are those needed to maintain one's health and to take appropriate actions to prevent and, if needed, respond to one's injury or illness. These skills include following safety rules, taking medicine, and using caution in daily practices. During this period, health and safety skills range in difficulty from using vocal expressions to indicate when infants are feeling ill or desire to be fed to showing a deeper understanding about emergency situations and understanding procedures related to their safety.

From birth to the 1st year of life. During this period, infants use vocal expressions (e.g., crying) to indicate when they are feeling ill or desire to be fed. They also swallow medicines when needed. Infants become more able to avoid bumping into objects as they crawl, to show another person their minor injuries, and to respond to an adult's vocal commands when they are nearing dangerous situations.

From the 1st to the 2nd year of life. As infants grow into toddlers at age 1, they are aware that patience is needed to allow basic medical procedures to be conducted. Additionally, they are aware of both hot and cold sensations and react accordingly.

Toddlers acknowledge that some objects or situations are dangerous. They inform others verbally when they feel ill. Toddlers also begin to develop a sense of hygiene or cleanliness.

From the 2nd to the 3rd year of life. At age 2, toddlers are aware of the location of their caregivers in proximity to themselves. They also identify differences in temperature and may request alternative clothing to accommodate temperature changes. Toddlers carry an item that may break while displaying the sustainable balance and coordination needed to not drop it.

From the 3rd to the 4th year of life. As toddlers grow into young children at age 3, they display additional awareness of dangers (e.g., of animals or when using playground equipment). Young children ask caregivers about potential dangers and remember to put on seat belts before riding in an automobile. In the last half of this period, young children generally consolidate health and safety skills rather than add many new skills.

From the 4th to 5th year of life. At age 4, young children show a deeper understanding of emergency situations and are able to understand procedures related to their safety. In the last half of this period, young children's previously developed health and safety skills are displayed more fluently, in an integrated fashion, and with less prompting from adults.

Leisure

Leisure skills include those needed to engage in and plan recreational activities. These skills include playing with caregivers and peers, participating in activities in and outside the home, and following rules when playing with others. During this period, leisure skills range in difficulty from picking up and being entertained by small toys for as long as 1 minute to enjoying simple board games with peers and caregivers.

From birth to the 1st year of life. During this period, infants begin to pick up and become entertained by small toys for as long as 1 minute. Additionally, they look at picture books with a caregiver and observe others interacting with objects (e.g., toys or games). Infants select and play games with adults or caregivers (e.g., peek-a-boo or rolling a ball). An infant's attention span while playing games may be as long as 5 minutes.

From the 1st to the 2nd year of life. As infants grow into toddlers at age 1, they participate in parallel play (e.g., playing alongside other children with minimal interaction). Toddlers also enjoy swinging and sliding at play areas. Toddlers have a favorite book they enjoy having read to them. Toddlers may also enjoy attending activities at locations other than their home and engaging in games or play without supervision (e.g., playing on the playground or visiting a relative).

From the 2nd to the 3rd year of life. At age 2, toddlers may watch a favorite television program or play a specific video game on a routine basis. They are also able to wait turns when playing a game with peers or adults. Toddlers express interest in saving small objects they find (e.g., stones, buttons, or other knickknacks).

From the 3rd to 5th year of life. As toddlers grow into young children at age 3, they invite others to attend an event at their home or elsewhere. At age 4, they participate in simple board games with peers and caregivers. During this period, many of young children's previously developed leisure skills are displayed more fluently, in an integrated fashion, and with more independence.

Self-Care

Self-care skills include those needed for personal care. They include proficiency in such areas as eating, toileting, and dressing. During this period, self-care skills range in difficulty from nursing, drinking, and eating to using the bathroom in private.

From birth to the 1st year of life. During this period, infants begin to develop various self-care skills including nursing, drinking, and eating willingly with little or no encouragement. If prompted by a caregiver, infants open their mouths or drink from a sippy cup with assistance. Infants also swallow soft foods (e.g., mashed carrots). Although infants initially wake every 3 or 4 hours to feed, they often sleep through most of night with few interruptions toward the end of this period. Infants continue to develop and improve existing self-care skills by requesting food when wanted,

drinking from a sippy cup without assistance, and feeding themselves dry food (e.g., cereal, crackers). Hygiene skills begin to develop as infants wash their hands with soap with assistance. Infants also generally enjoy being bathed. Self-dressing skills progress from first providing assistance when being dressed, to undressing themselves, to finally dressing themselves, albeit not during the 1st year of life. During the 1st year of life, infants develop the skills needed to assist caregivers when they are getting dressed (e.g., lifting arms or legs when needed).

From the 1st to the 2nd year of life. As infants grow into toddlers at age 1, they display many self-care skills, including washing their hands when needed and novice attempts to brush their teeth. Toileting skills are also developed, including the expression of displeasure when they are wet or soiled. Toddlers may be capable of sitting on a child-size toilet seat with little to no assistance. Dressing skills continue to be developed as toddlers remove their socks, pants, and shirt when needed. Additional toileting skill development includes being able to identify when they need to use the bathroom and not wetting or soiling themselves for several hours.

From the 2nd to the 3rd year of life. At age 2, toddlers eat food using a fork and spoon designed for their age. Dressing skills progress to being able to dress themselves, at first with help when buttoning and then buttoning by themselves. Concerning toileting, toddlers increasingly indicate their need to use the bathroom when questioned by a caregiver and then use the bathroom at regular times. Toddlers perform hygienic routines such as washing their face and body with little or no assistance. The numbers of toileting accidents are likely to be less than 1 per month as their skills continue to develop.

From the 3rd to the 4th year of life. As toddlers grow into young children at age 3, they become more proficient at performing previously acquired skills. For example, young children brush their teeth and rinse their mouth independently. They put their shoes on except for tying them. Their toileting skills include wiping themselves with little to no assistance after a bowel movement.

From the 4th to the 5th year of life. At age 4, young children develop additional independence in self-care skills as seen in their drying themselves after bathing, then dressing themselves by selecting and properly wearing clothes (e.g., not inside out). When eating, young children generally take amounts of food that are proportional to their appetite. Concerning toileting skills, young children use the bathroom in private.

Self-Direction

Self-direction skills include those needed for independence, responsibility, and self-control. These skills include making choices about one's food and clothing, starting and completing tasks, and following a daily routine. During this period, self-direction skills range in difficulty from engaging in self-soothing activities for 1 or more minutes to regulating their emotional responses when things do not go their way.

From birth to the 1st year of life. During this period, infants develop the skills necessary to self-soothe for 1 or more minutes before requiring attention. Infants adjust their emotional responses when picked up or spoken to by an adult. They entertain themselves for as long as 5 minutes, indicate interest in an object or person nonverbally (e.g., pointing), and express a choice when given a decision between two objects. They also begin to explore new areas (e.g., an unfamiliar room or other new situation) with or without encouragement from a caregiver.

From the 1st to the 2nd year of life. As infants grow into toddlers at age 1, they obey a caregiver's request to follow simple household rules. They also begin attempting to mimic caregiver-like actions (e.g., novice attempts to dress or feed themselves). Toddlers generally resist physically harming other children when upset, follow authoritative directions more readily, and persist at difficult tasks for longer periods of time.

From the 2nd to the 3rd year of life. At age 2, toddlers begin to request permission from a caregiver when needed before engaging in activities (e.g., when they would like to go outside). They are more inclined to work on a task by themselves and to request help only when needed. Toddlers are

better able to control their temper when a caregiver takes an object from them. Toddlers also concentrate on one activity for as long as 15 minutes.

From the 3rd to the 4th year of life. As toddlers grow into young children at age 3, they stop a fun event (e.g., play time) with little or no distress when told to by a caregiver. During the last half of this period, young children's self-direction skills generally consolidate rather than add new skills.

From the 4th to the 5th year of life. At age 4, young children are better able to control their emotional regulation when things may not go their way. Toward the end of age 4, young children's self-direction skills are displayed more fluently, in an integrated fashion, and with less prompting by caregivers.

Social

Social skills include those needed to interact socially and get along with other people. These skills include expressing affection, developing friendships, displaying and recognizing emotions, and providing assistance to others. During this period, social skills range in difficulty from expressing emotions vocally to caregivers to making verbal amends if one injures a peer or caregiver.

From birth to the 1st year of life. During this period, infants' social skills include the ability to express their emotions vocally to caregivers. For example, they may begin to express happiness when a parent or caregiver returns to them or holds them. Infants also begin to show physical movements intended to engage caregivers socially, such as lifting their arms to express the desire to be picked up. Infants develop the ability to recognize and respond differently to unfamiliar people. Their capacity to show affection to parents or caregivers also increases. They begin to imitate the actions of others (e.g., pretending to drive a vehicle) and are more willing to consciously share items with a caregiver.

From the 1st to the 2nd year of life. As infants grow into toddlers at age 1, they begin to follow social norms such acknowledging another child and saying "please" or "thank you" when prompted. They also display other social skills (e.g., sympathy for others when they are upset). Before age 2, toddlers seek peers of the same age group and begin to learn how to engage in play activities with them.

From the 2nd to the 3rd year of life. At age 2, toddlers exhibit an increased awareness of social norms when interacting with others, including offering assistance to others and waiting up to several minutes when are taking turns. Toddlers also begin to identify and regulate their emotions (e.g., to express when they are happy, sad, or angry and act accordingly). They sense when others are exhibiting familiar emotional expressions. Toddlers also begin to contribute to brief social discussions.

From the 3rd to the 5th year of life. As toddlers grow into young children at age 3, they are more aware of and attempt to make verbal amends if they injure or upset a peer, parent, or other caregiver. From the 4th to the 5th year of life, young children's previously developed social skills are displayed more fluently, in an integrated fashion, and with less prompting by caregivers.

Motor

Adaptive motor skills include fine and gross motor skills. Fine motor skills include the abilities that require a high degree of fine muscle control and precision (e.g., the use of fingers). Gross motor skills include the abilities required to control large muscles (e.g., crawling, walking, running). As previously noted, motor development typically develops in a proximal–distal and cephalo–caudal manner. Development typically begins with control of head movements and later extends to control of one's trunk and finally one's extremities (e.g., fingers and toes). During this period, motor skills range in difficulty from thrusting arms and legs during play to manipulating objects on clothing items.

From birth to the 1st year of life. Fine motor skills develop rapidly during this period. For example, infants clench their fists and move their hands to their mouth. They gain greater control of their arms and advance from just reaching to eventually reaching and grabbing an object. Before 6 months of age, infants tend to reach with one hand and grasp an object using their thumb in partial opposition to their fingers. Their later fine motor

skills enable them to use their entire hand to pick up small objects off the ground. They transfer an object between hands and use part of their thumb and fingers to grasp small objects. Infants use a palmar grasp when holding a crayon or pencil (e.g., holding or clenching an object with all four fingers).

Their gross motor skills also develop rapidly during this period. For example, they first thrust their arms and legs during play, among other activities, and gradually increase control of their head movements. Their further gross motor development enables them to gain mastery of their head movements and start to shift weight from one limb to another. At 6 months, infants pull themselves up as well as sit up without support for as long as 30 seconds. An infant's gross motor skills transition from sitting up with no support to adjusting to a crawling position and eventually to engaging in crawling. Infants later stand while holding onto an object, thus enhancing their ability to stand alone with no support. Infants also roll toys (e.g., a ball) and begin to develop the ability to throw a ball underhand to a caregiver.

From the 1st to the 2nd year of life. As infants grow into toddlers at age 1, their increased fine motor skills enable them to manipulate small objects (e.g., pick up small toys or coins). When given a crayon or pencil, toddlers first scribble spontaneously, using previously learned grasps to create strokes in any direction. At about the end of age 2, toddlers' fine motor skill development enables them hold a crayon or pencil using their fingers and part of their thumb when marking on paper. Toddlers are able to use a static tripod (i.e., thumb and two fingers) or quadrupod (i.e., thumb and three fingers) grasp during this period. These new grasps enable them to remove wrappings on small objects.

Toddlers' gross motor skills enable them to begin taking a few unassisted steps that later lead to their ability to walk short distances at first and haltingly without falling. In addition, toddlers begin first to walk up and later to walk down stairs one step at a time, first with help and later with little to no help from a caregiver. Their further gross motor development enables them to walk longer distances and balance on either foot. This more advanced sense of balance is exhibited when infants attempt to hit or kick a stationary ball.

From the 2nd to the 3rd year of life. At age 2, toddlers' fine motor skills further develop as they begin to hold the paper with one hand and draw with the other. They are able to catch objects thrown to them, albeit awkwardly at first. Their fine motor skill development allows them to draw straighter lines and curve shapes. Toddlers are also able to use scissors to cut paper.

Their gross motor skill development is apparent in their walking up and down at least three stairs with no support and engaging in jumping behaviors. Toddlers also develop the skills necessary to balance on either foot for at least 2 seconds without support and to peddle a tricycle.

From the 3rd to the 4th year of life. As toddlers grow into young children at age 3, their continued fine motor development is seen in their use of an advanced and more controlled grip when using a pencil or crayon to mark on paper. They are able to color within the outlines of shapes and objects.

Young children at this age gain further control of their ability to run and stop running when needed. Their advanced stair-walking skills enable them to walk up or down stairs while alternating feet and using no support. Young children also begin to imitate the postures of peers or other caregivers through social learning. They are able to carry objects a short distance (e.g., carrying a bag from a vehicle to the house).

From the 4th to the 5th year of life. At age 4, young children's fine motor skill development is seen in their ability to replicate written words when provided examples. They are also able to manipulate zippers on clothing items. During the last half of this period, young children's previously developed motor skills are displayed more fluently and in an integrated fashion.

IMPLICATIONS OF THE DESCRIPTION OF YOUNG CHILDREN'S DEVELOPMENT OF 10 ADAPTIVE SKILLS

An understanding of a somewhat detailed summary of children's adaptive behavior development is enlivened and thus enhanced by good theories of child development. Additionally, this somewhat detailed

summary may help verify and add to existing theory. An ultimate goal of good child development theory is to help define normal behavioral sequences in ways that help caregivers promote children's development.

Consistency of Adaptive Behavior Data and Theories

Adaptive skills encompass essential developmental features of infants and young children from birth through age 5. These skills enable infants to initiate communication with their caregivers, develop the motor skills necessary to move independently, and develop foundational social skills, among other skills. The review of normative data from commonly used standardized norm-referenced measures provides a wealth of information about the remarkable and fast-progressing development of adaptive skills during early childhood. An understanding of the relationship between theories of development and the data-based information enhances psychologists' understanding of both the theories and the data.

Drawing on Piaget's (1952) theory, one can envision infants' internal representation of their world into which they incorporate new experiences, leading to a continuous adaptation of prior mental representations of their world through the use of assimilation and accommodation. The previously reviewed information on the typical development of young children from birth to age 5 appears to support Piaget's emphasis on young children's sensorimotor and preoperational development—periods of rapid and discernible growth that provide a pathway to a better understanding of early childhood development. When examining the previously reviewed adaptive behavior data, young children adapt to their environment in many different ways. One of many examples is the way in which young children begin by holding small objects such as a pen or pencil in their entire hand and later adjust their grip to a static tripod that allows them to better accomplish their intended fine motor task.

Vygotsky's (1982) emphasis on the zone of proximal development underscores the reciprocal relationship between current and emerging areas of a young child's development together with the importance of guided development. Many of the previously reviewed communication, leisure, and social skills attest to this relationship. An infant's development is enhanced when attentive caretakers are aware of the stage of the young child's development, anticipate the forthcoming stage, and assist the infant to acquire skills characteristic of this stage. Thus, the richness and engagement of infants' social environment directly affects their development.

Gesell's (2007) emphasis on the biological basis of growth in infants and young children underscores the importance of both heredity and maturity in early childhood development. The skill areas previously reviewed generally develop in a uniform and integrative fashion. Gesell's emphasis on motor development through cephalo–caudal and proximal–distal control is supported by the data that indicate that infants first gain control of their head movements, followed by torso movements (e.g., rolling from their side to their back), leading to control of movement in their extremities (e.g., arms and legs), and finally developing control in their phalanges (e.g., fingers and toes).

Implications for Sequencing of Development

A child's development is characterized by the continuous integration of somewhat discrete skills into more complex yet general abilities that are increasingly displayed with greater ease and fluency. For example, an infant playing ball may at first be displaying only the development of fine motor skills. As the infant becomes older, the ball activity includes the development of communication and fine and gross motor skills, together with social and leisure skills.

Although most infants and young children follow a typical developmental sequence, the development of adaptive skills is not always continuous. The data from the three scales have suggested that, around age 5, children may require time during which they practice and integrate previously developed skills before progressing to acquire new skills. The preceding review identified several periods during early childhood development in which integration seems to take place.

Implications for Caregiver Understanding

Caregivers require knowledge of normal growth and development to understand and promote a child's development. Increased caregiver knowledge about child development leads to more optimal parenting behaviors, lower child–parent dysfunctional interaction, and lower perceived parental stress (Belcher, Watkins, Johnson, & Ialongo, 2007). Without this knowledge, caregivers and clinicians may have unrealistic developmental expectations, some too high (e.g., they believe normally developing children should be developing more quickly), and others too low (e.g., accepting developmental delays, given the belief the children will catch up later; Karraker & Evans, 1995). Parents of slow-developing children may also experience decreased well-being, including a sense of personal failure, concerns about later and permanent limitations, and frustrations about not knowing how to best intervene (Seltzer, Greenberg, Floyd, Pettee, & Hong, 2001).

INFLUENCE OF MENTAL RETARDATION–INTELLECTUAL DISABILITY AND OTHER DISABLING CONDITIONS ON ADAPTIVE BEHAVIOR

A diminished level of adaptive behavior has been viewed as evidence of mental retardation for centuries and continues to be used as an important marker of this disorder. People with mental retardation are expected to display diminished levels of adaptive behavior. Moreover, most clinical studies have focused on the adaptive behaviors of people with mental retardation.

Other clinical studies conducted during the standardization of the ABAS–II (e.g., see Harrison & Oakland, 2000, pp. 138–170; 2003) and the VABS–II (e.g., see Sparrow et al., 2005, pp. 137–158) investigated whether people with other disorders and disabilities also tend to display diminished levels of adaptive behavior and domains. These domains include the practical, social, and conceptual domains for the ABAS–II and communication, daily living skills, socialization, and motor skills domains for the VABS–II. These studies compared the adaptive behaviors of those with and without disabilities and disorders to determine whether these groups differ in adaptive behavior and thus whether the measures of adaptive behavior display desired properties (e.g., evidence of convergent and discriminant validity). Despite limitations in these studies (e.g., often small sample sizes and the nonrandom selection of people with the disorder), their results suggest that people with various disorders are likely to display diminished adaptive behaviors.

The following review examines adaptive behavior data from people with intellectual disability, developmental delays, known biological risk factors, motor and physical impairments, language disorders, autistic disorder, attention deficit/hyperactivity disorder, behavioral disorders and emotional disturbance, visual and hearing impairments, learning disabilities, Alzheimer's disease, and neuropsychological disorders.

People With Intellectual Disability

Most people with intellectual disability can be expected to display a pattern of adaptive behaviors that includes both strengths and weaknesses. Results from clinical studies that examine adaptive behavior strengths and weaknesses among persons with intellectual disability are reviewed in the sections that follow.

Ages 2 to 5. Infants and toddlers with intellectual disability may display diminished levels of adaptive behavior and skills at home and in daycare. Parent-reported data revealed that infants and toddlers with mild levels of intellectual disability had means of 66 on the total score and of less than 72 on three ABAS–II domains. Infants and toddlers with moderate levels of intellectual disability had means of 63 on total score and less than 69 on three domains. Two studies examined teacher-reported data. In one study, infants and toddlers with mild levels of intellectual disability had a total score mean of 67 (i.e., > 2 standard deviations) and means less than 72 for three ABAS–II domains. In the other study, infants and toddlers with moderate levels of intellectual disability had a total score mean of 65 and means less than 69 for three ABAS–II domains.

Ages 6 to 18. A study of parent-reported data on elementary and high school children found that

children with mild levels of intellectual disability displayed a total score mean of 66 and means less than 69 for the three VABS–II domains. Children with moderate levels of intellectual disability displayed a total score mean of 61, with the means for the three VABS–II domains all less than 64. Those children with severe levels of intellectual disability had a total score mean of 42 (i.e., > 4 standard deviations below the population mean) and means less than 45 for three VABS–II domains.

A study of teacher-reported data for children at these ages found those with mild levels of intellectual disability had a total score mean of 73 and means less than 71 for three ABAS–II domains. Children with moderate levels of intellectual disability had a total score mean of 59 and means less than 68 for three ABAS–II domains. Children with Down syndrome had a total score mean of 55, with three ABAS–II domains achieving means of less than 70.

Ages 19 to 86. Adults with mild levels of intellectual disability displayed a total score mean of 50 and means less than 57 for three VABS–II domains. Those with moderate levels of intellectual disability had total score means of 33 and means less than 41 for three VABS–II domains. Those with severe levels of intellectual disability had means of 20 for the total score and less than 23 for three domains.

Children with developmental delays ages 0 to 5. A study of parent-reported data indicated that infants and toddlers with developmental delays had a total score mean of 82 and means less than 86 for three ABAS–II domains. A study of teacher-reported data indicated that infants and toddlers with developmental delays had a total score mean of 84 and means less than 86 for three ABAS–II domains.

Children with known biological risk factors ages 0 to 2. A study of parent-reported data indicated infants and toddlers with known biological and physical conditions had a total score mean of 82 and means less than 87 for three ABAS–II domains. A study of teacher-reported data indicated infants and toddlers with known biological and physical conditions had a total score mean of 77 and means less than 81 for three ABAS–II domains.

Children With Motor and Physical Impairments

Most people with motor and physical impairments also can be expected to display a pattern of adaptive behaviors that includes both strengths and weaknesses. Results from clinical studies that examine adaptive behavior strengths and weaknesses among persons with motor and physical impairments are reviewed in the sections that follow.

Ages 0 to 5. Parent-reported data indicated that infants and toddlers with motor and physical impairments had a total score mean of 79 and means less than 87 for three ABAS–II domains. Motor and self-care skills were the lowest. Teacher-reported data indicated that infants and toddlers with motor and physical impairments had a total score mean of 76 and means less than 84 for three ABAS–II domains.

Ages 6 to 18. Teacher-reported data revealed that children between ages 6 and 18 with motor and physical impairments had a total score mean of 62 and means less than 74 for three ABAS–II domains.

Children With Receptive or Expressive Language Disorders Ages 2 to 6

Parent-reported data indicated that young children with receptive or expressive language disorders had a total score mean of 84 and means less than 87 for three ABAS–II domains. Teacher-reported data indicated that young children with receptive or expressive language disorders displayed a total score mean of 84 and means less than 87 for three ABAS–II domains.

Children With Pervasive Developmental Disorder Not Otherwise Specified Ages 3 to 5

Parent-reported data indicated that young children with pervasive developmental disorder not otherwise specified displayed a mean of 70 on total score and means less than 73 on three ABAS–II domains. Teacher-reported data indicated that young children with pervasive developmental disorder not otherwise specified had a total score mean of 66 and means less than 69 for three ABAS–II domains.

Children With Autistic Disorder

Family caretakers responsible for those with autistic disorder together with professionals who work with them typically recognize the important role of adaptive behaviors and skills to normalizing behaviors. Children, youth, and adults with this disorder can be expected to display a pattern of adaptive behaviors that includes both strengths and weaknesses. Results from clinical studies that examine adaptive behavior strengths and weaknesses among persons with autistic disorder are reviewed in the sections that follow.

Ages 3 to 5. Parent-reported data indicated that young children with autistic disorder had a total score mean of 64 and means less than 72 for three ABAS–II domains. Teacher-reported data indicated that young children with an autistic disorder had a total score mean of 67 and means less than 73 for three ABAS–II domains.

Ages 3 to 16. Parent-reported data indicated that children with autism and a verbal disorder had a total score mean of 66 and means less than 81 for four VABS–II domains. Parent-reported data indicated children with autism and a nonverbal disorder had a total score mean of 51 and means less than 67 for four VABS–II domains. Teacher-reported data indicated children with an autistic disorder had a total score mean of 54 and means less than 65 for three ABAS–II domains.

Children With Attention-Deficit/ Hyperactivity Disorder Ages 6 to 21

Parent-reported data indicated that children and youth with attention deficit/hyperactivity disorder had a total score mean of 91 and were in the low average range on three ABAS–II domains. Teacher-reported data indicated that children and youth with attention deficit/hyperactivity disorder displayed a total score mean of 77 and means less than 81 for three ABAS–II domains.

Children With Behavior Disorders and Emotional Disturbance Ages 6 to 18

Two studies examined parent-reported data on children with behavior disorders and emotional disturbance. One found a total score mean of 78 and means less than 82 for three ABAS–II domains, and the other found a total score mean of 86 and means less than 92 on three VABS-II domains. Teacher-reported data indicated that children with behavior disorders and emotional disturbance had a total score mean of 77 and means less than 81 for three ABAS–II domains.

Children With Visual Impairment Ages 6 to 18

Parent-reported data indicated that children with a visual impairment had a total score mean of 87 and means for the daily living skills and socialization VABS–II domains in the 80s.

Children With Hearing Impairment Ages 6 to 18

Parent-reported data on children with hearing impairment showed a total score mean of 90 and average VABS–II domain scores. Teacher-reported data on children with hearing impairment showed a total score mean of 93 and average ABAS–II domain scores.

Children With Learning Disability Ages 6 to 18

Two studies examined parent-reported data on children with learning disabilities. One found a total score mean of 88 and average ABAS–II domain scores, and the other found a total score mean of 95 and low average VABS–II domain scores. Three studies examined teacher-reported data on children with learning disability. One found a total score mean of 91 and low average ABAS–II domain scores. Another found a total score mean of 84 and means in the 80s for three domains. The third study found a total score mean of 87 and low average ABAS–II domain scores.

Adults With Alzheimer's Disease Ages 60 to 89

Adults with Alzheimer's disease, when rated by others, had a total score mean of 61 and means less than 75 on three ABAS–II domains.

Adults With Neuropsychological Disorders Ages 18 to 89

Adults with neuropsychological disorders, when rated by others, displayed a total score mean of 67

and means less than 80 for three ABAS–II domains. Adults with neuropsychological disorders, when rated by themselves, had a total score mean of 82 and means in the 80s on three ABAS–II domains.

Thus, people with various disabilities and disorders are likely to display diminished adaptive behaviors. Developmental and rehabilitation services are often intended to promote the independent display of behaviors associated with meeting a person's daily personal and social needs, including behaviors expected at home, school, and other social environments. Thus, adaptive behavior data may be critical when planning and monitoring interventions and evaluating progress of people with various disabilities and disorders.

ASSESSMENT OF ADAPTIVE BEHAVIOR

Measures of adaptive behavior are available in Canada, the United States, and a number of other Western countries to assist professionals in their work. These measures may assist in describing adaptive behavior and skills, understanding the pervasive influence disabilities and disorders may exert on adaptive behavior and skills, identifying strengths and weaknesses, and assisting in intervention planning and monitoring.

Describe Adaptive Behaviors and Skills Accurately

The first goal of assessment is to accurately describe behavior. An accurate description is needed when making a diagnosis, when engaged in determining strengths and weaknesses, and when planning and monitoring interventions. The assessment of adaptive behavior typically relies heavily on information from respondents, not from the examiner's direct observation of behavior. Thus, when assessing adaptive behavior, respondents must have an instrumental level of knowledge of the behaviors and skills measured by the instrument and complete the items honestly. Information from more than one knowledgeable respondent allows examiners to assess the reliability of the information obtained (Bothwick-Duffy, 2000; Tassé & Lecavalier, 2000).

Understand the Pervasive Influence Disabilities and Disorders May Exert on Adaptive Behavior and Skills

As noted earlier, the services of psychologists and others engaged in the diagnosis of mental retardation–intellectual disability or other disorders for which adaptive behavior data are important are guided by professional and legal standards. Additionally, information on adaptive behavior and skills will enhance their work with young children who display attention, autism, communication, conduct, elimination, feeding and eating, learning, motor skills, and pervasive developmental disorders (Harman et al., 2010; Oakland & Harrison, 2008). This information is also useful when working with older children and adolescents who display various disorders, including attention deficit/hyperactivity disorder, acquired brain injury, auditory or visual impairment, autism, developmental delays, emotional or behavioral disorders, learning disabilities, and physical impairments (Ditterline et al., 2008; Harrison & Oakland, 2003; Oakland & Harrison, 2008). Adaptive behavior information can assist professionals in their work with adults suspected of displaying one or more disorders associated with anxiety, acute stress adjustment, bipolar, dependency, depression, mood, psychosis, Parkinson's, postpartum, substance abuse, schizophrenia, sleep, and other disorders—to name a few—as they are likely to display impairments in their functional daily living skills.

Identify Strengths and Weaknesses in Light of Environmental Needs

Clinicians typically use various methods to determine possible strengths and weaknesses in a person's adaptive behavior profile. For example, mean scores, standard errors of measurement, and confidence intervals are examined when determining strengths and weaknesses. These data enable clinicians to review adaptive behavior profiles normatively (i.e., how a person compares with his or her peers) and ideographically (i.e., although a person's scores may be normatively below average, an ideographic review may reveal some scores to be higher than others, thus revealing personal strengths).

Clinicians also identify adaptive skill strengths and deficits in light of environmental requirements in an effort to determine what skills are most needed to function effectively and independently in one or more environments (e.g., home, school, work, and the community). Then they determine whether the skills are sufficiently developed to allow the person to function independently. Thus, some skills with low scores can be overlooked because these skills are not critical to successful functioning in the person's current environments. Clinicians are encouraged to focus on promoting those adaptive skills that are most needed in the person's environment and are within a person's zone of proximal development. Once acquired, these skills would enable the person to function more independently and successfully in his or her environment.

Intervention Planning and Monitoring

Intervention planning has become increasingly important to the work of professionals and others responsible for caring for people with special needs. Parents, professionals, and others expect assessment specialists to use their findings by making recommendations that stabilize or advance desired behaviors. Thus, professionals are asked to engage in an assessment process that will lead to intervention outcomes. Scales of adaptive behavior assess specific, practical, and functional skills that are amenable to change through interventions. Thus, information from these tests is especially useful when engaged in intervention planning and monitoring.

Intervention planning and monitoring efforts often require clinicians to examine behaviors at the item level rather than at the domain or GAC levels. Although information at the domain and GAC levels is often useful when making diagnoses, this information is unlikely to inform efforts to promote the development of specific adaptive skills that are needed to function independently and effectively in one's environment. Thus, those engaged in intervention efforts are encouraged to identify specific adaptive behaviors reflected at the item level. These specific behaviors are more amenable to change than those summarized by domain or GAC scores. For example, the ABAS–II provides an intervention planner that helps identify interventions designed to promote item-level behaviors.

Research

Efforts to better understand the impact of disabilities and disorders on a person's functional behaviors require increased research on their adaptive behavior and skills. As noted early, people with various disorders other than an intellectual disability are also likely to display functional life skill deficits. The extent to which efforts by rehabilitation specialists and others promote functional skill development that leads to improvement in daily living skills should also be subject to further research (Mpofu & Oakland, 2010a, 2010b).

REVIEW OF THREE ADAPTIVE BEHAVIOR MEASURES

The Buros *Mental Measurements Yearbook* cited 26 measures of adaptive behavior. Thus, clinicians have a wide array of adaptive behavior measures from which to choose. Three such measures that are reviewed in this section were selected because they provide a comprehensive assessment of adaptive skills and behaviors displayed in various environments and were normed nationally on a broad age range.

Adaptive Behavior Assessment System—Second Edition

The ABAS–II (Harrison & Oakland, 2003) provides an assessment of adaptive behavior and skills for people from birth through age 89. Five forms of the ABAS–II are available: Parent/Primary Caregiver Form (for ages 0–5), Teacher/Day Care Provider Form (for ages 2–5), Parent Form (for ages 5–21), Teacher Form (for ages 5–21), and an Adult Form (for ages 16–89). Parent forms are available in Spanish, and all five forms are available in Canadian French.

The ABAS–II was normed on 7,370 individuals from birth through age 89. Its standardization sample is representative of U.S. census data from 1999 through 2000 in reference to gender, race and ethnicity, parental education, and proportion of individuals with disabilities (Harrison & Oakland, 2003).

The ABAS–II presents items in accord with the AAMR's 1992 and 2002 definitions of adaptive behavior. This provision includes the following domains and skill areas: conceptual (including communication, functional academics, and self-direction), social (including social skills and leisure), and practical (including community use, home/school living, health and safety, and self-care) domains. Motor is assessed only for ages 0 to 5. In addition to the domains and skill areas, a GAC score is derived from all skill scores.

The ABAS–II demonstrates strong psychometric qualities. Internal consistency is high, with reliability coefficients of .85 to .99 for the GAC, three adaptive behavior domains, and skill areas. Test–retest reliability coefficients are in the .80s and .90s for the GAC, three domains, and skill areas (Harrison & Oakland, 2003). ABAS–II interrater reliability coefficients (e.g., between teachers, daycare providers, and parents) range from the .60s to the .80s for the skill areas and are in the .90s for the GAC. The ABAS–II's construct validity is strong, as evidenced by factor analyses (Aricak & Oakland, 2010; Wei et al., 2008). Its concurrent validity with the Vineland Adaptive Behavior Scales—Classroom Edition's Adaptive Behavior Composite is .82 (Harrison & Oakland, 2003). Clinical validity is also highly evident, demonstrating that the scales are assessing similar constructs.

Reviews of the ABAS–II noted several advantages over other measures (Burns, 2005; Meikamp & Suppa, 2005), including that the behavior domains align with the AAMR's (2002) recommendations; the scale allows for multiple respondents from multiple settings and an adult self-report; the ABAS–II allows one to anticipate the development of emerging behaviors. Moreover, the scale provides respondents the opportunity to answer each question without a trained interviewer present.

The *Mental Measurements Yearbook* provides two reviews of the ABAS–II. One review concludes that the measure's theory is sound and empirically supported, its norms large and sufficiently represented, and the GAC is adequately reliable for the scale's intended purpose (Burns, 2005). This review reported further that the data supporting the ABAS–II's reliability and validity are impressive, and efforts to link data to intervention planning are commendable. The review concluded that use of the ABAS–II could strengthen most comprehensive psychoeducational assessments. In short, Burns (2005) believed the ABAS–II was technically superior to most of its competitors. In addition to Burns's positive comments, he cautioned against using ABAS–II skill scores. A second review (Meikamp & Suppa, 2005) generally concurred with Burns's evaluation and suggested the need to increase the instrument's normative sample size despite the fact that the ABAS–II norm sample includes 7,370 individuals, the largest sample of any adaptive behavior.

Scales of Independent Behavior—Revised

The SIB–R; Bruininks et al., 1996) provides users with three forms: a Short Form, an Early Development Form, and the Full Scale Form. The Short Form is used as a screener for all ages. This scale contains items from the 14 subscales that make up the Full Scale Form. The Early Development Form is designed for use with children from infancy through age 6 or with older individuals with severe disabilities that place their functioning at developmental levels younger than age 8.

The SIB–R norming data were first collected in the 1980s to reflect 1980 census data and were originally used for the SIB. In the 1990s, further data were collected to reflect the 1990 census and were used to supplement the original SIB data to form norms for its revised form, the SIB–R. The 779 children who made up the standardization sample were stratified by sex, type of community, geographic region, race and ethnicity, and socioeconomic status.

The SIB–R measures adaptive behavior in four broad categories that include 14 skill areas: motor skills (including gross motor skills and fine motor skills), social interaction and communication skills (including social interaction, language comprehension, and language expression), personal living skills (including eating and meal preparation, toileting, dressing, personal self-care, and domestic skills), and community living skills (including time and punctuality, money and value, work skills, and home–community orientation). A Broad Independence score is derived from scores in these areas.

The SIB–R includes a Problem Behavior Scale that facilitates an assessment of problem behavior in three domains and eight problem areas: internalized maladaptive behavior (including hurtful to self, unusual or repetitive habits, and withdrawal or inattentive behavior), asocial maladaptive behavior (including socially offensive and uncooperative behaviors), and externalized maladaptive behavior (including hurtful to others, destructive to property, and disruptive behavior). A General Problem Behaviors score is derived from scores in these areas. The AAIDD views problem or maladaptive behaviors as different from adaptive behaviors. Thus, although clinicians may assess these behaviors, such behaviors should not inform judgments of a person's adaptive behavior and skills (Schalock & Braddock, 1999).

The SIB–R has high internal consistency, with most correlations in the high .80s and .90s. Interrater reliability is also high, generally in the .80s and .90s. Test–retest reliability is strong (>.95) for most forms. However, coefficients for the Maladaptive Behavior Scale are lower and generally range from .74 to .92. The reliability coefficients of the Short Form and Early Development Form are also low (Maccow & Zlomke, 2001).

The SIB–R displays concurrent validity (e.g. correlations mostly in the .90s) with the original version of the SIB. Also, correlations between the SIB–R Early Development Form and the Early Screening Profiles Self-Help and Social Profiles (from the VABS–II) range from .77 to .90.

Reviews of the SIB–R noted various positive features (Maccow & Zlomke, 2001). The SIB–R is easy to administer and score. It provides information about problem behaviors that may interfere with independent functioning. In addition, training objectives are provided at the end of each subscale to determine which of an individual's skills are most in need of improvement. The reviews of the SIB–R also listed some weaknesses, including an inability to measure adaptive skills through direct observation of individuals. The authors believe the criticism is unwarranted because information needed to complete measures of adaptive behavior comes from knowledgeable respondents, not from direct observations. The reliability of the Short Form and Early Development Form was questioned.

Vineland Adaptive Behavior Scales—Second Edition

Edgar Doll, while director of research at the Training School at Vineland (Vineland, NJ), developed a measure of social maturation, the Vineland Social Maturity Scale, and standardized it with people who were typically developing and people with mental deficiencies (Doll, 1936). Interventions designed to promote social development among people with mental retardation were somewhat common during the 1930s and 1940s. Thus, the Vineland Social Maturity Scale was developed, in part, to provide an objective norm-referenced assessment of the social development of children and adults. When examined in light of current concepts, this measure resembles features in current measures of adaptive behavior. The VABS, published in 1984, was a revision of the Vineland Social Maturity Scale (Oakland & Houchins, 1985).

The VABS–II (Sparrow et al., 2005) has four forms: a Survey Interview Form (for birth–age 90), Expanded Interview Form (for birth–age 90; recommended for younger ages or low-functioning individuals), Parent/Caregiver Form (age ranges not provided), and a Teacher Rating Form (for ages 3–21). The Survey Interview and Expanded Interview forms are administered by a professional using a semistructured interview format. The Parent/Caregiver Form uses a checklist procedure to assess the same content as the Survey Interview Form and may be used when an interview is not possible or for progress monitoring purposes. A respondent may complete the Teacher Rating Form independently.

The VABS–II was normed on 3,695 individuals from birth through age 90. Its standardization is based on a nationally representative norm group consistent with the 2001 U.S. population, including age, sex, race or ethnicity, socioeconomic status, geographic region, and educational placement.

The VABS–II provides various scores. The Adaptive Behavior Composite score is derived from scores on four domains: communication (including receptive, expressive, and written skills), daily living skills (including personal, domestic, and community), socialization (including interpersonal relationships, play and leisure time, and coping skills), and motor skills (including gross and fine motor).

The VABS–II also provides a measure of maladaptive behaviors, including internalizing and externalizing behaviors.

The VABS–II's internal consistency is high, generally ranging from .84 to .97 for the GAC, domains, and subdomains. Test–retest reliability is generally in the mid-.80s for the GAC. Interrater reliability for parents is in the .70s to .80s for the GAC. The VABS–II has concurrent validity with other related measures of adaptive behavior (e.g., correlations of .52 to .70 with the ABAS–II in overall composite scores). Although the VABS–II does not currently measure adaptive behavior as defined by AAMR, it has strong construct validity supporting its theoretical underpinnings, as evidenced by factor analysis studies (Sparrow et al., 2005). Correlations between the domain and Adaptive Behavior Composite scores from the VABS first edition and the VABS–II are in the .80 to .95 range.

Reviews of the VABS–II noted various positive features of the instrument (Stein & Widaman, 2010). The VABS–II is an improved version of the VABS, one that provides detailed and simple-to-perform administration and scoring instructions. The VABS–II has excellent internal consistency and test–retest reliability as well as firm evidence for content, concurrent, and construct validity. The norm group is impressive in size and an adequate representation of the population. The reviewers identified several weaknesses, including low levels of interrater reliability and difficulty in accurately assessing individuals who are at higher levels of adaptive behavior or skills. This latter issue is generic to all measures of adaptive behavior and may not be a problem in that the primary use of adaptive behavior measures is to identify possible weaknesses, not to identify those in the above average or gifted range.

SUMMARY

The concept and assessment of adaptive behavior have had a long history, one that dates at least to early Greek civilization. Although its importance has been long recognized, the formal measurement of adaptive behavior is relatively recent—perhaps best dated by the 1994 publication of the first edition of the VABS. The use of measures of adaptive behavior together with measures of intelligence in diagnosing mental retardation has been most common. However, the use of these two types of measure to diagnose mental retardation is also somewhat recent. AAMR's 1959 definition of mental retardation was the first to emphasize their joint use.

Psychologists' understanding of the pervasive impact that disabilities and disorders other than mental retardation may have on the development and display of adaptive behavior and related skills is increasing. Thus, clinicians are becoming more aware of the importance of acquiring information on an individual's ability to independently display behaviors associated with meeting his or her daily personal and social needs, including behaviors expected in domestic and social environments. These qualities may constitute the foundation for a quality life.

Adaptive behavior develops remarkably fast from birth through ages 5 or 6. Theory, research, and other forms of scholarship inform professionals as to its development during this critical period. Less is known about its development and decline during later ages. Legal and professional standards that govern diagnostic and intervention services by psychologists and other professionals acknowledge the importance of adaptive behavior and skills. Professionals are fortunate to have a number of measures of adaptive behavior to assist them in this work.

References

American Association on Intellectual and Developmental Disabilities.(2010). *Intellectual disability: Definition, classification, and systems of support* (10th ed.). Washington, DC: Author.

American Association on Mental Retardation. (1992). *Mental retardation: Definition, classification, and systems of support*. Washington, DC: Author.

American Association on Mental Retardation. (2002). *Mental retardation: Definition, classification, and systems of support* (10th ed.). Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

American Psychiatric Association. (2010). *DSM–5 development*. Retrieved from http://www.dsm5.org/ProposedRevisions/Pages/ proposed revision.aspx?rid=384

Apling, R. N., & Jones, N. L. (2005). *Congressional research service report for Congress, Individuals With Disabilities Education Act (IDEA): Analysis of changes made by P. L. 108–446*. Washington, DC: Library of Congress.

Aricak, O. T., & Oakland, T. (2010). Multigroup confirmatory factor analysis for the teacher form, ages 5–21, of the Adaptive Behavior Assessment System–II. *Journal of Psychoeducational Assessment, 28*, 578–584. doi:10.1177/0734282909350209

Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development, Third Edition: Technical manual*. San Antonio, TX: Harcourt Assessment.

Beadle-Brown, J., Murphy, G., & Wing, L. (2005). Long-term outcome for people with severe intellectual disabilities: Impacts of social impairment. *American Journal on Mental Retardation, 110*, 1–12. doi:10.1352/0895-8017(2005)110<1:LOFPWS>2.0.CO;2

Belcher, H., Watkins, K., Johnson, E., & Ialongo, N. (2007). Early Head Start: Factors associated with caregiver knowledge of child development, parenting behavior, and parenting stress. *Early Childhood Research Quarterly, 10*, 6–19.

Binet, A., & Simon, T. (1912). *A method of measuring the development of the intelligence of young children*. Lincoln, IL: Courier.

Bruininks, R., Woodcock, R., Weatherman, R., & Hill, B. (1996). *Scales of Independent Behavior—Revised*. Chicago, IL: Riverside.

Burns, M. K. (2005). [Review of the Adaptive Behavior Assessment System—Second Edition]. In R. Spies & B. Plake (Eds.), *The sixteenth mental measurements yearbook* (pp. 17–19). Lincoln, NE: Buros Institute of Mental Measurements.

Cole, M., & Cole, S. (1996). *The development of children*. New York, NY: Worth.

Council for Exceptional Children. (2004). *The new IDEA: CEC's summary of significant issues*. Arlington, VA: Author.

Daly, W. (2004). Gesell's infant growth orientation: A composite. *Journal of Instructional Psychology, 31*, 321–324.

De Ridder, D. T. D., Schreurs, K. M., & Bensing, J. (1998). Adaptive tasks, coping and quality of life of chronically ill patients: The case of Parkinson's disease and chronic fatigue syndrome. *Journal of Health Psychology, 3*, 87–101. doi:10.1177/135910539800300107

DeVries, R. (2008). Theory and practice in early childhood education. In T. L. Good (Ed.), *21st century education: A reference handbook*. New York, NY: Sage.

Ditterline, J., Banner, D., Oakland, T., & Becton, D. (2008). Adaptive behavior profiles of students with disabilities. *Journal of Applied School Psychology, 24*, 191–208. doi:10.1080/15377900802089973

Doll, E. (1936). Preliminary standardization of the Vineland Social Maturity Scale. *American Journal of Orthopsychiatry, 6,* 283–293. doi:10.1111/j.1939-0025.1936.tb05235.x

Emerson, E., Robertson, J., Gregory, N., Hatton, C., Kessissoglou, S., Hallam, A., . . . Netten, A. (2000). Quality and costs of community-based residential supports, village communities, and residential campuses in the United Kingdom. *American Journal on Mental Retardation, 105*, 81–102. doi:10.1352/0895-8017(2000)105<0081:QACOCR>2.0.CO;2

Gesell, A. (1930). *The guidance of mental growth in infant and child*. New York, NY: Macmillan.

Gesell, A. (1952). Autobiography. In E. G. Boring, H. Werner, H. S. Langfeld, & R. M. Yerkes (Eds.), *A history of psychology in autobiography* (Vol. 4, pp. 123–142). Worcester, MA: Clark University Press. doi:10.1037/11154-006

Gesell, A. (2007). *Studies in child development*. Minneapolis, MN: Jessen Press.

Gesell, A., Halverson, H. M., & Amatruda, C. S. (1940). *The first five years of life: A guide to the study of preschool children*. New York, NY: Harper & Brothers.

Greenspan, S. (2006). Functional concepts in mental retardation: Finding the natural essence of an artificial category. *Exceptionality, 14*, 205–224. doi:10.1207/s15327035ex1404_3

Greenspan, S., & Driscoll, J. (1997). The role of intelligence in a broad model of personal competence. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 131–150). New York, NY: Guilford Press.

Harman, J., Smith-Bonahue, T., & Oakland, T. (2010). Assessment of adaptive behavior development in young children. In E. Mpofu & T. Oakland (Eds.), *Rehabilitation and health assessment: Applying ICF guidelines* (pp. 333–352). New York, NY: Springer,.

Harries, J., Guscia, R., Kirby, N., Nettelbeck, T., & Taplin, J. (2005). Support needs and adaptive behaviors. *American Journal on Mental Retardation, 110*, 393–404. doi:10.1352/0895-8017(2005)110[393:SNAAB]2.0.CO;2

Harrison, P., & Oakland, T. (2000). *Adaptive Behavior Assessment System*. San Antonio, TX: Harcourt Assessment.

Harrison, P., & Oakland, T. (2003). *Adaptive Behavior Assessment System* (2nd ed.). San Antonio, TX: Harcourt Assessment.

Heber, R. (1959). A manual on terminology and classification in mental retardation. *American Journal of Mental Deficiency Monographs, 64*(2, Suppl.).

Hedegaard, M. (2005). The zone of proximal development for instruction. In H. Daniels (Ed.), *An introduction*

to Vygotsky (pp. 171–195). London, England: Routledge.

Individuals With Disabilities Education Improvement Act of 2004, Pub. L. 108-446, 20 U.S.C. § 1400 *et seq.* Retrieved from http://idea.ed.gov

Karraker, K., & Evans, S. (1996). Adolescent mothers' knowledge of child development and expectations for their own infants. *Journal of Youth and Adolescence, 25*, 651–666. doi:10.1007/BF01537359

Maccow, G., & Zlomke, L. (2001). [Review of the Scales of Independent Behavior–Revised]. In B. Plake & J. Impara (Eds.), *The fourteenth mental measurements yearbook* (pp. 1073–1077). Lincoln, NE: Buros Institute of Mental Measurements.

Matson, J. L., Rivet, T., Fodstad, J., Dempsey, T., & Boisjoli, J. (2009). Examination of adaptive behavior differences in adults with autism spectrum disorders and intellectual disability. *Research in Developmental Disabilities, 30*, 1317–1325. doi:10.1016/j.ridd.2009.05.008

McGrew, K., & Bruininks, R. (1989). The factor structure of adaptive behavior. *School Psychology Review, 18*, 64–81.

Meikamp, J., & Suppa, C. H. (2005). [Review of the Adaptive Behavior Assessment System—Second Edition]. In R. Spies & B. Plake (Eds.), *The sixteenth mental measurements yearbook* (pp. 19–21). Lincoln, NE: Buros Institute of Mental Measurements.

Mpofu, E., & Oakland, T. (Eds.). (2010a). *Assessment in rehabilitation and health.* Upper Saddle River, NJ: Merrill.

Mpofu, E., & Oakland, T. (Eds.). (2010b). *Rehabilitation and health assessment: Applying ICF guidelines.* New York, NY: Springer.

Nihira, K., Leland, H., & Lambert, N. (1993). *AAMR Adaptive Behavior Scale—Residential and Community* (2nd ed.). Austin, TX: Pro-Ed.

Oakland, T., & Algina, J. (2011). Adaptive Behavior Assessment System—II Parent/Primary Caregiver Form: Ages 0–5: Its factor structure and other implications for practice. *Journal of Applied School Psychology, 27*, 103–117.

Oakland, T., & Harrison, P. (2008). *Adaptive Behavior Assessment System—II: Clinical use and interpretation.* London, England: Academic Press.

Oakland, T., & Houchins, S. (1985). A review of the Vineland Adaptive Behavior Scales, Survey Form. *Journal of Counseling and Development, 63*, 585–586. doi:10.1002/j.1556-6676.1985.tb00689.x

Olley, J. G., & Cox, A. W. (2008). Assessment of adaptive behavior in adult forensic cases: The use of the ABAS-II. In T. Oakland & P. Harrison (Eds.), *Adaptive Behavior Assessment System—II: Clinical use and interpretation* (pp. 381–398). London, England: Academic Press. doi:10.1016/B978-012373586-7.00020-5

Piaget, J. (1924). Judgment and reasoning in the child. In H. Grober & J. Voneche (Eds.), *The essential Piaget* (pp. 89–118). New York, NY: Basic Books.

Piaget, J. (1951). *The psychology of intelligence.* London: Routledge & Kegan Paul.

Piaget, J. (1952). *The origins of intelligence in children.* New York, NY: Norton. doi:10.1037/11494-000

Piaget, J. (1963). The first acquired adaptions and the primary circular reaction. In M. Cook (Trans.), *The origins of intelligence in children* (pp. 47–152). New York, NY: Norton.

Piaget, J. (1971). The epistemology of elementary levels of behavior. In B. Walsh (Trans.), *Biology and knowledge: An essay on the relations between organic regulations and cognitive processes* (pp. 214–265). Chicago, IL: University of Chicago Press.

Piaget, J., & Inhelder, B. (1969). *The psychology of the child.* New York, NY: Basic Books.

Prasher, V. P., Chung, M., & Haque, M. S. (1998). Longitudinal changes in adaptive behavior in adults with Downs syndrome: Interim findings from a longitudinal study. *American Journal on Mental Retardation, 103*, 40–46. doi:10.1352/0895-8017(1998)103<0040:LCIABI>2.0.CO;2

Rubel, A., Reinsch, S., Tobis, J., & Hurrell, M. (1995). Adaptive behavior among very elderly Americans. *Physical and Occupational Therapy in Geriatrics, 12*, 67–78. doi:10.1080/J148v12n04_05

Sadurní, M., Perez Burriel, M., & Plooij, F. X. (2010). The temporal relation between regression and transition periods in early infancy. *Spanish Journal of Psychology, 13*, 112–126.

Schalock, R., & Braddock, D. (Eds.). (1999). *Adaptive behavior and its measurement.* Washington, DC: American Association on Intellectual and Developmental Disabilities.

Seltzer, M. M., Greenberg, J., Floyd, F., Pettee, Y., & Hong, J. (2001). Life course impacts of parenting a child with a disability. *American Journal on Mental Retardation, 106*, 265–286. doi:10.1352/0895-8017(2001)106<0265:LCIOPA>2.0.CO;2

Sparrow, S., Cicchetti, D., & Balla, D. (2005). *Vineland Adaptive Behavior Scales* (2nd ed.). Circle Pines, MN: AGS.

Stein, S., & Widaman, K. (2010). [Review of the Vineland Adaptive Behavior Scales –Second Edition]. In R. Spies, J. Carlson, & K. Geisinger (Eds.), *The eighteenth mental measurement yearbook* (pp. 677–684). Lincoln, NE: Buros Institute of Mental Measurements.

Tassé, M. J., & Lecavalier, L. (2000). Comparing parent and teacher ratings of social competence and problem behaviors. *American Journal on Mental Retardation, 105*, 252–259. doi:10.1352/0895-8017(2000)105<0252:CPATRO>2.0.CO;2

U.S. Code Service. (2007). *20 USCS § 1400: Individuals With Disabilities Education Act*. Retrieved from http://web.lexisnexis.com/universe/document?_m= 6af9d-78cf01c007d8151e26c38f4192dand_docnum=1and wchp=dGLbVlzzSkVband_ md5=412bb646231d76b75426f221f837233a

Van De Rijt-Plooij, H. H. C., & Plooij, F. X. (1992). Infantile regressions: Disorganization and onset of transition periods. *Journal of Reproductive and Infant Psychology, 10*, 129–149. doi:10.1080/02646839208403946

Van De Rijt-Plooij, H. H. C., & Plooij, F. X. (1993). Distinct periods of mother–infant conflict in normal development: Sources of progress and germs of pathology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 34*, 229–245. doi:10.1111/j.1469-7610.1993.tb00981.x

Vygotsky, L. (1978). Interaction between learning and development. In M. Cole (Ed.), *Mind in society: The development of higher psychological processes* (pp. 79–91). Cambridge, MA: Harvard University Press.

Vygotsky, L. (1987). The development of scientific concepts in childhood. In R. Rieber & A. Carton (Eds.), *The collected works of L. S.Vygotsky: Vol. 1. Problems of general psychology, including the volume* Thinking and Speech (pp. 167–242). New York, NY: Plenum Press. (Original work published 1934)

Wei, Y., Oakland, T., & Algina, J. (2008). Multigroup confirmatory factor analysis for the parent form, ages 5–21, of the Adaptive Behavior Assessment System-II. *American Journal on Mental Retardation, 113*, 178–186. doi:10.1352/0895-8017(2008)113[178:MCFAFT]2.0.CO;2

Woolf, S., Woolf, C., & Oakland, T. (2010). Adaptive behavior among adults with intellectual disabilities and its relationship to community independence. *Intellectual and Developmental Disabilities, 48*, 209–215. doi:10.1352/1944-7558-48.3.209

World Health Organization. (2001). *International classification of functioning, disability, and health*. Geneva, Switzerland: Author.

Zigman, W. B., Schupf, N., Urv, T., Zigman, A., & Silverman, W. (2002). Incidence and temporal patterns of adaptive behavior change in adults with mental retardation. *American Journal on Mental Retardation, 107*, 161–174. doi:10.1352/0895-8017(2002)107<0161:IATPOA>2.0.CO;2

CHAPTER 10

TESTING FOR LANGUAGE COMPETENCE IN CHILDREN AND ADOLESCENTS

Giselle B. Esquivel and Maria Acevedo

This chapter provides a multidimensional perspective on major theoretical and applied aspects of language competence assessment in children and adolescents. A brief review of language theories provides a basis for understanding the nature of language competencies and assessment approaches. Likewise, a description of the core anatomical areas in the brain linked to language, language processing, and speech production highlights the multifaceted foundations of language. Applied language assessment issues addressed in the chapter initially focus on typical populations exhibiting modal language development patterns. Emphasis is given to (a) assessment of structural literacy skills, such as phonological processing, receptive and expressive vocabulary, reading, and writing, and (b) functional language assessment of spoken language, pragmatics, and social communication competencies. Attention is also given to language assessment of second language learners (emerging bilinguals) operating at various levels of language proficiency in more than one language. Last, the chapter considers special populations with unique language characteristics such as students who are deaf or hard of hearing, individuals with speech and language disorders, and children with neurodevelopmental disorders. Conceptual and ethical considerations in test selection and testing procedures are incorporated throughout as an essential aspect of language competence assessment.

OVERVIEW OF LANGUAGE THEORIES

What is language? How is it acquired? Philosophers, linguists, and psychologists interested in the study of psycholinguistics have attempted to answer these kinds of questions for years. Although there is still no universally accepted definition of *language* (Kolb & Whishaw, 2003), various theories, conceptual models, and perspectives have emerged to address critical questions concerning language. Some of the major theories that have influenced psychologists studying language development are learning theory, nativism theory, and interactionist theory perspectives (Shaffer, 1985).

Learning Theory

Among the principal proponents of learning theory, Skinner (1957) argued that a child's language was shaped by adults reinforcing particular aspects of utterances that most resembled adult speech, thus increasing the likelihood of these utterances being repeated. This shaping would initially take place with babbling and would proceed to words, then to word combinations, and eventually to longer grammatical sentences. That is, a child's grammatical speech would evolve from caregivers reinforcing successive approximation of correct grammar (Shaffer, 1985). Learning theory emphasizes observational learning, modeling, and operant conditioning in language development and a behavioral approach to language instruction and assessment practices.

Nativism Theory

During the same time period, Chomsky (1957) challenged Skinner's (1957) traditional stimulus–response learning theory approach to language development with the theory of nativism, which posits that human beings have at birth an innate set

DOI: 10.1037/14049-010
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

of cognitive and perceptual schemas that are biologically programmed to acquire language (Greene, 1973; Shaffer, 1985). In *Syntactic Structures*, Chomsky developed the idea of a generative grammar model that assumes that grammatical structures are universal and provide the basis for the acquisition of language in general, although the idiosyncratic features of language are learned. Also, he incorporated the concept of how deep structures and surface levels of language connect in language development. Thus, this perspective gives emphasis to the concept that deep language structures are shared across languages and surface-level features are language specific. This perspective is consistent with current theories of second language acquisition and language transference (Cummins, 1984). Language assessment is focused on both surface levels (e.g., vocabulary competence) and grammatical structures of language (e.g., syntax). Likewise, bilingual assessment is based on the development of basic interpersonal communicative skills and cognitive academic language proficiency (CALP) in more than one language (Cummins, 1984).

Interactionist Theory

Cognitive theorist Piaget (1970) advanced the understanding of language by proposing an interactive perspective through which both biological processes and the linguistic environment are deemed to influence language acquisition (Shaffer, 1985). According to interactionist theory, the maturation of brain structure and functions leads to cognitive and language development. Consequently, children of the same chronological age should show similar speech patterns. In terms of environmental influences, as cognitive development matures and the child produces more sophisticated utterances, caregivers tend to increase the complexity of their communication and interaction with the child, thus facilitating greater language development (Piaget, 1970). The interactionist theory sets the stage for multifaceted methods of language assessment.

Communicative Competence Theory

Another major contribution made by Noam Chomsky (1965) was developing the distinction between linguistic competence and linguistic performance. Competence is based on knowledge of a language, whereas performance is actual speech production and expression. An individual's linguistic performance may or may not represent linguistic competence. That is, at times individuals make errors such as twists of the tongue that may not represent their language competence. In response to Chomsky's distinction between competence and performance, Hymes (1966) coined the term *communicative competence.*

Within the field of linguistics, the goal of teaching language is to reach communicative competence (Canale & Swain, 1980). In 1980, Canale and Swain defined *communicative competence* as consisting of three components: grammatical competence, sociolinguistic competence, and strategic competence. *Grammatical competence* is knowledge of the rules of phonology, morphology, lexical semantics, and syntax, which are also known as *language structural models.* At the phonological level of language, an individual is required to know the phonemes or the most basic distinct unit within a sound system or language (Castles & Coltheart, 2004). The English language has approximately 40 phonemes, which are used to differentiate among sounds. A *morpheme* is the most basic unit in language that has meaning. For example, the word *apples* consists of two morphemes: *apple,* which is considered a free morpheme, and *s,* which means "plural or many." *Lexical semantics* is the meaning of words and syntax in a string of words grammatically arranged to convey a thought.

Functionally, individuals use language as an effective means of communicating with others and understanding the pragmatics or the appropriate and effective social use of language (Pennington, 1991; Russell & Grizzle, 2008). Deficits in pragmatic language have been linked to autism spectrum disorders and internalizing and externalizing disorders (Russell & Grizzle, 2008). Canale and Swain (1980) referred to the functional use of language as *sociolinguistic competence.* Sociolinguistic competence is the capacity to use and interpret linguistic social meaning within a suitable communication situation. They divided sociolinguistic competence into two components: sociocultural competence and discourse competence. *Sociocultural competence* is

defined as the ability to use suitable social and cultural rules of speaking (Cohen & Olshtain, 1981). *Discourse competence* is having an understanding of the rules of cohesion and coherence. *Strategic competence* (Canale & Swain, 1980) is the communication strategies, which can be verbal or nonverbal, that are used when a communication breakdown occurs.

In addition to the theories and theoretical frameworks that have evolved regarding the definition of language and communicative competence, numerous studies have attempted to elucidate early skills in infancy and in young children that may have an impact on later language competence. Research has suggested that the early emergence of interpersonal skills plays a pivotal role in communication development. Typically developing infants and young children are able to follow another individual's gaze and to alternate gaze from the object to the person and point, if necessary, to check that the other person shares their focus of interest (Chiat & Roy, 2008; Trevarthen & Aitken, 2001). This ability is commonly referred to as *shared joint attention,* which has been noted to be important to language development (Baldwin, 1995). Beyond the ability to share joint attention and being responsive to verbal and nonverbal expressions, infants and young children need to recognize the purpose or the symbolic representation of the exchange (Chiat & Roy, 2008; DeLoache, 2004). Deficits in symbolic play as well as in joint attention (also referred to as *social cognition*) have been found to predict autism (Charman et al., 2005; Toth, Munson, Meltzoff, & Dawson, 2006). Chiat and Roy (2008) found that early social cognition was the strongest predictor of social communication, and early phonological processing skills were a strong predictor of later morphology and syntax. Communicative competence theory has had a strong influence on the way in which language is understood and assessed through current methods of assessment.

ANATOMY OF LANGUAGE

Before reviewing and summarizing the various areas included in testing language competence, it is important to briefly review the anatomy and neuroanatomy of language, language processing, and speech production. This brief overview does not cover all brain regions, subcortical areas, or sensory and motor pathways involved in language. Many texts provide inclusive detailed descriptions of the processes involved in language. The two texts referenced in this section provide more detailed information.

Regarding the hemispheres of the brain, the left hemisphere is dominant for language processing in approximately 95% of right-handed individuals and approximately 65% to 70% of left-handed individuals. The brain structure that connects both hemispheres, the corpus callosum, allows for processing connections between both hemispheres, enabling the nondominant hemisphere of the brain to also participate in the processing of language (Blumenfeld, 2002). A few of the core anatomical areas of the brain in which language is localized include Broca's area, which is roughly located within the left, lateral frontal cortex bordering the primary motor cortex. Broca's area is primarily the motor speech region of the brain. Wernicke's area, which is roughly located in the temporal lobe, is primarily responsible for the encoding and decoding of auditory–linguistic information. This latter area is in charge of understanding word meaning.

The anatomy of language processing, such as when one hears and repeats a word, is generally initiated by the auditory information reaching the primary auditory cortex or Heschl's gyrus; the information then travels to the adjacent association cortex and to Wernicke's area, where meaning is derived. Subsequently, information is transferred to Broca's area in the primary motor cortex through the sylvain fissure. This transfer is followed by the activation of sound sequences, which then allows an individual to produce words (Blumenfeld, 2002).

Regarding the production of sound, vocal folds or vocal cords located in the larynx oscillate when air pushes through them. Speech sound is then generated by formants, or three restricted frequency ranges that are specific to each vowel. Speech sound then passes through the pharyngeal, oral, and nasal cavities and then out through the mouth, where the tongue, teeth, and lips produce differential sounds.

In humans, the descent of the larynx has led to a broader range of formants (Kolb & Whishaw, 2005).

TESTING FOR LANGUAGE COMPETENCE

The assessment of language competence should be a comprehensive process, incorporating as much relevant information or data as possible. Assessment methods should be tailored to the client considering age, ethnocultural background, and specific referral questions. A multimodal approach should be followed based on information obtained from interviews, observations in various settings, and formal and informal measures (Shipley & McAfee, 2009). Formal measures are standardized or norm-referenced tests. Informal measures are also an important element in testing for language competence on the basis of receptive and expressive processes and include methods such as language sampling, checklists, curriculum-based information, and criterion-referenced and portfolio assessment. Shipley and McAfee (2009) have provided numerous resources such as forms and worksheets that can assist in informal language assessments. For example, they included a checklist for conducting an informal assessment of language, worksheets for recording a language sample, and guidelines as well as forms for assessing language development of children for parents and practitioners. It is beyond the scope of this chapter to review the process of conducting a curriculum-based assessment or how to complete portfolios in the different areas assessed. Hosp, Hosp, and Howell (2007) and Jones (2008) are two resources for completing a curriculum-based language assessment that includes guidelines and fidelity checklists.

It is also beyond the scope of this chapter to review the vast array of formal measures available that assess the different components of language. Although a review of every language competency measure is not warranted here, a few of the more commonly used standardized measures and subtests are reviewed. When selecting formal measures, practitioners need to consider the theoretical basis used to develop the measure, the psychometric properties, a literature review of studies evaluating the measure, the norming standards used, the examinee's ethnocultural background, and the skills that the measure purports to assess. The first few measures reviewed in this chapter include this information to provide the reader with a model of the pertinent information needed when reviewing a test. Moreover, the interested reader is also directed to peruse the following resources. The Buros *Mental Measurements Yearbook* is a resource that practitioners can use to obtain detailed reviews on specific tests, as is Nicolosi, Harryman, and Kreschek (2004, pp. 376–388), which includes a brief description of approximately 114 language tests.

The major areas of language assessment reviewed are phonological processing; receptive and expressive concepts, vocabulary, and language; pragmatics and communication competencies; language development competencies related to reading and writing; second language acquisition (conversational and academic proficiency); and spoken language. The assessment of sign language abilities in students who are deaf or hard of hearing and of speech and language disorders and language competencies in children with neurodevelopmental disorders such as autism spectrum disorder are also reviewed.

Phonological Processing and Preliteracy Skills

Phonological processing involves phonological awareness, phonological memory, and rapid retrieval of phonemes within a language (Preston & Edwards, 2007). *Phonological awareness* refers to the knowledge of the basic sound systems in a language, and *phonological memory* is the ability to retain and immediately recall phonemes (Preston & Edwards, 2007). To successfully and accurately perform phonological processing tasks, an individual has to be able to construct and retrieve symbols that represent phonemes and phoneme combinations (Preston & Edwards, 2007; Sutherland & Gillon, 2005; Swan & Goswami, 1997).

Overwhelming evidence has linked phonological processing skills and reading acquisition (M. J. Adams, 1990; Lundberg, Frost, & Petersen, 1988; Vellutino et al., 1996; Wagner & Torgesen, 1987). Proficiency in phonological skills has also been found to be indicative of preliteracy skills. That is, scores on phonological processing measures can be

predictive of reading success and word-level acquisition, and conversely, limited phonological processing skills are predictive of reading difficulties and a limited understanding of syntax (M. J. Adams, 1990; Ball & Blackman, 1991; Chiat, 2001). Furthermore, phonological processing skills have also been linked to spelling acquisition. Therefore, the assessment of phonological processing is fundamental to better understanding a child's potential regarding preliteracy and written language skills such as spelling (Bird, Bishop, & Freeman, 1995; Larrivee & Catts, 1999; Lewis, Freebairn, & Taylor, 2000, 2002).

The Comprehensive Test of Phonological Processing (Wagner, Torgesen, & Rashotte, 1999) is a widely used individually administered test of phonological processing. It is a norm-referenced test developed with a research-based process and a strong standardization program (Hintze, Ryan, & Stoner, 2003). The Comprehensive Test of Phonological Processing norming sample was representative of the population characteristics (i.e., gender, race, ethnicity, parent education, family income) reported by the U.S. Census Bureau in 1997. Item response theory modeling and item analysis were used in the validation and content development process (Hintze et al., 2003). Internal consistency reliability of the composite scores ranged from .83 (Phonological Memory) to .96 (Phonological Awareness), and test–retest reliability ranged from .70 for Rapid Naming to .94 for Alternate Rapid Naming (Wagner et al., 1999). Using confirmatory content analysis, the construct identification validity was established, suggesting three distinct and correlated phonological processing abilities: phonological awareness, phonological memory, and rapid naming. Thus, the Comprehensive Test of Phonological Processing has been established as a reliable and valid measure of phonological processing (Hintze et al., 2003).

Phonological awareness is an individual's knowledge of and access to his or her oral language sound structure (Mattingly, 1972). *Phonological memory* is an individual's ability to phonologically code information and temporarily store it in working memory or, more specifically, the phonological loop (Hintze et al., 2003). The phonological loop consists of the phonological store and the articulatory control process. This system briefly stores verbatim auditory information (Baddeley, 1992; Torgesen, 1996). Weakness in phonological memory has been associated with difficulty learning new spoken and written words (Gathercole & Baddeley, 1990; Gathercole, Willis, & Baddeley, 1991). Last, rapid naming is the efficient and quick retrieval of phonological information from long-term memory using visual information. Individuals with difficulty in reading fluency are expected to perform weakly on rapid naming subtests (Hintze et al., 2003).

The Comprehensive Test of Phonological Processing accommodates the wide age span from ages 5 to 24 and can be used to identify individuals who are functioning below age-level expectation in important phonological processes, to differentiate the areas of strength and weaknesses in the various areas of phonological processing, to document progress for individuals receiving intervention, and for research purposes.

Numerous measures focus primarily on assessing phonological awareness. Some of these measures include the Phonological Awareness Test—2 (Robertson & Salter, 2007); the Test of Phonological Awareness—Second Edition: Plus (Torgesen & Bryant, 2004); and the Dynamic Indicators of Basic Early Literacy Skills, Sixth Edition (Good & Kaminski, 2002). The latter is a widely used standardized measure that is individually administered and assesses alphabet understanding and phonological awareness (Kaminski & Good, 1996). It is primarily used for progress monitoring and in screening prereading skills of children from preschool age to third grade (Hintze et al., 2003). The Dynamic Indicators of Basic Early Literacy Skills, Sixth Edition was initially conceptualized as a downward extension of curriculum-based measures (Elliott, Lee, & Tollerson, 2001). In terms of validity, it has strong concurrent validity with the phonological awareness and phonological memory composites of the Comprehensive Test of Phonological Processing (Hintze et al., 2003).

A preliteracy measure that incorporates input from teachers and parents can be found in the Clinical Evaluation of Language Fundamentals—Preschool—Second Edition (CELF–Preschool–II; Semel, Wiig, & Secord, 2004). The Pre-Literacy Rating Scale is a checklist completed by the child's

parents, teacher, or both that provides further supplemental information regarding the skills that influence the development of reading and writing. It is intended for preschool- and kindergarten-aged children. Different aspects of the CELF–Preschool–II are discussed throughout this chapter in more detail.

Receptive Language

Measures of receptive language assess an individual's ability to understand the spoken and written word. Receptive language skills can be assessed at the sound level, the word level, the sentence level, and the narrative level. Sound Blending and Incomplete Words subtests of the Woodcock–Johnson III Tests of Cognitive Abilities—Normative Update (Woodcock, McGrew, & Mather, 2007) and the Spelling of Sounds subtest of the Woodcock–Johnson Tests of Achievement—Third Edition—Normative Update (Woodcock et al., 2007) are some measures of receptive language skills at the sound level. Sound Blending requires the individual to synthesize speech sounds to form a word, and Incomplete Words requires that an individual identify a complete word after being presented with a word with missing phonemes. Spelling of Sounds assesses an individual's ability to identify and spell words applying phonological knowledge.

At the word level, there is a difference between measures of graded random vocabulary such as the Peabody Picture Vocabulary Test, Fourth Edition (Dunn & Dunn, 2007), or the Receptive One-Word Picture Vocabulary Test (Brownell, 2000) and a functional preacademic vocabulary–language assessment such as the Bracken Basic Concept Scale—Third Edition: Receptive (Bracken, 2007b). Measures of basic concepts have been found to be important in establishing language competence because these measures reflect the vocabulary needed to function in a classroom setting (Bracken & Crawford, 2010). Moreover, the acquisition of basic concepts has been strongly correlated with vocabulary development and language development (Pecnyna Rhyner & Bracken, 1988; Zucker & Riordan, 1988). Moreover, preschool-age children with strong conceptual development were better with meaning making when presented with a novel word (Booth & Waxman, 2002).

The categories evaluated with the Bracken Basic Concept Scale—Third Edition: Receptive include colors, letters, numbers–counting, size–comparisons, shapes, directions–position, self- and social awareness, texture–material, quantity, and time–sequence. These categories are noted as a combination of comprehensive categories that are considered foundational knowledge for students to be able to communicate in school and learn in all of the required domains (Bracken & Crawford, 2010).

Regarding receptive vocabulary, a widely used measure of general vocabulary terms is the Peabody Picture Vocabulary Test, Fourth Edition, which is a norm-referenced, untimed, individually administered instrument that includes 228 items, in addition to training items. It is intended for use with people as young as age 2 years, 6 months, through age 90. The examinee is presented with four pictures and is required to identify the picture that corresponds to the stimulus word stated by the examiner. A similar test of receptive vocabulary is the Receptive One-Word Picture Vocabulary Test. It is also a norm-referenced test designed for individuals ages 2 to 18 years, 11 months, and it follows a similar administration format as the Peabody Picture Vocabulary Test, Fourth Edition. It was conormed with the Expressive One-Word Picture Vocabulary Test (Brownell, 2000) for meaningful comparisons that are discussed later.

Two of the most comprehensive validated and reliable language measures used widely are the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF–IV; Semel, Wiig, & Secord, 2003) and the CELF–Preschool–II. The preschool version is intended for use for children ages 3 to 6, and the CELF–IV used with individuals ages 5 to 21. The most recent version of the CELF–IV was linked to educational curriculum, which is convenient for teachers and clinicians. For the purpose of maintaining the organization of this chapter, the different subtests of the CELF will be discussed throughout the major language areas discussed.

Both versions of the CELF (IV and Preschool–II) have a subtest, Concepts and Following Directions, that measures receptive language at the sentence level. The examinee is required to point to the correct pictured objects in response to oral directions. For

the younger children, familiarization items are presented before the administration of the subtest. On the CELF–Preschool–II, the Sentence Structure subtest is used to evaluate the child's ability to interpret sentences of increasing length. In this subtest, the child is asked to point to the picture that illustrates a given sentence. For older children, the CELF–IV has a Semantic Relationships subtest, which also assesses receptive language comprehension skills at the sentence level. In this subtest, the student listens to a sentence and is then required to select two choices that answer the target question.

A measure of receptive language at the word and sentence level is the Test of Auditory Comprehension of Language, Third Edition (Carrow-Woolfok, 1998). It is a normed referenced measure based on language comprehension theory and is designed for children ages 4 to 13 years, 11 months. This measure assesses vocabulary, grammatical morphemes (using the context of a simple sentence), and elaborated phrases and sentences (measures the understanding of syntactically based word relations and sentence construction).

Regarding assessment of receptive comprehension of language at the narrative level, the CELF–IV's has a subtest, Understanding Spoken Paragraphs, in which the student is required to listen to spoken paragraphs and then has to respond to questions that target the main idea, details, sequence, and inferential and predictive information. Another measure that includes subtests that assess narrative comprehension is the Test of Narrative Language (Gillam & Pearson, 2004). This test is intended for children ages 5 to 11 years, 11 months; it is norm referenced and individually administered. In a test review of this measure, Hayward, Stewart, Phillips, Norris, and Lovell (2008), found that although the Test of Narrative Language has strong psychometric evidence, it is not a definitive measure of language skills. It measures only narrative skills and is best used as a supplemental assessment.

Regarding receptive narrative comprehension, the Test of Language Competence—Expanded Edition (Wiig & Secord, 1989) has a narrative comprehension subtest presented in three different formats: no picture, five sequenced pictures, and single picture. In the no-picture format, the examiner reads a story to the child who then has to answer comprehension questions. In the five-sequenced-pictures format, the examiner presents five sequenced pictures and reads a relevant story to the child. The child is then asked comprehension questions about the characters, events, and consequences in the story. In the single-picture format, the examiner reads a story while the child is presented with a single picture, and the child then answers questions.

Expressive Language

Measures of expressive language assess an individual's ability to produce language through the spoken word or in a written format. Expressive language measures add a further task requirement of word retrieval from memory (Brownell, 2000). As with receptive language skills, expressive language skills can be assessed at the sound level, word level, sentence level, and narrative level. The Oromotor Sequencing subtest of the Neuropsychological Assessment, Second Edition (Korkman, Kirt, & Kemp, 2007) measures expressive language at the sound level. The domains covered in the Neuropsychological Assessment are theoretically based, and the clinician may select subtests to administer dependent on the clinical or individual needs of the child identified either during the assessment process or at the time of referral. The second edition of the Neuropsychological Assessment is designed for use with children ages 3 to 16. Within the theoretically based language domain, the Oromotor Sequencing subtest requires the child to repeat articulatory sequences of sounds and tongue twisters. This subtest primarily measures oromotor coordination that underlies the sequential production of speech.

Another subtest of the Neuropsychological Assessment, Speeded Naming, assesses the child's ability to rapidly access semantic information and produce familiar words to identify numbers, shapes, size, and letters in alternating patterns. This measure also assesses the automaticity of expressive language. On this measure, the child has to initially rapidly name a series of letters and numbers that are randomly presented. As the test progresses, the child is then required to name shapes, size, and color of the visually presented stimuli.

Commonly used measures of receptive language at the word level include the Expressive Vocabulary Test, Second Edition (Williams, 2007) and Expressive One-Word Picture Vocabulary Test. The Expressive One-Word Picture Vocabulary Test is a norm-referenced individually administered measure designed for individuals ages 2 to 18 years, 11 months. Scores on the Expressive One-Word Picture Vocabulary Test can be compared with scores on the Receptive One-Word Picture Vocabulary Test described in the previous section because there is equivalence in the norms for both measures. The Expressive One-Word Picture Vocabulary Test consists of a set of 170 pictures representing objects, actions, or concepts, which the student is required to name.

The Expressive Vocabulary Test, Second Edition, is a norm-referenced test that used the same norm sample as the Peabody Picture Vocabulary Test, Fourth Edition, such that direct comparisons between these two measures can also be made. That is, a comparison between the Expressive Vocabulary Test and the Peabody Picture Vocabulary Test is a way of screening for aphasia (difficulty understanding or producing spoken or written language) or other expressive language impairments. Moreover, this comparison may assist in determining areas of strengths and weaknesses. For example, if an individual scores significantly higher in receptive than expressive language, the difference in scores can suggest a word-finding or word-retrieval difficulty. The Expressive Vocabulary Test, Second Edition, consists of 190 items and is designed for individuals who are ages 2 to 90. The examiner presents a picture and reads the stimulus question, and the examinee must then provide the label for the picture or provide a synonym for a word that applies to the pictured context. The recent edition of this measure provides updated illustrations that are sensitive to the racial and ethnic differences found in the United States.

Similarly, the Bracken Basic Concept Scale—Third Edition: Receptive and Bracken Basic Concept Scale—Third Edition: Expressive (Bracken, 2007a) were designed to allow differentiation of receptive concept development and expressive development. Testing formats used the same normative sample to contrast receptive and expressive abilities, which is relatively rare, especially when they yield empirical data on whether the differences are statistically significant and provide information on the proportion of the population that has a difference of any given score magnitude.

At the sentence level, expressive language can be assessed by subtests such as the CELF–IV's Formulated Sentences and Recalling Sentences. It is important to note that other measures include repeating or recalling sentences; however, because of the breadth of this chapter, the aforementioned measures are the ones reviewed in this chapter. In Recalling Sentences, the examiner orally presents the student with a sentence, and the student is then asked to imitate what the examiner has stated. In Formulating Sentences, the child is presented with visual stimuli using a targeted word or phase, and he or she has to create a sentence.

In terms of assessing expressive narrative language, the Test of Language Competence—Expanded Edition, previously reviewed, has an oral expressive component in the same three formats: no picture, five sequenced pictures, and single picture. The Oral Narration subtest with the no-picture format is administered right after the Narrative Comprehension subtest (no-picture format). The child is required to retell the same story in manner in which it was orally presented by the examiner. In the Oral Narration subtest, five-sequenced-pictures format, the child is presented with five pictures and has to verbally create a story that corresponds to the sequence of pictures. Last, in the single-picture format, the child is required to create a story that is relevant to the picture.

Language Development Competencies Related to Reading and Writing

What are the essential early language competencies related to reading and writing? Within the literature, a vast number of investigations have linked areas of early language competencies to reading; however, fewer studies have investigated the early competencies necessary for writing. Regarding early reading skills, phonological processing, which was previously reviewed, has been linked to reading skills. The reader is referred to the Phonological Processing and Preliteracy section of this chapter for discussion of

the measures that correspond to these areas. Regarding reading fluency, recent studies have suggested that the retrieval of phonological information from long-term memory is an important factor explaining reading fluency (Barth, Catts, & Anthony, 2009). Moreover, McCutchen and Perfetti (1982) regarded access to phonemic information in memory as critical to higher level semantic and syntactic processes that can also assist with fluency. In addition to early reading acquisition, phonemic processing has also been implicated in reading comprehension because phonemic codes are active as the comprehension of sentences takes place (McCutchen, Dibble, & Blount, 1994).

In addition to phonological processing skills, knowledge of the basic concept of letters has also been identified as a future marker for reading development in English- and non-English-speaking children (M. J. Adams, 1990; Lyytinen et al., 2004; Muter & Diethelm, 2001). That is, the child's knowledge in early elementary school years of basic concepts, such as naming upper- and lowercase letters, sound–symbol association, sentence recall, and naming speed have been linked to later reading skills (Denton & West, 2002; Hooper, Roberts, Nelson, Zeisel, & Fannin, 2010; Lonigan, Burgess, & Anthony, 2000; West, Denton, & Germino-Hauskin, 2000). The previously reviewed Bracken Basic Concept Scale—Third Edition: Receptive and Bracken Basic Concept Scale—Third Edition: Expressive are comprehensive measures of receptive and expressive basic concept acquisition, respectively, and the CELF–Preschool–II also includes a brief measure of basic concept skills.

Although evidence has supported reading, writing, and spelling as integrated processes (cross-sectional, instructional, and longitudinal studies; Bear, Ivernizzi, Templeton, & Johnston, 2004; Moats, 2000), they are generally considered separate functional systems (Berninger & Richards, 2002; Hooper et al., 2010). It is important, however, to note that in the early preliteracy stages of literacy development, some overlap exists in the skills required for reading and writing (Snow, Burns, & Griffin, 1998). For example, in the area of writing, phonemic processing skills have been linked to spelling and later writing skills. Additional predictors of writing skills include knowledge of the basic concept of upper- and lowercase letter names, letter writing along with writing first names, and writing dictated and copied letters and numbers (Molfese, Beswick, Molnar, & Jacobi-Vessels, 2006). In a recent study, Hooper et al. (2010) found that core language abilities as well as prereading skills assessed just before kindergarten are predictive of written language skills in Grades 3 to 5. The CELF–4 and CELF–II–Preschool also assess core language abilities.

One of the measures that incorporates writing first and last name as well as copying letters is the Oral and Written Language Scales (Carrow-Woolfolk, 1996). It is a theoretically based, individually administered measure of listening comprehension and oral and written expression. This measure is intended for children ages 3 to 21.

Spoken Language Competencies

A competent speaker of a language has semantic knowledge, phonological knowledge, and grammatical knowledge of a particular language (Byrnes & Wasik, 2009; Hoff, 2001). For there to be spoken language competence, Byrnes and Wasik, (2009) noted that in addition to these three language abilities, there also has to be a connection to social competence. They highlighted five additional aspects of spoken language competence that include social competence as a desire to communicate with others, reciprocity and turn taking, a desire to get along with others, respect for others, and a lack of egocentrism (Ninio & Snow, 1999).

One measure of spoken language competence is the Comprehensive Assessment of Spoken Language (Carrow-Woolfolk, 1999). This measure is grounded in theory; it is an individually administered, norm-referenced measure designed for individuals ages 3 through 21. No reading or writing is required, and the examinee either responds verbally or by pointing. This measure includes 15 subtests that assess four language structures categories: Lexical/Semantic (word knowledge and use of words or phases), Syntactic (knowledge and use of grammar), Supra-Linguistic (understanding of language in which the meaning is not directly available such as indirect request and sarcasm), and, as previously

noted, Pragmatics (awareness of effective and appropriate language in situational context). Moreover, this measure provides descriptive analysis worksheets that enable practitioners to target specific skills areas.

Another measure of spoken language is the Test of Language Development—Primary, Fourth Edition, and Test of Language Development—Intermediate, Fourth Edition (Hammill & Newcomer, 2008). Both are norm-referenced measures that are individually administered. The primary version was designed for young children ages 4 to 8 years, 11 months, and the intermediate version is intended for individuals between the ages of 8 through 17 years, 11 months. The Test of Language Development—Primary consists of subtests that measure semantics and syntax, listening, organizing, speaking, and overall language ability. The Test of Language Development—Intermediate includes subtests that measure semantics and grammar, listening abilities, organizing abilities, and speaking abilities. As previously noted, there is an important social component to measuring spoken language competence. Measures of social competence or pragmatics would be a good supplement to the Test of Language Development—Primary and the Test of Language Development—Intermediate.

Pragmatics or Social Communication Competencies

In addition to the structural components that contribute to the emergence of language, there is also an important social component. For example, young children have been found to learn the names of objects to which adults are explicitly attending (Baldwin, 1991; MacWhinney, 1998). Furthermore, early word learning has been strongly associated with the important role of the mutual gaze between the mother and child (Akhtar, Carpenter, & Tomasello, 1996; Tomasello & Akhtar, 1995). Therefore, the early emergence of interpersonal skills plays a pivotal role in communication development. Specifically, the effective and appropriate use of language in a social context requires pragmatic language competencies (Russell & Grizzle, 2008). Deficits in this area of language development have been linked to child and adolescent disorders such as autism spectrum disorder (Lord, 1993; Mawhood, Howlin, & Rutter, 2000; Rapin, 1996; Russell & Grizzle, 2008) and to a lesser extent to attention-deficit/hyperactivity disorder (Bishop & Baird, 2001; Russell & Grizzle, 2008) as well as internalizing and externalizing disorders (Im-Bolter & Cohen, 2007).

The assessment of pragmatic language has increasingly been incorporated into language measures (C. Adams, 2002). Russell and Grizzle (2008) conducted an extensive review of the various pragmatic language assessments such as structured participant observations, diagnostic measures, behavioral checklists, and questionnaires and made recommendations on test selection. The Teacher Assessment of Student Communicative Competence (Smith, McCauley, & Guitar, 2000), the Children's Communication Checklist—2 (Bishop, 2006), the Pragmatic Profile, and the Observational Rating Scale of the CELF–IV ranked highest in terms of strong content validity in the group of observation instruments and questionnaires. Of these, the Children's Communication Checklist and the Pragmatic Profile are considered good screening instruments when used along with structured observation data. These instruments also allow for multiple informants to achieve a more ecologically valid measure of pragmatic language. Another rating scale of pragmatics is provided by the CELF–IV and the CELF–Preschool–2. These measures incorporate a supplemental test, the Pragmatic Profile, on which the examiner can use information provided from the teacher, the parents, or both regarding the child's social language.

More narrowly targeted measures of pragmatic language include the Comprehensive Assessment of Spoken Language, specifically, the pragmatic judgment and nonliteral language subtests. Last, a comprehensive, norm-referenced measure of pragmatic language is the Test of Pragmatic Language, Second Edition (Phelps-Terasaki & Phelps-Gunn, 2007). It includes content based on Norris and Hoffman's (1993) situational–discourse–semantic language theory. This theoretical model includes situational context, discourse context, and semantic context. Phelps-Terasaki and Phelps-Gunn (2007) adapted the situational–discourse–semantic model to focus on pragmatic language. This measure was designed for use with children ages 6 to 18, and it has seven

core subcomponents that provide information on such pragmatic issues as physical context, audience, topic, purpose, visual–gestural cues, abstractions, and pragmatic evaluation. The examinee is presented with a visual illustration along with a verbal prompt and asked to create a verbal response to the dilemma presented.

Second Language Acquisition Competencies: Conversational–Academic Proficiency

A second language acquisition process has four stages: preproduction, early production, speech emergence, and intermediate fluency (Hearne, 2000; Ortiz & Kushner, 1997; Rhodes, Ochoa, & Ortiz, 2005; Roseberry-McKibblin, 2002). Briefly, the first, or preproduction, stage generally consists of the child's first 3 months of exposure to the new language and is characterized as the silent period in which an individual is focusing on comprehension, generally giving yes-or-no responses in English or one-word responses. During the second, or early production, stage, the individual focuses more on comprehension and uses one- to three-word phrases and responses; this period lasts between 3 and 5 months beyond preproduction. The third, or speech-emergent, stage generally extends beyond the second stage for another 6 months to 2 years. During this period, increased comprehension and expanded vocabulary occur, and oral responses can include recalling and retelling as well as comparing and sequencing. Stage 4, or the intermediate fluency stage, lasts from 2 to 3 years beyond Stage 3. During this time, improved comprehension, more extensive vocabulary, and fewer grammatical errors occur. Oral responses during this latter stage include predicting, giving opinions, and summarizing (please refer to Hearne, 2000, and Rosenberry-McKibblin, 2002, for more in-depth detailed descriptions of these stages). In addition to reviewing the stages of the second-language acquisition process, it is also important to differentiate between interpersonal and academic language proficiencies.

Language proficiency is an individual's performance in understanding and using language in formal and informal social and academic settings. There are two types of language proficiencies: basic interpersonal communication skills and CALP (Cummins, 1984). Basic interpersonal communicative skills is basic language proficiency that generally takes 2 or 3 years to acquire, and it is used in informal social settings. CALP is generally attained within 5 to 7 years, and this category of language proficiency is required to perform well in school (Cummins, 1984). Evidence in the literature has supported that the amount of time it takes an individual to acquire a second language is affected by the level of CALP in the first or native language (Thomas & Collier, 1997). Furthermore, Thomas and Collier (1997) found that in general the more schooling a child had in the first language, the higher the second language achievement.

When assessing a student whose primary language is not English, it is extremely important to use both formal and informal methods (Rhodes et al., 2005). It is important that the evaluator obtain the student's basic interpersonal communicative skills and CALP through the school's second language department (Rhodes et al., 2005). If the proficiency levels are dated or do not exist, the evaluator can, with knowledge of the second language, assess these levels. One formal measure that is widely used, exists in English and Spanish, and assesses the CALP level is the Woodcock–Muñoz Language Survey—Revised (Woodcock, Muñoz-Sandoval, Ruef, Alvarado, & Ruef, 2005), which measures CALP within six levels of competencies (i.e., listening, speaking, comprehension, reading, writing, and oral language). This measure has strong theoretical underpinnings and psychometric properties. In addition to the CALP, this measure provides information regarding oral language dominance, monitoring growth or change for both languages, readiness for English-only instruction, and determining eligibility for bilingual education services.

Informal methods of assessing language proficiency include but are not limited to observations in structured and unstructured settings; interviews with parent, teacher, and student; questionnaires; teacher rating scales; and language samples. Observations of the child's social and academic language in structured and unstructured settings provide critical information regarding how the child communicates (Lopez, 1997). The evaluator can

focus on receptive language skills such as understanding teacher directions, the vocabulary being used, class discussion, and so on. The evaluator can focus on expressive skills by observing appropriate use of vocabulary, describing events, appropriate use of tense, and so forth (Lopez, 1997).

Regarding interviews with parents, teachers, and students, Rhodes et al. (2005) developed interview forms that can be replicated from their book *Assessing Culturally and Linguistically Diverse Students: A Practical Guide*. Further, Rhodes et al. (2005) also provided information regarding specific questionnaires and rating scales. Moreover, the collection of language samples should also be recorded as part of the assessment process and can be done by means of a conversation between the examiner and examinee or the examinee and teacher. Pragmatic and structural features can then be analyzed once recorded (Mattes & Omark, 1991). Another interview resource for estimating language use with various interactors is in Mattes and Omark's (1991) *Speech and Language Assessment for the Bilingual Handicapped.*

Sign Language Abilities in Students Who Are Deaf or Hard of Hearing

Numerous distinct sign languages exist throughout the world. American Sign Language is considered a visual–gestural or visual–manual modality language that is the natural language of individuals who are deaf or hard of hearing in the United States (Miller, 2008). It is considered as complex as any speech-based language, with its own grammatical rules and other linguistic features (Miller, 2008).

Regarding the assessment of language and sign language for those who are deaf or hard of hearing, not only is the research limited, but so are the assessment tools. One published measure that is recommended for use with children who are deaf or hard of hearing is the Behavior Rating Instrument for Autistic and Other Atypical Children, Second Edition (Ruttenberg, Wenar, & Wolf-Schein, 1991), which is an observational rating instrument that consists of nine scales: Relationship to an Adult, Communication, Drive for Mastery, Vocalization and Expressive Speech, Sound and Speech Reception, Social Responsiveness, and Psychobiological Development. This measure also includes expressive gesture and sign language and receptive gesture and sign language. The Behavior Rating Instrument for Autistic and Other Atypical Children, Second Edition, was designed for children who are on the autism spectrum as well as low-functioning students who are deaf and blind. In an article regarding assessments of language and other competencies, Cawthon and Wurtz (2009) discussed the value of alternate assessments such as portfolios and checklists for school-aged children who are deaf or hard of hearing.

Speech and Language Disorders

Typical language relies on the connections between sensory information and symbolic associations, motor skills, memory, and syntactical patterns (Kolb & Whishaw, 2003). Speech is one of the modalities for expressing language (Bowen, 2009). Speech disorders affect oral motor output and thus involve difficulties with producing speech sounds or with voice quality. Speech disorders are considered to be a type of communication disorder in which typical speech is disrupted (Bowen, 2009).

One model of differentiating among speech disorders that has psycholinguistic underpinnings was proposed by Dodd (1995). He identified four subtypes of speech disorder: phonological delay, phonological disorder (consistent and inconsistent deviant), articulation disorder, and childhood apraxia of speech. Briefly, he described phonological delay as involving typical development in phonological rules and processes; however, these abilities or skills are characteristic of a chronologically younger child. *Consistent deviant phonological disorder* is when a child has impaired understanding of the phonological system with developmental errors and unusual processes. *Inconsistent deviant phonological disorder* is when the child presents delays and variability in speech production of the same words equal to or greater than 40% of the time; when the child cannot produce acceptable phonemes, it is differentiated as an articulation disorder. *Apraxia of speech* is an impaired ability to plan the oral movements required for speech resulting in errors of prosody and speech sound production (Bowen, 2009).

Disorders of language involve difficulty in the processing of linguistic information and can affect expression of language, the capacity to understand language, or both. Thus, a language disorder can be expressive, mixed receptive–expressive language, or a communication disorder such as a phonological disorder or stuttering (American Psychiatric Association, 2000). Moreover, language disorders can be developmental or acquired.

Developmental language disorders are usually evident from the initial stages of language development and can occur secondary to other disorders such as autism, cerebral palsy, and so forth. Acquired disorders of language generally result from a brain injury. Many of the previously described (formal and informal) test and subtests are used in assessing speech and language disorders. Diagnoses of these types of disorder are made by speech–language pathologists and therapists, and some are made by neuropsychologists. It is important for practitioners, other than the specialist trained to diagnose these types of disorders, to become familiar with the criteria used within either the school system or the diagnostic manual to more effectively screen and refer individuals for a comprehensive speech and language assessment. Both Shipley and McAfee (2009) and McCauley (2001) are helpful resources for gaining a better understanding of developing speech and language competencies.

Children With Autism and Other Neurodevelopmental Disorders

Developmental language acquisition delays are accounted for in children by a range of neurodevelopmental disorders such as autistic spectrum disorder (ASD), which includes pervasive developmental disorder, not otherwise specified; autism; and Asperger syndrome (American Psychiatric Association, 2000). Additionally, neurodevelopmental disorders that affect language acquisition include Down syndrome, Prader-Willi syndrome, Williams syndrome, and Fragile X syndrome. With the exception of Asperger syndrome, delays in language are one of the first features recognized in ASD (Luyster & Lord, 2009). Pragmatic language competency deficits have been noted to be most symptomatic of children with ASD (Lord, 1993; Mawhood et al., 2000; Rapin, 1996; Russell & Grizzle, 2008). The trajectory of language competence varies for individuals with ASD. For example, in a longitudinal study, Lord, Risi and Pickels (2004) found that approximately half of a sample of 1,000 individuals on the spectrum were initially classified as language impaired; 40% were verbally fluent and 45% had functional, but not completely intact, language.

The formal diagnosis of autism includes as one criterion a qualitative impairment in communication. The Autism Diagnostic Observation Schedule (Lord, Rutter, DiLavore, & Risi, 2008) is a widely used, individually administered, semistructured standardized measure of communication, social interaction, and play designed for use with individuals who may have an ASD. Lord et al. (2008) noted that expressive language level is most likely the strongest predictor of outcome in ASD. This measure consists of four modules that range from one for individuals not having any speech to one for those who are verbally fluent.

Within each module, there are numerous activities in which certain behaviors of interest in the ASD diagnosis are likely to appear. The instrument's psychometric properties, such as the normative sample and validity and reliability studies, are extensively reviewed in the manual. In addition to this measure, other rating scales are included to assist in the diagnosis of autism, which are intended for parents or caregivers and teachers. A widely used rating scale designed to distinguish individuals with autism from those with developmental delays who do not have autism is the Childhood Autism Rating Scale, Second Edition (Schopler, Van Bourgondien, Wellman, & Love, 2010), which consists of a 15-item scale that provides information regarding general autistic behaviors.

Down syndrome and Prader-Willi syndrome are chromosomal disorders that result in secondary language disorders. Specifically, individuals with Prader-Willi syndrome have associated features of speech articulation deficits, and individuals with Down syndrome tend to have developmental challenges in receptive and expressive language and in articulation. That is, individuals with Down syndrome

tend to have good receptive vocabularies and generally present a desire to communicate; however, they tend to have difficulty with syntax and morphology (Rondal, 1995). Many measures that assess these areas have previously been reviewed in this chapter, and again, specialists in assessing language should be the ones conducting the comprehensive language evaluation. Regarding Fragile X syndrome, some children with this syndrome also present with some form of speech and language delay. An abnormality in the FMRI gene is the cause of Fragile X syndrome (Kolb & Whishaw, 2003). Because the speech and language development of individuals with Fragile X syndrome varies widely, broad generalizations cannot be made. However, it has been noted in the literature that girls with Fragile X syndrome tend to evidence fewer speech and language disorders than boys. The assessment of competencies for individuals with this disorder should also be conducted by a speech–language specialist.

CONCLUSION

Language is an important and broad area of assessment for a wide variety of populations. This chapter briefly covered a theoretical and applied multidimensional perspective on testing language competencies in children and adolescents. Theory and research are proposed to form the basis of more valid language assessments. Likewise, knowledge of language theory provides the basis for practitioners to use when selecting appropriate measures. Assessment of language competence should be multimodal, including formal and informal measures and multiple informants, and it should be conducted across a number of relevant settings.

Recognizing the need for assessing different domains and aspects of language, the authors reviewed a variety of widely used measures of structural and functional language competence as they applied to either literacy skills (receptive or expressive language, etc.) or pragmatics and social communication. Moreover, important factors to consider when assessing second language learners were included as a critical aspect of language assessment given the increase of children whose native language is not English or who are emerging bilinguals. Last, language areas important to special populations were reviewed with an acknowledgment of the need to develop more appropriate language assessment measures. In sum, future directions for testing language competence include ongoing research on language processes and the development of evidence-based measures and assessment procedures appropriate for learners with unique language characteristics and neglected aspects of language such as social emotional communicative skills.

References

Adams, C. (2002). Practitioner review: The assessment of language pragmatics. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 36*, 289–305. doi:10.1080/13682820110055161

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child Development, 67*, 635–645. doi:10.2307/1131837

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Baddeley, A. D. (1992). Working memory: The interface between memory and cognition. *Journal of Cognitive Neuroscience, 4*, 281–288. doi:10.1162/jocn.1992.4.3.281

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development, 62*, 875–890. doi:10.2307/1131140

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 131–158). Hillsdale, NJ: Erlbaum.

Ball, E. E., & Blackman, B. A. (1991). Does phonemic awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly, 26*, 49–66. doi:10.1598/RRQ.26.1.3

Barth, A. E., Catts, H. W., & Anthony, J. L. (2009). The component skills underlying reading fluency in adolescent readers: A latent variable analysis. *Reading and Writing, 22*, 567–590. doi:10.1007/s11145-008-9125-y

Bear, D. R., Ivernizzi, M., Templeton, S., & Johnston, F. (2004). *Words their way: Word study for phonics, vocabulary, and spelling instruction.* Upper Saddle River, NJ: Merrill/Prentice Hall.

Berninger, V. W., & Richards, T. L. (2002). *Brain literacy for educators and psychologists.* New York, NY: Academic Press.

Bird, J., Bishop, D. V. M., & Freeman, N. H. (1995). Phonological awareness and literacy development in children with expressive phonological impairments. *Journal of Speech and Hearing Research, 38*, 446–462.

Bishop, D. V. M. (2006). *Children's Communication Checklist—2*. San Antonio, TX: Harcourt Assessment.

Bishop, D. V. M., & Baird, G. (2001). Parent and teacher report of pragmatic aspects of communication: Use of the children's communication checklist in the clinical setting. *Developmental Medicine and Child Neurology, 43*, 809–818. doi:10.1017/S0012162201001475

Blumenfeld, H. (2002). *Neuroanatomy through clinical cases*. Sunderland, MA: Sinauer Associates.

Booth, A. E., & Waxman, S. R. (2002). Word learning is "smart": Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition, 84*, B11–B22. doi:10.1016/S0010-0277(02)00015-X

Bowen, C. (2009). *Children's speech sound disorders*. New York, NY: Wiley.

Bracken, B. A. (2007a). *Bracken Basic Concept Scale—Third Edition: Expressive*. San Antonio, TX: Pearson.

Bracken, B. A. (2007b). *Bracken Basic Concept Scale—Third Edition: Receptive*. San Antonio, TX: Pearson.

Bracken, B. A., & Crawford, E. (2010). Basic concepts in early childhood educational standards: A 50-state review. *Early Childhood Education Journal, 37*, 421–430. doi:10.1007/s10643-009-0363-7

Brownell, R. (2000). *Expressive One-Word Picture Vocabulary Test*. Los Angeles, CA: Western Psychological Services.

Byrnes, J. P., & Wasik, B. A. (2009). *Language and literacy: What educators need to know*. New York, NY: Guilford Press.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47. doi:10.1093/applin/1.1.1

Carrow-Woolfolk, E. (1996). *Oral and Written Language Scales*. San Antonio, TX: Pearson.

Carrow-Woolfolk, E. (1998). *Test of Auditory Comprehension of Language* (3rd ed.). San Antonio, TX: Pearson Assessments.

Carrow-Woolfolk, E. (1999). *Comprehensive Assessment of Spoken Language*. San Antonio, TX: Pearson Assessments.

Castles, A., & Coltheart, M. (2004). Is there a causal link from phonological awareness to success in learning to read? *Cognition, 91*, 77–111. doi:10.1016/S0010-0277(03)00164-1

Cawthon, S. W., & Wurtz, K. A. (2009). Alternate assessment use with students who are deaf or hard of hearing: An exploratory mixed-methods analysis of portfolio, checklists, and out-of-level test formats. *Journal of Deaf Studies and Deaf Education, 14*, 155–177. doi:10.1093/deafed/enn027

Charman, T., Taylor, E., Drew, A., Cockerill, H., Brown, J., & Baird, G. (2005). Outcome at 7 years of children diagnosed with autism at age 2: Predictive validity of assessments conducted at 2 and 3 years of age and pattern of symptom change over time. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 46*, 500–513. doi:10.1111/j.1469-7610.2004.00377.x

Chiat, S. (2001). Mapping theories of developmental language impairment: Premises, predictions and evidence. *Language and Cognitive Processes, 16*, 113–142. doi:10.1080/01690960042000012

Chiat, S., & Roy, P. (2008). Early phonological and sociocognitive skills as predictors of later language and social communication outcomes. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 49*, 635–645.

Chomsky, N. (1957). *Syntactical structures*. Hawthorne, NY: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Hawthorne, NY: Mouton.

Cohen, A. D., & Olshtain, E. (1981). Developing a measure of sociocultural competence: The case of apology. *Language Learning, 31*, 113–134. doi:10.1111/j.1467-1770.1981.tb01375.x

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. San Diego, CA: College Hill.

DeLoache, J. S. (2004). Becoming symbol-minded. *Trends in Cognitive Sciences, 8*, 66–70. doi:10.1016/j.tics.2003.12.004

Denton, K., & West, J. (2002). *Children's reading and mathematics achievement in kindergarten and first grade*. Washington, DC: National Center for Education Statistics.

Dodd, B. (1995). *Differential diagnosis and treatment of children with speech disorders*. London, England: Whurr.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). San Antonio, TX: Pearson.

Elliott, J., Lee, S. W., & Tollerson, N. (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Literacy Skills—Modified. *School Psychology Review, 30*, 33–49.

Gathercole, S. E., & Baddeley, A. D. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language, 29*, 336–360. doi:10.1016/0749-596X(90)90004-J

Gathercole, S. E., Willis, C., & Baddeley, A. D. (1991). Nonword repetition, phonological memory, and vocabulary: A reply to Snowling, Chiat, and Hulme. *Applied Psycholinguistics, 12*, 375–379. doi:10.1017/S0142716400009280

Gillam, R. B., & Pearson, N. A. (2004). *Test of Language Comprehension*. East Moline, IL: LinguiSystems.

Good, R. H., & Kaminski, R. A. (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Greene, J. (1973). *Psycholinguistics: Chomsky and psychology*. Middlesex, England: Penguin Education.

Hammill, D. D., & Newcomer, P. L. (2008). *Test of Language Development—Intermediate* (4th ed.). San Antonio, TX: Pearson.

Hayward, D. V., Stewart, G. E., Phillips, L. M., Norris, S. P., & Lovell, M. A. (2008). Test review: Test of Narrative Language (TNL). In D. V. Hayward, G. E. Stewart, L. M. Phillips, S. P. Norris, & M. A. Lovell, *Language, phonological awareness, and reading test directory* (p. 2). Edmonton, Alberta, Canada: Canadian Centre for Research on Literacy & Canadian Language and Literacy Research Network. Retrieved from http://www.ualberta.ca/~lphillip/documents/Introduction%20to%20the%20Test%20Reviews.pdf

Hearne, D. (2000). *Teaching second language learners with learning disabilities: Strategies for effective practice*. Oceanside, CA: Academic Communication Associates.

Hintze, J., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review, 32*, 541–556.

Hoff, E. (2001). *Language development* (2nd ed.). Belmont, CA: Wadsworth/Thomson Learning.

Hooper, S. R., Roberts, J. E., Nelson, L., Zeisel, S., & Fannin, D. (2010). Preschool predictors of narrative writing skills in elementary school children. *School Psychology Quarterly, 25*, 1–12. doi:10.1037/a0018329

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2007). *The ABCs to CBM: A practical guide to curriculum-based measurement*. New York, NY: Guilford Press.

Hymes, D. H. (1966). Two types of linguistic relativity. In W. Bright (Ed.), *Sociolinguistics* (pp. 114–158). The Hague, the Netherlands: Mouton.

Im-Bolter, N., & Cohen, N. J. (2007). Language impairment and psychiatric comorbidities. *Pediatric Clinics of North America, 54*, 525–542. doi:10.1016/j.pcl.2007.02.008

Jones, C. J. (2008). *Curriculum-based assessment: The easy way to determine response-to-intervention*. Springfield, IL: Charles C Thomas.

Kaminski, R. A., & Good, R. H., III. (1996). Towards a technology for assessing basic literacy skills. *School Psychology Review, 25*, 215–227.

Kolb, B., & Whishaw, I. Q. (2003). *Fundamentals of human neuropsychology* (5th ed.). New York, NY: Worth.

Korkman, M., Kirk, U., & Kemp, S. (2007). *Neuropsychological assessment* (2nd ed.). San Antonio, TX: Pearson.

Larrivee, L. S., & Catts, H. W. (1999). Early reading achievement in children with expressive phonological disorders. *American Journal of Speech-Language Pathology, 8*, 118–128.

Lewis, B. A., Freebairn, L. A., & Taylor, H. G. (2000). Follow up of children with early expressive phonology disorders. *Journal of Learning Disabilities, 33*, 433–444. doi:10.1177/002221940003300504

Lewis, B. A., Freebairn, L. A., & Taylor, H. G. (2002). Correlates of spelling abilities in children with early speech sound disorders. *Reading and Writing, 15*, 389–407. doi:10.1023/A:1015237202592

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. *Developmental Psychology, 36*, 596–613. doi:10.1037/0012-1649.36.5.596

Lopez, E. C. (1997). The cognitive assessment of limited English proficient bilingual children. In D. P. Flannagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 506–516). New York, NY: Guilford Press.

Lord, C. (1993). The complexity of social behaviour in autism. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from autism* (pp. 292–316). Oxford, England: Oxford University Press.

Lord, C., Risi, S., & Pickels, S. (2004). Trajectory of language development in autism spectrum disorders. In M. L. Rice & S. F. Warren (Eds.), *Developmental language disorders: From phenotypes to etiologies* (pp. 7–29). Mahwah, NJ: Erlbaum.

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (2008). *Autism Diagnostic Observation Schedule*. Los Angeles, CA: Western Psychological Services.

Lundberg, I., Frost, J., & Petersen, O. P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23*, 263–284. doi:10.1598/RRQ.23.3.1

Luyster, R., & Lord, C. (2009). Word learning in children with autism spectrum disorder. *Developmental Psychology, 45*, 1774–1786. doi:10.1037/a0016223

Lyytinen, H., Aro, M., Eklund, K., Erskine, J., Guttorm, T. K., Laakso, M. L., . . . Torppa, M. (2004). The development of children at familial risk for dyslexia: Birth to school age. *Annals of Dyslexia*, *54*, 184–220. doi:10.1007/s11881-004-0010-3

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology, 49*, 199–227. doi:10.1146/annurev.psych.49.1.199

Mattes, L. J., & Omark, D. (1991). *Speech and language assessment for the bilingual handicapped* (2nd ed.). Oceanside, CA: Academic Communication Associates.

Mattingly, I. G. (1972). Reading, the linguistic process, and linguistic awareness. In J. Kavanagh & I. Mattingly (Eds.), *Language by ear and by eye* (pp. 133–147). Cambridge, MA: MIT Press.

Mawhood, L., Howlin, P., & Rutter, M. (2000). Autism and developmental receptive language disorder—A comprehensive follow-up in early adult life. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 41*, 547–559. doi:10.1111/1469-7610.00642

McCauley, R. J. (2001). *Assessment of language disorders in children.* Mahwah, NJ: Erlbaum.

McCutchen, D., Dibble, E., & Blount, M. M. (1994). Phonemic effects in reading comprehension and text memory. *Applied Cognitive Psychology, 8*, 597–611. doi:10.1002/acp.2350080606

McCutchen, D., & Perfetti, C. A. (1982). The visual tongue-twister effect: Phonological activation in silent reading. *Journal of Verbal Learning and Verbal Behavior, 21*, 672–687. doi:10.1016/S0022-5371(82)90870-2

Miller, K. R. (2008). American sign language: Acceptance at the university level. *Language, Culture and Curriculum, 21*, 226–234. doi:10.1080/07908310802385899

Moats, L. C. (2000). *Speech to print: Language essentials for teachers.* Baltimore, MD: Brookes.

Molfese, V. J., Beswick, J., Molnar, A., & Jacobi-Vessels, J. (2006). Alphabetic skills in preschool: A preliminary study of letter naming and letter writing. *Developmental Neuropsychology, 29*, 5–19. doi:10.1207/s15326942dn2901_2

Muter, V., & Diethelm, K. (2001). The contribution of phonological skills and letter knowledge to early reading development in a multilingual population. *Language Learning, 51*, 187–219. doi:10.1111/1467-9922.00153

Nicolosi, L., Harryman, E., & Kreschek, J. (2004). *Terminology of communications disorders: Speech-language.* Baltimore, MD: Lippincott Williams & Wilkins.

Ninio, A., & Snow, C. E. (1999). The development of pragmatics: Learning to use language. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of child language acquisition* (pp. 347–383). San Diego, CA: Academic Press.

Norris, J., & Hoffman, P. (1993). *Whole language intervention for school-age children.* San Diego, CA: Singular Press.

Ortiz, A. A., & Kushner, M. I. (1997). Bilingualism and the impact on academic performance. *Academic Difficulties, 6*, 657–679.

Pecnyna Rhyner, P. M., & Bracken, B. A. (1988). Concurrent validity of the Bracken Basic Concept Scale with language and intelligence measures. *Journal of Communication Disorders, 21*, 479–489. doi:10.1016/0021-9924(88)90018-4

Pennington, B. F. (1991). *Diagnosing learning disorders: A neuropsychological framework.* New York, NY: Guilford Press.

Perfetti, C. A., & McCutchen, D. (1982). Speech processes in reading. In N. Lass (Ed.), *Advances in speech and language* (pp. 237–269). New York, NY: Academic Press.

Phelps-Terasaki, D., & Phelps-Gunn, T. (2007). *Test of Pragmatic Language, Second Edition.* East Moline, IL: LinguiSystems.

Piaget, J. (1970). Piaget theory. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology* (3rd ed., Vol. 2, pp. 703–732). New York, NY: Wiley.

Preston, J. L., & Edwards, M. L. (2007). Phonological processing skills of adolescents with residual speech sound errors. *Language, Speech, and Hearing Services in Schools, 38*, 297–308. doi:10.1044/0161-1461(2007/032)

Rapin, I. (1996). Practitioner review: Developmental language disorders: A clinical update. *Journal of Speech and Hearing Disorders, 52*, 105–199.

Rhodes, R. L., Ochoa, S. H., & Ortiz, S. O. (2005). *Assessing cultural and linguistically diverse students: A practical guide.* New York, NY: Guilford Press.

Robertson, C., & Salter, W. (2007). *The Phonological Awareness Test—2.* East Moline, IL: LinguiSystems.

Rondal, J. A. (1995). *Exceptional language development in Down syndrome: Implications for the cognitive-language relationship.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511582189

Roseberry-McKibblin, C. (2002). *Multicultural students with special language needs* (2nd ed.). Oceanside, CA: Academic Communication Associates.

Russell, R. L., & Grizzle, K. L. (2008). Assessing child and adolescent pragmatic language competencies: Towards evidence-based assessment. *Clinical Child and Family Psychology Review, 11*, 59–73. doi:10.1007/s10567-008-0032-1

Ruttenberg, B. A., Wenar, C. W., & Wolf-Schein, E. G. (1991). *Behavior Rating Instrument for Autistic and Other Atypical Children* (2nd ed.). Wood Dale, IL: Stoelting.

Schopler, E., Van Bourgondien, M. E., Wellman, G. J., & Love, S. R. (2010). *Childhood Autism Rating Scale* (2nd ed.). Los Angeles, CA: Western Psychological Services.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluations of Language Fundamentals* (4th ed.). San Antonio, TX: Pearson.

Semel, E., Wiig, E. H., & Secord, W. A. (2004). *Clinical Evaluations of Language Fundamentals, Preschool* (2nd ed.). San Antonio, TX: Pearson.

Shaffer, D. R. (1985). *Developmental psychology: Theory, research, and applications*. Monterey, CA: Brooks/Cole.

Shipley, K. G., & McAfee, J. G. (2009). *Assessment in speech and language pathology: A resource manual* (4th ed.). Clifton Park, NY: Delmar/Cengage Learning.

Skinner, B. F. (1957). *Verbal behavior*. New York, NY: Appleton-Century-Crofts. doi:10.1037/11256-000

Smith, A. R., McCauley, R., & Guitar, B. (2000). Development of the Teacher Assessment of Student Communicative Competence (TASCC) for grades 1 through 5. *Communication Disorders Quarterly, 22*, 3–11. doi:10.1177/152574010002200102

Snow, C. E., Burns, S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academies Press.

Sutherland, D., & Gillon, G. T. (2005). Assessment of phonological representations in children with speech impairment. *Language, Speech, and Hearing Services in Schools, 36*, 294–307. doi:10.1044/0161-1461(2005/030)

Swan, D., & Goswami, U. (1997). Phonological awareness deficits in developmental dyslexia and the phonological representations hypothesis. *Journal of Experimental Child Psychology, 66*, 18–41. doi:10.1006/jecp.1997.2375

Thomas, W. P., & Collier, V. P. (1997). *School effectiveness for language minority students*. Washington, DC: National Clearinghouse for Bilingual Education.

Tomasello, M., & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development, 10*, 201–224. doi:10.1016/0885-2014(95)90009-8

Torgesen, J. K. (1996). A model of memory from an information processing perspective: The special case of phonological memory. In G. R. Lyon (Ed.), *Attention, memory, and executive functions* (pp. 157–184). Baltimore, MD: Paul H. Brookes.

Torgesen, J. K., & Bryant, B. R. (2004). *Test of Phonological Awareness—Second Edition: Plus*. East Moline, IL: LinguiSystems.

Toth, K., Munson, J., Meltzoff, A. N., & Dawson, G. (2006). Early predictors of communication development in young children with autism spectrum disorder: Joint attention, imitation, and toy play. *Journal of Autism and Developmental Disorders, 36*, 993–1005. doi:10.1007/s10803-006-0137-7

Trevarthen, C., & Aitken, K. J. (2001). Infant intersubjectivity: Research, theory, and clinical applications. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 42*, 3–48. doi:10.1111/1469-7610.00701

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R., & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88*, 601–638. doi:10.1037/0022-0663.88.4.601

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101*, 192–212. doi:10.1037/0033-2909.101.2.192

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing*. Austin, TX: Pro-Ed.

West, J., Denton, K., & Germino-Hausken, E. (2000). *America's kindergarteners* (NCES 2000 –070). Washington, DC: National Center for Education Statistics.

Wiig, E. H., & Secord, W. (1989). *Test of Language Competence—Expanded*. San Antonio, TX: Pearson.

Williams, K. T. (2007). *Expressive Vocabulary Test* (2nd ed.). San Antonio, TX: Pearson.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). *Woodcock–Johnson III Tests of Achievement—Normative Update*. Rolling Meadows, IL: Riverside.

Woodcock, R. W., Muñoz-Sandoval, A. F., Ruef, M. L., Alvarado, D. G., & Ruef, M. L. (2005). *Woodcock–Muñoz Language Survey—Revised*. Itasca, IL: Riverside Publishing.

Zucker, S., & Riordan, J. (1988). Concurrent validity of new and revised conceptual language measures. *Psychology in the Schools, 25*, 252–256. doi:10.1002/1520-6807(198807)25:3<252::AID-PITS2310250305>3.0.CO;2-S

CHAPTER 11

TEST USE WITH CHILDREN ACROSS CULTURES: A VIEW FROM THREE COUNTRIES

Thomas Oakland, Solange Muglia Wechsler, and Kobus Maree

Test development and use constitute the flagship activities of applied psychology. They are context dependent and occur uniquely within various cultural contexts. Thus, an understanding of test development and use with children and youth requires an understanding of the cultural contexts within which tests are created and used. This chapter begins with a discussion of some conditions that influence test development and use with children generally, including the size and nature of a country's population together with population trends, the degree of racial–ethnic and social diversity, whether standardized tests normed on children and youths are available, and the existence of national and international guidelines for test development and use.

The chapter then examines the status of test development and use in three disparate countries: South Africa, Brazil, and the United States. These countries were selected to provide a range of comparative information on how educators and psychologists in these countries with large multicultural populations are providing assessment services to children and youths through test development and use.

Testing practices in the three countries differ. Yet professionals in these countries are striving to design and offer assessment services consistent with national and international standards and guidelines. One goal of this chapter is to contrast the ways in which test specialists have responded similarly and differently to the needs and demands of their countries, including their recognition of important issues that influence test use with children from low-income and minority backgrounds.

POPULATION TRENDS

The world's population is large and continuing to grow. The world's population was estimated to exceed 7 billion in 2011 and continues to expand, with an annual growth rate of 1.4% (United Nations, 2009). The five countries with the largest populations, in order of magnitude, are the People's Republic of China (1.5 billion), India (1.2 billion), United States (310,000 million), Indonesia (240,000 million), and Brazil (190 million). The population of sub-Saharan Africa, also known as Black South Africa, is approximately 800,000 million and is expected to double by 2050. South Africa's population is currently 49 million.

Children younger than age 16 make up approximately 27% of the world's population—about 2.2 billion (U.S. Census Bureau, 2010), which approximates the total population of the four most populous countries listed in the preceding paragraph with the exception of India. The distribution of the mean age of the population across countries is also not uniform. The largest numbers of children and youth, as well as the largest numbers of children per family, are found in Africa, the Middle East, and Southeast Asia—regions that generally have fewer economic and educational resources to promote children's development into successful adults.

The larger number of children per family typically strains a family's resources and its ability to provide basic needs (e.g., nutritious food, potable water, durable clothing, and sturdy shelter) and is likely to limit the attainment of basic and higher education that is more commonly provided to

DOI: 10.1037/14049-011
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

children raised in smaller families. The effects of these and other resource restrictions may be seen most immediately among women through early marriage and domestic work and among men through limited educational and vocational opportunities, possibly leading to higher levels of social unrest.

INTERNATIONAL VARIATION IN DIVERSITY

Countries also differ in their historic and current levels of diversity, as exemplified by such characteristics as race and ethnicity, socioeconomic status, age, gender, and sexual orientation. Historically, some countries have had fairly homogeneous racial and ethnic populations (e.g., those in the Middle East and Scandinavia), and others have been diverse. For example, immigrants and their descendants, mainly from Europe, constitute the majority population in most countries in the Americas, from Canada to Argentina. In contrast, with the exception of Liberia and South Africa, fewer people immigrated to Black South Africa. Thus, those populations are more homogeneous and indigenous.

Socioeconomic status generally reflects one's level of education, job, and wealth and may be based on family heritage. *Family heritage* generally includes one's family history together with one's ethnicity and socioeconomic status. These characteristics may contribute to social stratification and thus social advantages and restrictions. Race, ethnicity, and differences in socioeconomic status may also result in the formation and maintenance of social groupings that reflect differences in political, economic, and social power. Children from lower class families who live in countries characterized by greater social stratification and social restrictions often have fewer opportunities to improve their socioeconomic status.

One hundred years ago, people generally knew in which counties or regions specific racial and ethnic groups mainly, if not exclusively, resided. For example, Arabs were generally known to reside in the Middle East; Blacks in sub-Saharan Africa; Chinese in China; Indians in India; and Whites in Canada, Russia, the United States, and Europe. Both legal and illegal migration have changed this traditional landscape, with the result that many countries now display a multiracial and multiethnic character (e.g., France, Germany, Sweden).

Moreover, 100 years ago one could generally define a country's prevailing monocultural characteristics according to three levels: shared and dominant biological and physical qualities (Level 1); shared and dominant religion, history, legal structure, language, values, goals, beliefs, and attitudes (Level 2); and shared and dominant preferences for foods, dress, dating, marriage, and recreation (Level 3; Oakland, 2005). Those monocultural characteristics are far less dominant today. Global migration has created multicultural settings throughout the world, mainly in cities in which shared qualities are less clear and differences are more abundant.

Countries in which cultural differences are new and prominent generally have two concurrent, competing belief systems. Some people view diversity as a country's strength and welcome an influx of cultural qualities that add to the prevailing character of the country. However, others feel diversity threatens historical roots as well as traditional and tested methods of living. Educators are often caught in the attitudinal crossfire that exists between these two competing belief systems.

Educators typically display two important temperament qualities (Lawrence, 1982): They are practical and organized. Educators who value these qualities are generally dedicated to public institutions, including schools, marriage, religion, and other strong institutions that are seen as binding the country together. Tried-and-true traditions are generally valued and maintained through education. The public appreciates educators' dedication to these qualities and expects them and the schooling process to promote Level 2 qualities (i.e., a shared language and an appreciation for and embrace of the country's history, legal structure, values, goals, beliefs, and attitudes). In short, the authors believe that educators generally emphasize conformity more than diversity.

Although newly arrived immigrants to any country cannot be expected to display shared biological and physical qualities with the core culture, they can be expected to acquire, appreciate, and embrace Level 2 qualities as time passes. Newly arrived

immigrants may also be advised to minimize differences that set them apart from the majority by adopting local standards and traditions for foods, dress, dating, marriage, and recreation. Teachers model appropriate behaviors, and schools provide the context for a natural acculturation of young immigrants.

TEST DEVELOPMENT AND USE WITH CHILDREN

Tests have been described as the flagship of applied psychology (Embretson, 1996). Their development and use may constitute psychology's most important technical contribution to the behavioral sciences (Oakland, 2009). Tests are used to describe current behaviors and qualities and to predict future behaviors. Test results assist guidance and counseling services; help establish educational or therapeutic intervention methods; evaluate student progress; screen for special needs; contribute to the diagnosis of disabling disorders; help place people in jobs or programs; assist in determining whether people should be credentialed, admitted or employed, retained, or promoted and are used for various administrative and planning purposes as well as for research.

Test use in some form is universal. Tests are used in virtually every country, with newborns through older adults, and most commonly with students. The ubiquitous teacher-made tests exemplify tests' universality (Hambleton, Bartram, & Oakland, 2011; Hambleton & Oakland, 2004; Oakland, 2004).

Some years ago, test specialists from 44 countries provided information on test development and use for children and youths (Hu & Oakland, 1991; Oakland & Hu, 1991, 1992, 1993). Among the 455 tests identified, the most commonly cited were measures of intelligence (39%), personality (24%), and achievement (10%). Among commonly used tests, 46% were developed in countries other than where they were used. That is, most tests were mainly from the United States and were imported for use in other countries. Tests imported for use came mainly from one of five countries: United States (22%), United Kingdom (7%), Germany (7%), France (5%), and Sweden (5%).

These results are not surprising. The following five general qualities needed for a country to develop and use tests frequently may not be present in many countries: a perception that tests serve important social and personal functions, positive attitudes toward test use (e.g., to favor meritocracy over egalitarianism and individualism over collectivism), a national population of sufficient size and stability to support test publishing, a testing industry responsible for test development (and test adaptation) and marketing, and universities that teach students to use and develop tests (e.g., a specialty commonly called *psychometrics*).

These five general qualities are more commonly found in countries and regions that have more abundant tests (e.g., Australia, Canada, Western Europe, the United States) and are less commonly found in countries and regions that have fewer tests (e.g., Africa, Central and South America, India, Indonesia, the Middle East, the People's Republic of China; Oakland, 2009). Thus, although test use is universal, its use among the world's 220 or more countries is uneven.

INTERNATIONAL PROFESSIONAL GUIDELINES THAT INFLUENCE TEST DEVELOPMENT AND USE

Technology, including the use of tests, has an international impact. Thus, efforts to establish professional guidelines for test development and use, as with other technologies, often require the involvement of international organizations. The International Test Commission (ITC) has assumed leadership for establishing and promulgating guidelines governing test development and use that are thought to be applicable to most countries and cultures. Its guidelines are summarized in the following sections and can be found online (http://www.intestcom.org). The International Standards Organization is also becoming more active in setting international standards for test users.

International Test Commission Guidelines on Test Adaptation

Test adaptation guidelines were developed to help overcome the common tendency for people in emerging countries to obtain standardized tests used in developed countries and merely translate them

into the local language for use. Thus, the test adaptation guidelines provide assistance to people attempting to transform a test intended to be used with one population into one suitable for use with a different population in terms of language, culture, and other differences.

International Test Commission Guidelines on Computer-Based and Internet-Delivered Testing

Testing technology now has an international reach in large measure through the use of the Internet. The legitimate use and potential abuse of computer-based testing are generally well known within the testing industry. Current and potential abuse warrants standards or guidelines for test administration, security of tests and test results, and control of the testing process. Therefore, the ITC established international guidelines on computer-based and Internet-delivered testing that address these and other relevant issues (Bartram & Hambleton, 2006; ITC, 2005).

International Test Commission Guidelines for Test Use

Test use guidelines discuss the fair and ethical use of tests. Their intent is to provide an internationally agreed-on framework from which standards for training and test-user competence and qualifications can be derived. The ITC guidelines underscore five ethical principles important in test use: the need to act in a professional and ethical manner, to ensure those who use tests have desired competencies, to be responsible for test use, to ensure test materials are secure, and to ensure test results are confidential.

Emerging International Test Commission Guidelines

The ITC is committed to developing other methods to improve test development and use internationally. These methods include the development of a test-taker's guide to technology-based testing as well as guidelines on assessing people with language differences, scoring and reporting test data, methods that survey health-related issues, and test security. The ITC may establish guidelines for professional preparation programs that prepare professionals engaged in test development.

More on International Ethical Guidelines That Affect Test Development and Use

People engaged in test development and use and whose work has influence beyond their own country might reasonably be expected to understand and abide by the ethical codes in the countries affected by their work.[1] The importance of ethical codes was underscored centuries ago through the Code of Hammurabi (1795–1750 BC), the earliest known code to specify desired personal and professional behaviors. The Hippocratic Oath (500–400 BC), an ethical code for the medical profession, helped establish a tradition that professional behaviors should be based on overarching moral principles. Modern-day ethical principles spring, in part, from the Nuremberg Code of Ethics in Medical Research, a code developed to regulate experiments involving humans after the Nazi atrocities during World War II.

A partial listing of ethical codes from more than 50 international psychological associations can be found online (http://www.iupsys.net/index.php/ethics/codes-of-ethics-of-international-organizations). Recurring themes found among these ethics statements include the following five broad principles applicable to test service providers.

1. *Promote beneficence and nonmaleficence:* Test service providers strive to enable others to derive benefit from their professional services. Minimally, they strive to do no harm.
2. *Promote competence:* Test service providers should restrict their work to their areas of expertise, established through initial and continued academic and professional preparation.
3. *Promote fidelity and responsibility:* Test service providers work to establish and maintain trust in the services they provide and in their profession.
4. *Promote integrity:* Test service providers are committed to the expression and promotion of accuracy, honesty, and truthfulness through their professional behaviors.
5. *Promote respect for people's rights and dignity:*

[1] Those interested in a further discussion of ethics are advised to refer to Leach and Oakland (2007, in press), Oakland et al. (2012), and Byrne et al. (2009).

> Test service providers strive to promote and respect the dignity and worth of all people. They acknowledge an individual's rights to privacy, confidentiality, and self-determination and acknowledge cultural, individual, and role differences associated with age, gender, gender identity, race, ethnicity, culture, national origin, religion, sexual orientation, disability, language, and socioeconomic status.

ACHIEVEMENT COMPARISONS AMONG MEMBER COUNTRIES OF THE ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

The importance of schooling in promoting civic, personal, and academic development is acknowledged universally. These qualities serve as bellwethers for a country's future well-being, a climate that requires a civil and well-educated population. Thus, efforts to better understand a country's academic development and to compare its academic development with that of similar developed countries may have an impact on a country's future well-being.

The importance of testing to these efforts is demonstrated, in part, by the Program for International Student Assessment (PISA). This program provides a worldwide evaluation of 15-year-old students' scholastic performance. PISA was first conducted in 2000 and has been repeated every 3 years in 27 or more developed or emerging countries (National Center for Education Statistics, n.d.). The Organisation for Economic Co-operation and Development sponsors and coordinates the PISA effort in conjunction with the International Association for the Evaluation of Educational Achievement. The goal of these efforts is to improve educational policies and outcomes internationally.

PISA assesses achievement in reading, mathematics, and science (Table 11.1). Top-performing 15-year-olds generally reside in Australia, Canada, Finland, New Zealand, and Japan. The lowest performing 15-year-olds generally reside in Italy, Greece, Mexico, Portugal, and Turkey. U.S. 15 year-olds were ranked 15th among 27 countries in reading, 21st among 30 countries in science, and 24th among 29 countries in math. Believing that the data would reflect poorly on their country, some countries declined to participate in one or all of the assessments, and others declined to release their national data.

A CLOSER LOOK AT TEST DEVELOPMENT AND USE IN THREE COUNTRIES

The following three sections of this chapter describe the status of test development and use with children and youths in three countries: South Africa, Brazil, and the United States. Some qualities that influence test development and use with children in these countries include their history, size and nature of their population, degree of racial and ethnic diversity, and socioeconomic stratification. These countries were selected for their differences in these qualities.

Test Development and Use With Children in South Africa

Test development in South Africa is still in its infancy. The vast majority of tests currently in use in this country has been imported and adapted for use in this context. Furthermore, the chances of seeing an improvement in the situation are rather small. The costs involved in developing homegrown tests are simply too high. Furthermore, very few researchers venture into this field. It should also be mentioned that, as is frequently the case in developing countries elsewhere in the world, this situation can be explained to a large extent by looking at historical events that shaped the practice of mental health in this country.

Demographic and economic diversity. The South African population of 49 million people includes approximately 39 million Blacks, 5 million Whites, 4.5 million Coloureds (people of mixed origin), and slightly more than 1 million Indians or Asians (Statistics South Africa, 2009). Thus, the overwhelming majority of South Africans are Black. Despite their majority, the concept of minority group is used sociopolitically to refer to Black South Africans and excludes true minority groups of Coloureds, Indians, and Asians (Siberia, Hlongwane & Makunga, 1996).

South Africa, a self-described "rainbow nation," displays diversity in its races, tribes, creeds, and 11 official languages (nine of which are indigenous;

TABLE 11.1

Mean Achievements of Organisation for Economic Co-operation and Development Member Countries in Three Achievement Areas

Country	Mean achievement
Reading[a]	
Finland	546
Canada	534
New Zealand	529
Australia	528
Ireland	527
South Korea	525
United Kingdom	523
Japan	522
Sweden	516
Austria	507
Belgium	507
Iceland	507
Norway	505
France	505
United States	504
Denmark	497
Switzerland	494
Spain	493
Czech Republic	492
Italy	487
Germany	484
Hungary	480
Poland	479
Greece	474
Portugal	470
Luxembourg	441
Mexico	422
Mathematics[b]	
Finland	544
South Korea	542
Netherlands	538
Japan	534
Canada	532
Belgium	529
Switzerland	527
Australia	524
New Zealand	523
Czech Republic	516
Iceland	515
Denmark	514
France	511
Sweden	503
Austria	506
Germany	503

Country	Mean achievement
Mathematics (*cont.*)	
Ireland	503
Slovakia	498
Norway	495
Luxembourg	493
Poland	490
Hungary	490
Spain	485
United States	483
Italy	466
Portugal	466
Greece	445
Turkey	423
Mexico	385
Science[c]	
Finland	563
Canada	534
Japan	531
New Zealand	530
Australia	527
Netherlands	525
South Korea	522
Germany	516
United Kingdom	515
Czech Republic	513
Switzerland	512
Austria	511
Belgium	510
Ireland	508
Hungary	504
Sweden	503
Poland	498
Denmark	496
France	495
Iceland	491
United States	489
Slovakia	488
Spain	488
Norway	487
Luxembourg	486
Italy	475
Portugal	474
Greece	473
Turkey	424
Mexico	410

Note. A 9-point difference is sufficient to be considered significant.

[a]Evaluated in 2000. [b]Evaluated in 2003. [c]Evaluated in 2006.

Rainbow Nation, n.d.). Most South Africans speak more than one language and can communicate in English. English, often used in official business, is the first language, used by only 10% of South Africans. Zulu, Xhosa, Afrikaans, Pedi, Tswana, English, and Sotho (in that order) have the most speakers (Rainbow Nation, n.d.). Most South Africans are Christian. The South African constitution guarantees freedom of religion and speech.

Diversity's impact on test development and use. Apartheid was officially overturned in South Africa in 1994. However, many conditions that influenced test development and use during apartheid linger. For example, vast differences in school quality continue between White and Black students as well as for students living in poor rural townships or who attend inner-city schools. The quality of educational services is considerably lower than that provided to students from affluent families (the vast majority of whom are White despite the fact that the economic situation of many Black families has consistently been improving). Some social scientists believe tests of crystallized abilities (e.g., achievement and similar abilities that are typically acquired through education) serve to maintain these apartheid-established differences (Sehlapelo & Terre Blanche, 1996; Stead, 2002; Vandeyar, 2007). Issues associated with race and socioeconomic status are intertwined, including deprivation that is steadily worsening among the Black and, to an increasing extent, White populations, manifested by malnutrition, educational deprivation, poverty, unemployment, widespread reliance on government grants for survival, and high crime levels.

Historically, international trends that emphasized applied research over basic research have been favored in South Africa (Painter & Terre Blanche, 2004). Psychological testing, as one example of applied research, was introduced at the beginning of the 20th century when South Africa was a British colony. European and North American scholarship and technology have generally been accepted in South Africa and provided the basis for its nascent research. This scholarship and technology was most relevant to White South Africans (Stead & Watson, 1998). Measures of aptitude, achievement, and personality were generally developed for use with the White population along with some essentially parallel test forms developed for Black students, including some tests with separate racial norms. Few tests have been designed specifically for South Africa's diverse and largely Black population (Claassen, 1995; Foxcroft, 1997; O. Owen, 1991).

Assessments commonly used by assessment specialists. Tests that assess emotions, personality, interests, and academic and vocational aptitudes, as well as visual–motor functioning, are commonly used in South Africa (Foxcroft, Paterson, Le Roux, & Herbst, 2004). As in other countries, test use patterns differ across the various specialties in psychology. For example, clinical psychologists use tests mainly to assess intellectual, personality, and neuropsychological functioning. Counseling psychologists, however, use tests primarily in vocational assessments to assess career interests, intelligence, and personality. Industrial psychologists use them mainly to assess career interests, intelligence, and personality as well as occupational potential and employment suitability. Educational (i.e., school) psychologists mostly conduct psychoeducational assessments that focus on intelligence and other aptitudes as well as academic achievement, emotional adjustment, career development, personality types, and visual–motor functioning (Foxcroft et al., 2004).

Given the diverse South African population, it is surprising that none of the most popular tests has been standardized for use in a multicultural context. Some tests, including the Bender Visual Motor Gestalt Test (Bender, 1938) and the Developmental Test of Visual–Motor Integration (Beery, Buktenica, & Beery, 2010) have been imported and used without first conducting full-scale national normative studies, thus depriving practitioners of access to appropriate South African norms. In addition, some tests are outdated and were not developed to be used with the vast majority of South Africans. The latter include tests of intelligence such as the Senior South African Individual Scale—Revised (Van Eeden, 1991) and personality tests such as the Nineteen Field Interest Inventory (Fouché & Alberts, 1986). Thus, most tests are either outdated or have never

been normed on the diverse South African population, thus limiting their practical utility (Foxcroft et al., 2004; Nell, 1990, 1994; K. Owen, 1998; O. Owen, 1991).

The various assessment methods used for identifying, diagnosing, and intervening with individuals who have psychological problems are delineated in the Scope of Practice of Psychologists (South Africa Department of Health, 2010). The methods outlined in this document include assessments of cognitive, personality, emotional, and neuropsychological functioning of students for the purposes of diagnosing psychopathology; identifying and diagnosing barriers to learning and development; guiding psychological interventions to enhance, promote, and facilitate learning and development; informing therapeutic interventions in relation to learning and development; and referring students to appropriate professionals for further assessment or intervention.

How educational or school psychologists address cultural diversity in their assessments. White examiners are more common than Black examiners throughout South Africa (Claassen & Schepers, 1990). This racial imbalance is the result of few Blacks becoming trained as specialists in test use. Many reasons exist for Blacks not being trained in the use of tests. For example, fewer Blacks enter university programs, in part because of the lower achievement levels among those who graduate from high school compared with their White peers. Furthermore, Blacks generally do not view psychology as a preferred field of study. As a result, very few promising Black students major in psychology at the university level. Some writers have proposed that this imbalance may be exacerbated because few White vocational counselors speak a Black language and thus do not effectively communicate job opportunities to Blacks (Watson & Fouche, 2007).

Although an increasing number of Black psychologists have received psychometric training in recent years, such training tends to promote assessment methods commonly used in the United States and other Western countries and are thus more relevant to South Africa's White citizens than its Black citizens. These Western-centric models have been described as being inconsistent with the language and other cultural characteristics of lower socioeconomic status Blacks and thus need to be changed to accommodate the needs and characteristics of the Black majority population (Berry, 1985).

Another problem contributing to the assessment racial divide is that most White psychologists in South Africa reportedly feel underprepared to work with or assess students from a cultural group other than their own (Ruane, 2010). The South African National Department of Education (2008) drafted a National Strategy on Screening, Identification and Support to address issues of cultural diversity in assessments. This strategy delineates the processes and protocols used to screen, identify, and assess special needs students and to reduce the number who are placed in special schools away from their homes, without their primary support system. This strategy was designed to promote an inclusive local education and training system to address the cultural challenges psychologists experience as they assess a diverse student body.

Systemic problems associated with conducting assessments that are not indigenous to test use. The following conditions characterize assessment practices in South African schools and constitute some of the major problems impeding assessment services for South African students (Eloff, Maree, & Ebersöhn, 2006). The profession of psychology adheres largely to a quantitative (i.e., positivist) approach to understanding the world. Despite radical postapartheid public policy changes, psychological services, including the use of tests, have changed little since 1994. The main use of tests in South Africa is to meet the needs of individuals, not of groups.

A focus on individuals is inconsistent with the prevailing South African concept of *ubuntu*, which underscores a culture of sharing. Thus, group assessment is likely to tap into the feeling of esprit de corps that typifies African culture. Eliciting individual narratives in a group context implies that children are guided to write their stories while the counselor facilitates the direction of these stories. Counselors may explore children's stories as a means through which their experiences are revealed, leading to building and mapping their future (Maree, 2010b).

Psychologists often view group assessment based on qualitative approaches (e.g., assessment characterized by the use of nonstandardized techniques such as the life line, collage, drawing, and other narrative techniques) with skepticism, believing such methods lack needed reliability and validity and are thus less useful than scores obtained from standardized instruments. This narrow viewpoint persists. Although group testing is generally more cost effective than individually administered testing, analyzing a large number of individual narratives requires a high level of personal exposition and is time consuming and consequently costly. Furthermore, few psychologists have received adequate training in the use of narratives or similar forms of qualitative assessment.

Efforts by South African professionals to act as facilitators and agents of change are generally regarded as inadequate by parents and educators. For example, networking, referral, and collaboration between education departments, university departments of education and psychology, and educators are insufficient and ineffective. Training and service delivery models are needed that emphasize team interactions, including the need to go beyond traditional disciplinary and departmental borders when collaborating within organizations and communities (Biggs, 1997). The current district and school-based support teams are seen by some as being unsuccessful in providing assessment and intervention opportunities for students and devising programs to facilitate students' development (Eloff et al., 2006).

Need to facilitate equitable and fair testing practices in South Africa. Four years after the demise of apartheid, K. Owen (1998) concluded that mounting problems related to psychological testing (e.g., financial constraints, unmanageable psychologist:student ratios, politicians' reservations about testing) might spell the end of psychological testing in South Africa. The attainment of equitable and fair testing practices in South Africa was thought by some to be limited in the then-foreseeable future because of practical limitations. For example, the goal to develop separate forms of tests for all of South Africa's different cultural groups is not practical or feasible.

South Africa may benefit from an assessment approach that resembles the universal design for learning movement (Hanna, 2005), in which a flexible approach to learning addresses the unique needs of each learner. In such an approach, emphasis is placed on identifying students' general and unique learning needs in light of the curriculum and then providing corresponding services. This universal design movement calls for the use of multiple instructional methods because no one approach works effectively for all students. Innovative approaches to assessment, teaching, and learning are needed that underscore "the need for inherently flexible, customizable content, assignments, and activities" (Hanna, 2005, p. 3).

Despite the views held by some that psychological testing in South Africa does not have a viable future, in reality tests based primarily on North American and Eurocentric approaches may still be used with various South African cultural groups (K. Owen, 1998) provided they are adapted to reflect the country's multicultural characteristics and evidence the validity and relevance that warrant their use (Paterson & Uys, 2005).

Need to rethink the current paradigm underpinning test development and use in South Africa. The current paradigm underpinning test development and use in South Africa needs to be reviewed and revised. The applied uses of test data make up only one component of an entire system that needs evaluation and change. Possible remedies to the extant paradigm include increased focus on systems change, organizational analysis, whole-school development, and the adoption of constructivist perspectives that link emotional development, teaching, learning, and assessment. Reliance on contemporary developments in psychology are needed in South African schools to facilitate, among others, an understanding of students' life stories, dynamic assessment that helps to identify and nurture students' strong points (De Beer, 2006; Maree, 2010a, 2010c), and emotional intelligence assessment that facilitates students' identification and management of their emotions. This reorientation implies a movement away from a narrow focus on student needs and deficits and toward the belief that all students possess strengths and assets that should form the basis of effective intervention (Eloff et al., 2006). A focus on students' needs and deficits should

not be regarded as antithetical to the collection of information that informs effective interventions. Quite the opposite: An approach is needed that emphasizes the importance of developing students' strengths as enthusiastically as it addresses weak points in a remediating manner.

Students from poor families are at a special disadvantage in South Africa because of their limited exposure to some of the cultural content found in many cognitive tests. Although tests are intended to address the needs of these diverse students, the largely Western-based content and technological language found in these instruments often make them unsuitable for this population.

South African psychology is attempting to strike a balance between assessment methods commonly used in Western countries and those that may be in greater harmony with South African realities and culture (Sehlapelo & Terre Blanche, 1996), including a theoretical framework that combines both quantitative and qualitative methods of assessment. For example, the Career Interest Profile (Maree, 2010b) combines both quantitative assessment of students' interests and qualitative assessment by means of a narrative supplement that facilitates discussion and analysis of interests through students' life story. The dearth of recently developed tests for use in South Africa, and the even fewer tests appropriate for use with all South African cultural groups, prompted some psychologists to request the South African Professional Board of Psychology to allow professionals to use tests not registered by the Health Professions Council of South Africa, the agency responsible for certifying test suitably (Foxcroft et al., 2004; Paterson & Uys, 2005). Psychologists may be able to use unregistered tests if they can justify their use with evidence that the tests meet acceptable overall standards. Others have advocated for an overarching code of fair testing to guide test practices (Foxcroft, 1997).

Role of language in test development and use. Language plays a crucial role in the development and use of tests in South Africa. As noted previously, South Africans collectively speak 11 official languages, nine of which are indigenous and, although not widely used, nevertheless important to those individuals who use them. Thus, the wide number of languages used requires psychologists and others to be sensitive to language differences as well as language deficits when developing and using tests and to be alert when these differences and deficits may form a barrier to fair test use. Before the 1990s, tests developed in South Africa used one of the two former official languages, English and Afrikaans—languages used mainly by Whites. The restricted range of languages used in most tests led some to resist test use, given the belief that tests were and remain unfair and their use prejudicial (Paterson & Uys, 2005).

Possible remedies to language problems and education-related biases include the use of English or Afrikaans content and norms based on education and language proficiency levels for people with 10 or more years of schooling as well as separate norms for those with fewer than 10 years of schooling (Foxcroft, 2004; K. Owen, 1998; Stead, 2002). Professionals should ensure an examinee's test performance is not attenuated by language differences.

Role of academic acculturation in test performance. Additionally, when assessing students' cognitive abilities, professionals should also ensure an examinee's background has provided an opportunity to acquire the knowledge assessed on the test. That is, the use of measures of achievement assumes students have been exposed to the content measured by these measures. For example, an assessment of basic addition and subtraction assumes students have been exposed to this mathematical content at school or home. School or home environments in impoverished rural and township areas are often austere, thus limiting children's exposure to information commonly included on cognitive tests. Racial minority status further exacerbates this limitation because children from racial minority groups are less often exposed to certain cultural content. In contrast, children living in more affluent environments are more likely to be exposed to the breadth of information assessed on cognitive tests.

Role of rapport in test performance. Professionals attempt to maximize the likelihood that an assessment is valid by creating and maintaining rapport that fosters good relationships between an examiner

and examinee. Trust is central to this relationship. South Africa's apartheid past together with the country's current racial divide have created feelings of ill will between many Whites and Blacks. As a result, Black children taking tests may feel more comfortable with and perform better when tested by an examiner from their cultural and linguistic background whom they trust. This belief comes mainly from clinical practice and needs empirical review.

Role of standardized testing practices in test performance. Another problem lies in whether standardized tests are administered in the standardized fashion, as intended. For example, when standardizing a test, students are generally seated individually at desks in a classroom and complete their work in a setting that is quiet and devoid of other distractions. Thus, when administering the standardized test, examiners strive to replicate these conditions. However, in reality, test administration in South Africa often occurs under vastly different conditions. For example, tests may be administered outside of classrooms, devoid of desks, and instead require students to sit on the ground while completing them. Alternatively, tests may be administered in large, overcrowded classrooms, with two or more students sharing a desk, under distracting conditions that defy adequate supervision.

In summary, South African students, confronted with change and its radical impact, face various challenges in postapartheid South Africa. These challenges cannot be addressed by the traditional (i.e., quantitative) approach to counseling alone because this approach does not integrate contextual factors and the personal meaning an individual attaches to life experiences, including decision making, during the counseling process. Efforts to help ensure tests are used appropriately with South Africa's minority populations must address issues pertaining to creating a balance between qualitative and quantitative methods (see earlier discussion), ensuring that tests display suitable psychometric qualities, determining whether the student has facility with the language used in the test, ensuring students are comfortable working with professionals who differ in race, and administering tests in a standardized fashion.

Efforts to prohibit the use of intelligence tests in South African schools. Educational authorities have largely dispensed with intelligence testing in schools and generally disparage its assessment. For example, in 1995, the National Education Department of South Africa placed an unofficial moratorium on the use of group intelligence tests in schools with children from all racial and ethnic groups. This moratorium created a widespread vacuum in schools that deprived some students, their parents, and educators of test data that may have been important to the students' receiving appropriate educational services. Examples include academic and vocational counseling for entrance into programs for gifted students and ruling out possible mental retardation when making school-based diagnoses. This moratorium did not affect assessment services provided by private practitioners (i.e., those whose services are more readily available to children from affluent families). Therefore, most South African students (i.e., those from predominantly low-income homes) have less access to state-of-the-art assessment services and instead receive limited and inferior services provided mainly by school districts.

Recommendations. A paradigm shift is needed in the provision of testing services in South Africa—one that levels the playing field between students from economically deprived families and students from affluent families. Individual and group assessment methods are needed, especially in under-resourced schools in rural and township areas. Additional efforts are needed to identify educational problems early, thereby leading to interventions that may prevent problems and promote development to help reduce the current 50% student attrition between the first and 12th grades. The announcement by the Minister of Basic Education (Motshekga, 2010) that assessments will be conducted in Grades 3, 6, and 9 in literacy (in home language and first additional language) and mathematics is a step in the right direction.

The following approaches may help psychologists better address cultural diversity when conducting assessments in the South African context (Eloff et al., 2006). Existing professional resources (e.g., existing school support services at the district and other

levels) need to be improved, and the services of educational psychologists in particular need to be increased. Important issues of equity, access, and redress may be addressed by encouraging or even compelling (as is currently the case with many occupational practitioners in South Africa) psychologists and teachers to perform community service by working in township and remote rural schools immediately after their training. Increased salaries and tuition forgiveness may be offered as incentives to add additional professional resources to those schools most in need of services. Additionally, psychologists and teachers need appropriate training for their work in rural schools, including how best to work with large and understaffed classes. Instruction from staff who are knowledgeable of and experienced in these topics is also needed. Initial efforts to promote social change and school development should be evaluated and subsequently modified on the basis of concrete evidence provided by departments of education. Social change and school development strategies should include prevention, group work, empowerment or enablement, parental guidance, community involvement and networking, referral, and collaboration. The employment of assessment specialists to work and reside in rural areas is also needed.

The work of all South African psychologists should be guided by three broad principles found in their ethics code (Professional Board for Psychology, 2006) as well by South African policy that guided the country from an apartheid state to a full democracy: to promote equity and access and to redress grievances. The principle of promoting equity suggests that specific groups should not receive privileges on the basis of individual characteristics, such as gender, resources, culture, language, or race. To redress grievances in South Africa, psychologists should understand and be committed to finding ways to emend historic and existing unevenness in the provision of psychological services in South Africa. Assessment practices should provide redress for past erroneous ways and promote access by striving to make testing services more available.

Given the importance of test development and use viewed against the backdrop of South Africa's many cultures, reforms in assessment practices are needed to cater to the country's unique social composition and to ensure quality and international comparability. Furthermore, in light of the multicultural and dynamic contexts in which tests are used, merely training personnel to work more wisely with defective tests and perhaps within a defective assessment model, although currently necessary, is only a temporary solution.

Thus, an open, empathetic, best-practice approach, including the use of existing imported and locally developed tests and the development of new tests that take South Africa's multicultural and dynamic context into account is in the country's best interests and may be the best viable option. The success of these efforts could empower the Psychological Society of South Africa, a member of the ITC, to "shape international guidelines related to testing and test use and stay in touch with the cutting-edge issues in testing and assessment" (*Annual Report of the Psychological Society of South Africa,* 2007, p. 9).

Last, a paradigm shift is needed in test development to ensure that more culturally sensitive tests are designed to reflect concepts and items that make more sense to examinees. This shift can be achieved through the involvement of researchers who understand and share the cultural and linguistic forms of culturally different groups. Tests should not be based on a one-size-fits-all philosophy and should instead reflect an understanding of the examinee's culture, including her or his school, home, recreation, and working environments.

Test Development and Use With Children in Brazil

Demographic and economic diversity. An understanding of test development and use in Brazil is enhanced by first providing a somewhat broad understanding of the country and its characteristics. Brazil's population of 190 million includes 57% Whites, 33% Mulattos, 10% Blacks, 0.7% Asians, and 0.1% Indians (Instituto Brasileiro de Geografia e Estatística [IBGE], 2010). The population's geographic distribution is irregular. For example, the southeast region (e.g., the states of São Paulo, Rio de Janeiro, Minas Gerais, and Espírito Santo) includes 50% of the population. São Paulo, Brazil's largest city

and located in this region, has approximately 12 million inhabitants. The northeastern states have fewer people, and the northwestern states have the fewest. Portuguese is the country's official language.

Prevailing conditions associated with socioeconomic inequalities among Brazil's population constitute one of its greatest challenges. An estimated 10% of the richest people control 50% of the country's wealth. Approximately 23% of families have monthly incomes of less than $230 (IBGE, 2009a). Although Brazil has the eighth largest economy in the world, its extreme income inequality contributes to problems of social exclusion and economic growth. In 2006, White workers earned on average 40% more than Black or mixed-race workers with the same level of schooling (Ministério da Educação e Cultura, 2008).

Mortality rates among infants and children are high throughout the country, yet dropping, with a rate of 24% nationally and 34% in the northeastern and poorest states (IBGE, 2009b). In Brazil, as in most other countries, education levels are closely related to family income and infant mortality rates (De Barros, Henriques, & Mendonça, 2000), thus underscoring the importance of educational policies in the country's development.

Educational assessment and challenges. Brazil's basic educational system includes elementary (ages 7–14) and high school (ages 15–17). Elementary education is mandatory for all children and has a 97% registration rate (IBGE, 2009b). State and municipal governments financially support public education. Approximately 80% of students, mainly from low-income homes, attend public schools, whereas others, mainly from more affluent families, attend private schools (IBGE, 2009a)

The quality of education in public schools is problematic. For example, approximately 19% of students repeat one or more grades, and 14% of students drop out before completing 6 years of schooling (Ministério da Educação e Cultura, 2008). Nationally, children attend school an average of 7.4 years—8.1 years in the southern region and 6.2 years in the northeastern region. Dropout rates increase with each grade level and are highest among low-income families (Instituto de Pesquisa Econômica Aplicada, 2008).

Low achievement, especially among public school students, is worrisome. Data from PISA were acquired in three academic areas: reading, mathematics, and science. Brazilian students performed below international averages in 2000, 2003, and 2006, albeit showing small improvements during these years. For example, the 2006 evaluation results, reported on a scale ranging from 1 (*low*) to 6 (*high*), found high percentages of Brazilian students at Level 1: 56% in reading, 75% in math, and 61% in sciences (see http://enem.inep.gov.br). Brazilian students ranked 54th among 57 countries on their literacy skills, thus highlighting the country's problematic educational system.

Academic achievement is likely to be better understood and promoted through knowledge of students' assessed achievement. However, Brazilian teachers have difficulty evaluating their students' academic achievement because of a lack of standardized achievement tests. Furthermore, public schools are not legally required to use school psychologists, thus depriving schools of information on individual differences in the learning process (Prette, 2008; Wechsler, 1996).

Several efforts have been made to have school psychologists officially recognized and to provide psychological services in public systems since the 1991 formation of the Brazilian School and Educational Psychology Association (Wechsler, 1996). However, only a few professionals who work at health centers provide public services to children and youths. The nature of their services ranges somewhat broadly and is not limited to education.

Test use in Brazil: The three waves. The history of test development and use in Brazil can be characterized in three chronological waves. During the first wave, the importance of tests was recognized, and they became used somewhat widely in educational, clinical, and organizational assessment. Most tests were imported from the United States and Europe. The second wave, the fall of psychological tests, occurred in response to criticisms of tests' lack of scientific rigor as well as political views that tests were culturally unfair and overlooked Brazil's diversity and socioeconomic inequalities. During the third and current wave, the scientific merits of

tests have been recognized by society as a result of test authors' development of tests that are more reflective of the country's cultural characteristics and needs.

The reversal in acceptance that occurred from the second to the third waves was due, in part, to efforts by various Brazilian psychologists to adapt and construct tests to reflect Brazilian reality. The Brazilian Institute of Psychological Assessment (IBAP), founded in 1997, has helped coordinate these efforts. The work of the *Conselho Federal de Psicologia* (Federal Council of Psychology; CFP), Brazil's national psychological association, in establishing and promulgating standards for test quality, has also had a decisive influence on test quality (Wechsler, 2009).

First wave: Importance of testing. Tests were introduced to Brazil in the late 19th century, largely in schools (Angelini, 1995). By 1914, a psychological laboratory was organized in São Paulo and later affiliated with the first educational and psychological undergraduate programs offered by the University of São Paulo. Foreign psychologists from Italy, France, Spain, Russia, and Poland contributed to these initial efforts (Pfromm Netto, 1996; Wechsler, 2001).

The importance of tests in educational, clinical, and personnel assessment was widely recognized by the late 1940s (Pasquali, 2010; Penna, 2004), and an infrastructure to support their use was forming. For example, an institute for vocational guidance was established in Rio de Janeiro. A scientific society was formed, the *Sociedade Brasileira de Psicotécnica,* later renamed the *Sociedade Brasileira de Psicologia Aplicada* (Brazilian Society of Applied Psychology). The first scientific journal, *Arquivos Brasileiros de Psicotécnica* (*Brazilian Psychometric Archives*) published research on tests use; it was later renamed the *Arquivos Brasileiros de Psicologia* (*Brazilian Psychological Archives;* Pessoti, 1988).

The period between 1950 and 1960 was very productive for those interested in testing. Funds from industrial agencies (*Servico Nacional de Aprendizagem Comercial, Servico Nacional de Aprendizagem Industrial*) were invested into developing tests for personnel assessment, along with the development of a few group intelligence tests (e.g., General Intelligence–G36, Non-Verbal Intelligence Test–INV) and aptitude batteries (e.g., *Bateria Fatorial CEPA*). These tests were used for many years. Moreover, the use of several psychological tests (e.g., general intelligence, personality, aptitude) to obtain a driver's license became legally required, thus acknowledging the importance of testing and expanding the market for psychological services as well as for test development and use nationally.

Second wave: Tests are not highly regarded. Shortly after 1960, a second period, lasting approximately 20 years, began in which tests were not considered sufficiently important to warrant financial and professional investments. Because of negative attitudes toward testing, work on test development or revisions largely halted. As a result, Brazilian norms or other test adaptations were not available for a long time, which led to greater use of theory-driven projective measures to assess psychological dysfunction (Wechsler, 2001).

Criticisms of test use during this period came from both scientific and political sources. Scientists criticized tests for not being constructed or adapted in light of Brazilian culture, leading to the belief that such tests could not effectively measure Brazilians' cognitive and personality qualities (Noronha, 2002). Prevailing political views also did not favor test use. Tests highlight individual and group differences and were viewed as being antithetical to prevailing socialist and collectivist views. Considering Brazil's huge socioeconomic differences, along with lower test scores for those from lower socioeconomic status levels, test use was seen as favoring the more privileged individuals and groups while overlooking the needs of the less privileged ones (Instituto Brasileiro de Avaliacao Psicologica, 2002).

These negative views had a deleterious effect on test production as well as on psychology students' interest in taking testing courses (Alves, 2009). Classes promoting psychotherapy skills were seen as more relevant to preparation for a professional career than classes dedicated to assessment. Students of the era preferred qualitative methods that emphasized an understanding of the entire person to quantitative methods that seemingly led to more narrow views (Hutz, 2009, 2010).

Third wave: The testing movement progresses. Brazilian psychologists began responding to criticisms of the testing enterprise by investing effort in test construction and adaptation. Starting in the 1980s, university-based laboratories were formed: the first in Brasilia, Brazil's national capital (Federal University of Brasilia), followed by three others in the state of São Paulo (State University of São Paulo, Pontifical Catholic University, and University of São Francisco) and one in the south in Rio Grande do Sul (Federal University of Porto Alegre). Today, more research groups are interested in test development, and the number of laboratories has increased and are present in every Brazilian region. Additionally, the growth in test development and use has led to hiring more professors to offer courses that prepare professionals to develop and use tests.

An increased interest in the scientific qualities of tests led to the founding of IBAP in 1997. Its four national conferences each attract approximately 1,000 participants, many of whom are students and young professionals who discuss their test-related research, thus demonstrating a degree of support that suggests a bright future for test development and use in Brazil. The association's scholarly journal, *Avaliação Psicologica* (*Psychological Assessment*), features research on test construction and adaptation in every issue (Wechsler, 2009). An international journal of psychological assessment is also being developed by IBAP to advance test construction and use throughout Latin American countries.

The Federal Council of Psychology (CFP) regarded the founding of IBAP as extremely positive. The IBAP formed a national commission of professionals with expertise in psychological assessment. Later, in 2001, CFP assumed leadership, with Brazilian Institute on Psychological Assessment's (IBAP's) assistance, in establishing national standards for test development, quality, and use. Guidelines on test use approved in 2000 by the ITC helped form the basis of these standards. In 2003, CFP adopted federal regulations that require that all tests used in the country have empirical evidence of their validity, reliability, and norms relevant to Brazil. This regulation emphasizes the need to adapt tests in light of the Brazilian environment before using tests in Brazil.

All existing and new psychological tests must be evaluated under these new rules, and a national commission was convened to evaluate tests in light of the new standards. The titles of approved tests are listed on CFP's website (http://www2.pol.org.br/satepsi/sistema/admin.cfm) to inform psychologists and the public. Psychologists whose practices disregard the requirement to use only approved tests may face sanctions under the profession's ethics code (CFP, 2003). Tests not listed may be used only for research. In addition, currently approved tests need to be reviewed every 15 years on the basis of new evidence of the test's validity and norms (CFP, 2010).

A somewhat large and increasing number of tests have been adapted or created since this resolution, with 210 tests submitted for examination by the national commission between 2003 and 2010. Among those submitted, 114 tests were approved, 77 were disapproved, and 19 remain under review, thus indicating the impact of this regulation on Brazil's existing tests (Anache & Correa, 2010).

The third wave of test use in Brazil, highlighting the potential value of tests, has also influenced educational research. National and state leaders have increasingly recognized the value of data acquired through large-scale testing programs when forming and reviewing educational policies. Two major government exams evaluate public elementary school achievement: the System for Assessing Basic Education, which is administered in randomized samples of schools in each state, and Prova Brazil, a more broadly administered national exam. Both exams assess student yearly achievement in language and mathematics in the fourth and eighth grades and at the end of high school. The National Assessment for Middle Education, another large-scale educational assessment, examines public middle school students in language, humanities, biological sciences, mathematics, and technology. Test scores on this exam are usually required for entry into public universities (See http://enem.inep.gov.br; Ministério da Educação e Cultura, 2010). Results from these assessments are reviewed for the purpose of establishing and revising educational policies, including academic content.

An emerging fourth wave: A vision of the future. Leaders in government, education, and industry

have increasingly recognized the value of test data to guide decision making about the country's future. The evolution of the country's testing infrastructure, described in reference to the preceding three waves, provides a firm foundation for this work. However, much remains to be done.

Most tests used in Brazil are designed for typically developing children between ages 6 and 12. Additional tests are needed for children with special needs as well as for younger and older children (e.g., preschool and adolescent populations). Assessing younger and older populations could provide important information on developmental progress, thereby aiding in the identification of children whose development differs from that of typically developing children. Moreover, industrial and commercial organizations need more tests for use in personnel assessment, including assessment of skills and personality characteristics related to work productivity in Brazilian business contexts (Wechsler, 2010).

Universities must continue efforts to improve their research infrastructures leading to test development and evaluation as well as the preparation of those who will develop and use the next generation of tests. Companies that publish and distribute tests will have to consider the increasing number of university-based test development and research laboratories and increase their liaison with this scientific community to be able to publish tests developed in Brazil. Finally, tests that assess specific educationally relevant qualities, established by educational policies, must be developed and made available to teachers and other educational personnel to help them assess student achievement in various subjects and grades.

Test Development and Use With Children in the United States

Demographic and economic diversity. The U.S. population is approximately 340 million, among which 20% are between ages 0 and 14. The country's population is mainly urban (82%) and White (80%, among whom 15% are Hispanic), together with 13% Black; 4% Asian; and smaller percentages of American Indians, Alaska natives, native Hawaiian, and other Pacific Islanders. Students are mandated to remain in school until age 16. Currently, more women than men are entering and graduating from colleges and universities. An estimated 99% of the U.S. adult population is literate (Central Intelligence Agency, n.d.).

Between July 1, 2005, and July 1, 2006, Hispanic and Latino Americans accounted for almost half (1.4 million) of the national population growth of 2.9 million. Immigrants and their U.S.-born descendants are expected to provide most of the U.S. population increase in the decades ahead. In addition to the influx of Hispanics as a result of immigration patterns, the high birth rates among Hispanics foretell a growing number and percentage of Hispanic students in the U.S. population.

Median family incomes in the United States vary by race and ethnicity, with an estimated median income of $58,000 for Asians, $49,000 for non-Hispanic Whites, $34,000 for Hispanics, and $30,000 for Blacks (U.S. Census Bureau, 2010). Approximately 1% of the U.S. population is classified as super-rich, 5% as rich (e.g., having an estate worth more than $1 million), 44% as middle class (e.g., college educated with an individual annual income of between $40,000 and $57,000), 39% as working class (e.g., high school educated with an individual annual income between $26,000 and $40,000), and 11% as poor (e.g., some high school, individual annual income less than $18,000 or chronically unemployed). Religious affiliations in the United States commonly include Protestant (51%), Roman Catholic (24%), Mormon (2%), other Christian (2%), Jewish (2%), Buddhist (<1%), Muslim (<1%), other or unspecified religions (2%), unaffiliated (12%), and none (4%). Thus, the U.S. population displays diversity in race and ethnicity, socioeconomic status, and religious affiliations. Examples of the influence of race, ethnicity, and socioeconomic status on test development and use are provided in the discussion that follows (Central Intelligence Agency, n.d.; U.S. Census Bureau, 2010).

The United States is also diverse in language. English is the dominant language (82%), and Spanish (11%), other Indo-European languages (4%), and those used by Asian/Pacific Islanders (e.g., Mandarin, Cantonese, Tagalog, Vietnamese) account for

approximately 3%. Additional languages spoken in various U.S. cities and states include Arabic, German, Greek, Italian, and Polish. Immigrant students in some urban school districts have collectively reported using more than 120 languages at home. In all, an estimated 337 languages are spoken or signed in the United States (*Languages of the United States,* n.d.).

Five qualities needed for a strong testing industry. As noted in the introduction to this chapter, a country needs five qualities to have a vibrant testing industry and to use tests frequently: a perceived need for tests that serve important social and personal issues, a positive attitude toward test use (e.g., favoring meritocracy over egalitarianism and individualism over collectivism), a national population of sufficient size to support such an industry, a testing industry responsible for test development (or test adaptation) and marketing as well as universities that offer programs that teach students to develop tests (e.g., psychometrics) and use them professionally. The United States has these five qualities, which has thus given rise to a vibrant testing industry and frequent test use. These five qualities are discussed more fully next.

A perceived need for tests that serve important social and personal issues. The United States is technology savvy. Tests form part of this technology. Thus, parents, educators, and policymakers commonly see value in having children tested to better understand current behaviors and other qualities, estimate future behaviors, assist guidance and counseling services, establish intervention methods, evaluate progress, screen for special needs, diagnose disabling disorders, and help place youths in jobs or programs. In fact, federal and state laws increasingly require the use of tests and other data-gathering methods to assist in making these decisions with children. Adults commonly use tests to become credentialed, admitted or employed, retained, or promoted.

However, tests are not always used to address important social and personal issues. As with other tools, tests are likely to fall short of their goal of providing information that serves the individual and the public when they are administered by test users who are not well trained, who take shortcuts when performing their work, or whose intentions are contrary to personal and public good.

Positive attitudes displayed by the country toward test use. Prevailing attitudes in the United States generally favor meritocracy over egalitarianism (e.g., providing resources to people on the basis of objective standards rather than equally to everyone) and are based on valuing individualism over collectivism (e.g., decisions should be based on knowledge of an individual's relevant personal qualities, not on a person's race, socioeconomic status, gender, family name, or other group characteristic). These and similar positive attitudes create a climate in the United States that is favorable to test development and use.

Although prevailing attitudes have generally fostered a positive climate for test development and use, concerns remain. Some people question the overall value of clinical services, including testing services, to children. For example, parents may refuse to have their children tested individually in school, given their belief that such information may be used in ways that will not serve their children's best interests.

Children and people of other groups that have been abused or otherwise marginalized (e.g., people with disabilities or in foster care) constitute legally protected classes. The ability of these groups to advocate for themselves may be limited. Thus, adults often take a special interest in members of these legally protected groups and may be skeptical about whether testing or the use of test results is in their best interests. Blacks and other minority groups also constitute a protected class that has not historically fared well in the testing environment.

During the 1970s, tests fell into some disfavor in the United States, in part because of charges that tests discriminated against Blacks (Oakland, 1977). Test use with Blacks and other minorities faced legal challenges (Oakland & Gallegos, 2005). Issues addressed by the courts included whether tests limit Blacks' educational and vocational opportunities and whether tests are sufficiently reliable or valid for use with Black children, especially intelligence tests. Given that tests of cognitive ability often yield higher mean scores for some racial and socioeconomic groups and lower mean scores for other groups,

some explained measured differences as resulting from test bias. The origin of test bias was attributed to the assumption that tests were developed by and reflect the values of White middle-class people who knew or possibly cared little about minority group lifestyles. Additionally, tests are intended to statistically discriminate between people on the basis of the amount of the attribute assessed, which bothers those who believe everyone is or should be equal (i.e., an egalitarian philosophy). Thus, test practices used with Black children were thought to deserve special scrutiny.

The National Center for Fair and Open Testing (http://www.FairTest.org) is the leading test critic in the United States. Over the past 30 years, the center and other groups have charged that

- test data do not improve decision making;
- testing unnecessarily invades one's privacy;
- tests are so flawed that laws should prohibit their use;
- too much time and money are spent on testing;
- people easily cheat on tests and fake their scores, thereby diminishing the tests' utility;
- only institutions, not people, benefit from test use;
- human behavior is too complex to assess it accurately;
- some complex decisions (e.g., admission to colleges) are made only on the basis of test scores;
- ability tests do not assess higher order cognitive abilities;
- multiple-choice tests reduce the promotion of creativity and deep thinking;
- teachers' grades provide a more accurate evaluation of student performance than tests,
- tests provide information that can be readily obtained from other, more reliable sources;
- achievement tests assess qualities unrelated to what children actually learn in school;
- the impact of test use is too pervasive and results in educators tailoring curricula in light of year-end tests; and
- important fundamental qualities (e.g., character, creativity, curiosity, friendliness, kindness, loyalty, emotional intelligence, and obedience) are underassessed and thus overlooked and undervalued.

Phelps's (2009) *Correcting Fallacies About Educational and Psychological Testing* addressed allegations that test use with children and youths does not serve the public interest. Phelps and Gottfredson (2009) attributed opposition to testing as mainly the result of those individuals or groups who make unsubstantiated or false claims (i.e., claims that lack scientific evidence) that are then disseminated by the media and thus accorded unwarranted veracity. Phelps marshaled considerable evidence that supports the value of test use with children and thus dismissed the validity of most of the claims promulgated by the National Center for Fair and Open Testing.

Test data can have various levels of impact on society. For example, some tests (e.g., teacher-made achievement tests) will generally have little permanent effect on an individual or society. Teacher-made tests are referred to as *low-stakes tests*. In contrast, some high-stakes tests (e.g., those used for college admissions, to certify professional competencies, or to diagnose psychiatric disorders) generally have a larger and more permanent influence on people's lives. Public attitudes toward test use may differ considerably depending on whether the data are used for low- or high-stakes decisions. Generally, greater opposition to testing arises when test data are used to make high-stakes decisions.

Public attitudes tend to serve as a bellwether of the degree of acceptance of professional services, including testing services. Although public attitudes toward test use are not uniformly positive in the United States, people often see value in the testing enterprise and believe that tests are generally accurate, objective, and fair. Objective tests are also commonly accepted to provide a standardized and efficient way to collect and interpret useful information, thereby avoiding the subjective biases that often occur when decisions are based on informal information (Phelps, 2009).

National legislation may also be used as a bellwether of the public's general acceptance of professional services. Legislation generally mandates only those services seen as generally serving the public interest. Legislation governing education and services to those with special needs, along with application for social security benefits, typically mandates the use of tests, thus providing support

for the belief that proper test use generally serves the public interest.

A national population of sufficient size to support such an industry. The development of any product, including tests, requires a sufficiently large and stable market for its purchase and application. Given that tests are typically developed for use in one country, that country's population should be sufficiently large and stable to warrant the cost of test development. Many countries are too small or experience dramatic population changes, perhaps even declines, that preclude the development of a viable testing industry. The United States is sufficiently large and growing to support a large and vibrant testing industry. In fact, it is generally recognized that the United States is the testing industry's international leader.

An industry responsible for test developing and marketing tests. Toward the beginning of the 20th century, psychologists had no publishers available to develop or market their test products. Believing a need and market existed for their work, James McKeen Cattell and two of his former students, Robert Woodworth and Edward Thorndike, founded the Psychological Corporation in New York in 1921. This corporation grew to become one of the largest and preeminent test publishers in the world (Sokal, 1981). The sale of the Psychological Corporation in 2007 for almost $1 billion underscored its then current and future value.

Efforts to develop and market tests have resulted in a large and dynamic testing industry in the United States, one sufficiently large to warrant the formation of the Association of Test Publishers (http://www.testpublishers.org/mc) to represent industry interests. Members of the Association of Test Publishers include more than 100 U.S. corporations and many publishers from other countries. Some of the larger corporate members associated with the practice of psychological and educational testing include American College Testing, Consulting Psychologists Press, California Test Bureau/McGraw-Hill, College Board, Educational Testing Service, Multi-Health Systems, Psychological Assessment Resources, Pearson Assessment, Pro-Ed, Psychological Corporation/Harcourt Assessment, Riverside Publishing, and Western Psychological Services. Competition among the somewhat large number of testing companies has helped in the creation and marketing of thousands of tests. Competition may be lessening because of mergers and acquisitions (i.e., the purchase of one company by another company or organization); several of the largest test companies have merged during the past 2 decades. Such mergers may foretell the development of fewer new tests and greater reliance on revising existing tests.

Psychologists in the United States generally respect copyright laws by using test materials purchased from test publishers rather than relying on photocopied test materials. These practices demonstrate support of the testing industry by compensating test authors and companies for their intellectual and commercial work. Unfortunately, the illegal and often unethical practice of photocopying or in other ways reproducing test materials is common in many emerging countries, thus limiting the development and growth of test development in these countries.

The growth of the U.S. testing industry reflects the generally positive attitudes toward and frequency of test use in this country. The testing industry serves three broad professional markets: those working in commerce and industry (e.g., human resources specialists, industrial and organizational psychologists), those working in education (e.g., school administrators, educators, counselors, school psychologists), and those providing clinical services (e.g., clinical and counseling psychologists, occupational and physical therapists, speech pathologists, and rehabilitation specialists, some of whom also work in educational settings). The first and second markets are particularly strong in the United States. The number of standardized tests available in English is in the thousands. The number of tests assessing diverse constructs, abilities, and behaviors for children and youths listed in U.S. test publishers' catalogs exceeds 300. Thus, the number of tests for use with children and youths seems adequate in light of current needs.

Universities that offer programs that teach students to develop tests. In the United States, students can acquire a specialization in test development and use at the graduate level. The country has a number of

educational programs to prepare psychometricians and others to develop tests, typically at the doctoral level as well as hundreds of programs to prepare specialists to use tests, typically at the master's, specialist, or doctoral levels. Test users typically carry the title of school counselor, psychologist (with specializations in clinical, counseling, neuro-, or school psychology), physical and occupational specialist, and speech–language pathologist. Doctoral-level psychometrics students are generally well prepared in basic measurement and advanced statistics (Rossen & Oakland, 2008).

A further note on academic attainment among U.S. children. As previously noted, PISA data from the Organisation for Economic Co-operation and Development (i.e., industrialized) countries showed that U.S. students performed 15th among 27 countries in reading, 21st among 30 countries in science, and 24th among 29 countries in math. Another international study that compared U.S. fourth- and eighth-grade students with same-grade peers from a larger number of countries, many of which were less developed, showed that U.S. students performed about average in science and lower in math (International Association for the Evaluation of Educational Achievement, 2003). Compared with their international peers, U.S. fourth graders ranked 13th in mathematics achievement among 24 participating countries, and U.S. eighth graders ranked 25th among 44 participating countries. Compared with their international peers, U.S. fourth-grade students ranked 17th in science achievement among 24 participating countries, and eighth-grade students ranked 33rd among 44 participating countries.

Since the 1970s, numerous school reform efforts at the local, state, and national levels have attempted to improve achievement, especially among low-achieving students. These efforts have generally not resulted in higher achievement. Data from the U.S. Department of Education's National Assessment of Educational Progress have revealed that students' achievement is linked consistently and strongly with their family's economic status as reflected in whether they are eligible for free or reduced-price lunches or are ineligible for lunch subsidies (common indicators of socioeconomic status). Students who qualify for free lunch generally perform lower in reading, math, and science than those who qualify for reduced-price lunch, who in turn perform lower than those who pay full price for lunch. These differences occur despite increased educational resources directed to students in poverty over the past several decades. Among students receiving either free or reduced-price lunch, numerically most are White. However, proportionate to their representation in the general population, the largest proportion of students receiving free or reduced-price lunch are Black and Hispanic—the segment of the U.S. population expected to increase the most.

Challenges for those conducting assessment services cross-culturally. The United States faces many challenges with respect to test use. These challenges include developing closer links between assessment and intervention, instituting methods that inform parents about their children's educational development in ways that lead to more shared responsibility for educating them, improving testing resources for English language learners, and limiting the costs associated with assessment.

Tests constitute a technology that should be used as long as it serves important personal, institutional, or social needs. Tests generally pass this test. Changing conditions within the United States can result in a call for changes in testing services. For example, the need for educational reforms in the United States remains urgent. Those engaged in this reform effort are interested in developing and implementing assessment services that better assist in assessing and guiding academic and behavioral development among students. Attention to the needs of low-achieving and poorly performing students is most urgent. The following two efforts are consistent with efforts to link assessment and interventions.

Curriculum-based measurement. Curriculum-based measurement uses frequent (e.g., at least weekly) and continuous measurement of attainment of desired achievement in light of the curriculum used with the student (Deno, 1985; Shinn, 1989). The focus in curriculum-based measurement is on the degree to which a student is progressing through the curriculum, not how one student compares with his or her peers, and

on linking assessment and instruction directly. Curriculum-based instruction has several advantages over instruction based on norm-referenced assessment: Curriculum-based measures focus on the curricular materials used in the student's classroom, assessment data have direct instructional implications, repeated measurement monitors the student's attainment and retention of achievement, assessment data can be charted and are sensitive to change, and curriculum-based tests avoid norm-referenced comparisons and racial, ethnic, and socioeconomic status contrasts.

Curriculum-based measurement is commonly used in response-to-intervention efforts. Response-to-intervention strategies integrate assessment and intervention to maximize student achievement and reduce behavior problems. Assessment practices in a response-to-intervention framework help identify students at risk for lower learning outcomes or behavior problems, monitor student progress, provide evidence-based interventions, and adjust the intensity and nature of the interventions in light of a student's progress. See the National Center on Response to Intervention's website (http://www.rti4success.org) for more details on response to intervention.

Authentic assessment. *Authentic assessment* is "the systematic recording of developmental observations over time about the naturally occurring behaviors and functional competencies of young children in daily routines by familiar and knowledgeable caregivers in the child's life" (Bagnato & Yeh Ho, 2006, p. 16). In contrast to curriculum-based measurement that focuses on classroom performance, authentic assessment focuses on the naturally occurring behaviors of young children as seen by their parents or other caregivers. Bagnato, Neisworth, and Pretti-Frontczak (2010) identified eight overarching standards for developmentally appropriate authentic assessment materials and practices. Their functional implications are evident: Measures that support authentic assessment must (a) be acceptable to those who use them (e.g., parents and other caregivers view the measures as having social validity and social worth), (b) be authentic (e.g., the measures sample naturally occurring behaviors evidenced in daily situations), (c) foster collaboration (e.g., parent–professional and interdisciplinary teamwork), (d) be evidence based (e.g., the test materials are designed, developed, and field validated for young children, especially those with special needs), (e) be multifactored (e.g., data are collected using multiple methods from multiple sources in reference to behaviors displayed in various naturally occurring settings), (f) be sufficiently sensitive to change (e.g., items are sequentially arranged and sufficiently dense to provide a graduated scoring of young children's performance), (g) reflect universality (e.g., allow for the identification of both underlying strengths and needs), and (h) offer utility (e.g., provide data in sufficient detail to lead to the identification of evidence-based interventions).

The implementation of curriculum-based and authentic assessment methods is consistent with efforts to implement a somewhat new wave of assessment services that might better assist efforts to promote children's academic and behavioral development, not merely compare them with peers. A number of tests are being developed or revised to provide suggestions on how to use test data at the item and subtest levels to promote children's development in adaptive behavior, cognitive, and social–emotional development. These efforts are consistent with efforts to develop direct links between assessment and intervention and thus are commendable and should continue.

Consider the needs of parents as important educators. The degree of parent support for and involvement in their children's development, including education, may constitute the single most important social ingredient influencing their children's success. Many parents assume an active role as educators early in their children's lives (e.g., teaching basic number and reading skills as well as developing their children's elaborative language and grammar skills before school entrance). Such parents often remain vigilant to needed support throughout their children's schooling. In contrast, other parents provide little support or involvement from the onset or later. Parents who are minimally involved in their children's education may assume that families are responsible for socializing their children and that schools are responsible for educating them.

Efforts are needed to revise the assessment process in the United States to better support parents who are involved in their children's education and to encourage and aid parents who are minimally involved. Unlike those U.S. parents who are not actively involved in educating their children, the Maoris in New Zealand realize as a society that they, not their children's teachers or other professionals, are ultimately responsible for educating their children, and thus Maori parents feel a need to be instrumentally involved in their children's development. Professionals in New Zealand are to assist and serve as consultant to parents (Annan, 2010).

Address the assessment needs of English language learners. A large percentage of U.S. children speak a first language other than English. The testing industry has responded to this demographic by developing tests, largely measures of intelligence, that are either published in Spanish or involve nonverbal administrative methods. The number and type of tests for English language learners should be increased and improved. Additionally, the United States has a large and growing minority student population, yet has limited numbers of assessment specialists from minority communities, including those who speak a second language. Continued efforts to prepare more testing specialists from these communities, especially those who are bilingual, are needed.

Align assessment practices with standards. The profession of psychology in the United States has well-established and respected standards for developing and using tests (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) as well as ethical standards governing test use (American Psychological Association, 2010). Although U.S. psychologists are committed to these practice standards, the policies and practices of educational institutions and other organizations (e.g., insurance companies and other third-party payers) often view assessment practices as much by financial considerations as by professional standards.

Psychologists providing high-stakes assessment practices in schools and elsewhere generally strive to be comprehensive, which results in time-consuming and expensive evaluations. These detailed assessment practices generally reflect the perceived importance of assessing multiple traits and abilities with various assessment methods that use information from various sources in an effort to describe and understand behavior in multiple settings and over time. These principles are often under attack by those organizations that control or influence the finances that govern payment for such services (e.g., insurance companies, school districts, parents). Assessment specialists are increasingly encouraged to spend less time testing, to use short-form tests whenever possible, or to provide their services pro bono. Assessment specialists are also encouraged to use computer-generated reports and standard templates instead of relying on more traditional comprehensive and individually tailored reports.

CONCLUSIONS

The development and use of standardized tests may constitute psychology's most important technical contribution to the behavioral sciences. The value of tests, when properly used, lies in their provision of objective, standardized, valid, reliable, and efficient methods to inform decision making. Test data can assist professionals and those with whom they work to describe current level of functioning, estimate future functioning, assist guidance and counseling services, help establish intervention methods, evaluate progress, screen for special needs, diagnose disabling disorders, help place people in jobs or programs, assist in determining whether people should be credentialed, admitted or employed, retained, or promoted for administrative and planning purposes as well as for the conduct of research.

Despite the advantages of using tests, test development and use have various inherent limitations. For example, information from standardized tests is never perfectly reliable or valid. Thus, high-stakes decisions may be improved through the use of other highly reliable and valid sources of supplemental information. Over time, individual tests become outdated and need regular and expensive revision and renorming. In some situations, assessed test content may be insufficient or unsuitable to ade-

quately reflect desired constructs and traits. Although the testing industry is sometimes well regulated, in some locales tests may be administered by people who have little knowledge of proper test use and interpretations. Thus, although tests may have considerable value, professionals must remain vigilant to ensure that the test's technical qualities and the personnel who use tests are adequate.

Test development and use, to be relevant, must change with changing national conditions. For example, since apartheid ended, leaders in South Africa have attempted to develop assessment models and methods consistent with the prevailing South African concept of *ubuntu* to better reflect a culture of sharing. The combined use of qualitative and quantitative methods is often thought to be superior to the use of quantitative methods only in most countries.

Leaders in Brazil recently established a national infrastructure in which tests are reviewed for their adequacy. A number of Brazilian universities have established programs that prepare professionals to be psychometricians and test users as well as to develop tests. These somewhat recent actions occurred in response to Brazil's need for and reliance on testing. Leaders in the United States continue to use many existing tests while adding tests and revising assessment processes intended to have more practical applications that help promote child growth and development, including education. Thus, to remain relevant, test services must strive to serve the public good rather than expect the public to accommodate to its existing services.

Testing practices between countries will and should differ. Differences in their histories, populations, national goals and priorities, resources, and other qualities create conditions that require different needs for tests. For example, the desire for tests to help identify and respond to children's special education needs will be more important in countries that offer special education and related services than in countries that are merely struggling to provide universal general education. Testing practices also differ according to the degree that a country's population is multicultural and multilingual. Professionals attempting to align tests with a somewhat large national population with widely varying multicultural and multilingual characteristics, such as in South Africa, face serious challenges. Such alignment may not be possible with existing or foreseeable resources.

The primary conditions that influence testing in South Africa, Brazil, and the United States also differ. These influences include their unique histories, the maturity of the national testing infrastructure, the size and proportion of the population that differs by race and socioeconomic status, the number of languages commonly spoken, and the number of tests for use with children and youths. Other conditions are similar across these three countries, including national professional associations promoting the development of tests and their wise use, a current or emerging infrastructure that supports test development and use, attempts to respond to socioeconomic and racial differences that may affect test use, and efforts to align test use with important national issues, especially those that involve student development, including education.

The science and art of test development and use began to emerge somewhat strongly and widely in some Western countries during the late 1940s. Thus, these efforts are relatively young, less than 70 years old in their current and widely used applied form. As noted in the discussion of test use in South Africa, Western-centric models and methods, although used in other countries, may be unsuitable.

New assessment models and measures are needed in all three countries. New tests will be developed, in part, in response to changing national needs and desires. For example, emerging countries will increasingly develop their own models and measures to reflect their local needs and conditions. The outcomes of the efforts made in emerging counties may lead to improvements in ways to develop, norm, standardize, and validate tests nationally, regionally, and internationally.

Many issues that affect the initial or continued development of a testing infrastructure in one country are common to other countries. Somewhat common cross-national issues warrant the attention of national (e.g., British Psychological Society, American Psychological Association) and international (e.g., International Union of Psychological Sciences, International Association of Applied Psychology) professional associations that have expressed an

abiding interest in testing. The continued efforts by the ITC to address common and cross-national issues can lead to substantial improvements in test development and use, especially in countries that initially lack resources to address them. This belief is validated, in part, by the sizable influence the ITC's guidelines have had on Brazil's developing testing infrastructure. The ITC's efforts to accredit preparation programs that prepare professionals to engage in test development and to develop additional guidelines could influence future developments in a positive manner.

References

Alves, I. C. B. (2009). Reflexões sobre o ensino da avaliação psicológica na formação do psicólogo [Reflections on psychological assessment teaching for psychologists]. In C. S. Hutz (Org.), *Avanços e polêmicas em avaliação psicológica* (pp. 217–246). São Paulo, Brazil: Casa do Psicologo

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2002, Amended June 1, 2010)*. Retrieved from http://www.apa.org/ethics/code/index.aspx

Anache, A. A., & Correa, F. B. (2010). As politicas do Conselho Federal de Psicologia para a avaliação psicológica [Federal Council of Psychology's policies for psychological assessment]. In Conselho Federal de Psicologia (Eds.), *Avaliação psicologia: Diretrizes na regulamentação da profissão* (pp. 19–30). São Paulo, Brazil: Conselho Federal de Psicologia.

Angelini, A. L. (1995). Abertura do I Encontro de Técnicas de Exame Psicológico, Ensino, Pesquisa e Aplicações [Lecture at the 1st Convention of Psychological Examination: Teaching, Research and Applications]. *Boletín de Psicología*, *45*, 9–18.

Annan, J. (2010). Test use by educational psychologists in New Zealand. *International Association of Applied Psychologists Bulletin, 22*(2–3), 15–17.

Annual Report of the Psychological Society of South Africa. (2007). Retrieved from http://64.233.183.104/search?=cache:MP-gtrRiDVY:www.psyss.com/documents/AnnualReportfinal.pdf

Bagnato, S., Neisworth, J., & Pretti-Frontczak, K. (2010). *LINKing authentic assessment and early childhood interventions: Best measures for best practices* (2nd ed.). Baltimore, MD: Brookes.

Bagnato, S., & Yeh Ho, H. (2006). High stakes testing with preschool children. *KEDI International Journal of Educational Policy, 3*, 23–43.

Bartram, D., & Hambleton, R. (2006). *Computer-based testing and the internet: Issues and advances.* Hoboken, NJ: Wiley.

Beery, K. E., Buktenica, N. A., & Beery, N. A. (2010). *Beery-Buktenica Developmental Test of Visual–Motor Integration—Sixth Edition.* San Antonio, TX: Pearson.

Bender, L. (1938). *A visual-motor Gestalt test and its clinical use* (Research Monograph No. 3). New York, NY: American Orthopsychiatric Association.

Berry, J. W. (1985). Learning mathematics in a second language: Some cross-cultural issues. *For the Learning of Mathematics, 5*(2), 18–23.

Brannigan, G. G., & Decker, S. L. (2006). *Bender Visual Motor Gestalt Test—Second Edition.* Rolling Meadows, IL: Riverside.

Byrne, B. M., Oakland, T., Leong, F. T. L., van de Vijver, F. J. R., Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3*, 94–105. doi:10.1037/a0014516

Central Intelligence Agency. (n.d.) *The world factbook.* Washington, DC: Author. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/index.html

Claassen, N. C. W. (1995, September). *Cross-cultural assessment in the human sciences.* Paper presented at the Human Sciences Research Council, Pretoria, South Africa.

Claassen, N. C. W., & Schepers, J. M. (1990). Group differences in academic intelligence based on differences in socio-economic status. *South African Journal of Psychology, 20*, 294–302.

Conselho Federal de Psicologia. (2003). *Caderno especial de resoluções: Resolução CFP002/2003* [Special compendium of legal decisions: Resolution CFP 002/2003]. Brasília, Brazil: Author.

Conselho Federal de Psicologia. (2010). *Avaliação Psicológica: Diretrizes na regulamentação da profissão* [Psychological assessment: Guidelines for professional regulation]. Brasília, Brazil: Author.

De Barros, R. P., Henriques, R., & Mendonça, R. (2000). Education and equitable economic development. *Economia, 1*, 111–144.

De Beer, M. (2006). Dynamic testing: Practical solutions to some concerns. *South African Journal of Industrial Psychology*, *32*(4), 8–14.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.

Eloff, I., Maree, J. G., & Ebersöhn, L. E. (2006). Some thoughts on the perceptions on the role of educational psychologists in early childhood intervention. *Early Child Development and Care, 176*, 111–127. doi:10.1080/03004430500209522

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341–349. doi:10.1037/1040-3590.8.4.341

Fouché, F., & Alberts, N. F. (1986). *Manual for the 19 Field Interest Inventory (19 FII)*. Pretoria, South Africa: Human Sciences Research Council.

Foxcroft, C., Paterson, H., Le Roux, N., & Herbst, D. (2004). *Psychological assessment in South Africa: A needs analysis*. Pretoria, South Africa: Human Sciences Research Council.

Foxcroft, C. D. (1997). Psychological testing in South Africa: Perspectives regarding ethical and fair practices. *European Journal of Psychological Assessment, 13*, 229–235. doi:10.1027/1015-5759.13.3.229

Foxcroft, C. D. (2004). Planning a psychological test in the multicultural South African context. *South African Journal of Industrial Psychology, 30*(4), 8–15.

Hambleton, R., Bartram, D., & Oakland, T. (2011). Technical advances and guidelines for improving testing practices. In P. Martin, F. Cheung, M. Kyrios, L. Littlefield, M. Knowles, B. Overmier, & J. M. Prieto (Eds.), *The IAAP handbook of applied psychology* (pp. 338–361). London, England: Blackwell.

Hambleton, R., & Oakland, T. (2004). Advances, issues, and research in testing practices around the world. *Applied Psychology, 53*, 155–156. doi:10.1111/j.1464-0597.2004.00165.x

Hanna, E. I. (2005). *Inclusive design for maximum accessibility: A practical approach to universal design* (PEM Research Report 05–04). San Antonio, TX: Pearson Education Measurement

Hu, S., & Oakland, T. (1991). Global and regional perspectives on testing children and youth: An international survey. *International Journal of Psychology, 26*, 329–344. doi:10.1080/00207599108246857

Hutz, C. S. (Ed.). (2009). *Avanços e polêmicas em avaliação psicológica* [Questions and progress in psychological assessment]. São Paulo, Brazil: Casa do Psicólogo.

Hutz, C. S. (Ed.). (2010). *Avanços em avaliação psicológica e neuropsicológica decrianças e adolescentes* [Progress in psychological and neuropsychological assessment for children and adolescents]. São Paulo, Brazil: Casa do Psicólogo.

Instituto Brasileiro de Avaliação Psicológica. *Em defesa da avaliação psicológica: Manifesto* [In defense of psychological assessment: Petition]. Retrieved from http://www.ibapnet.org.br/avalpsi_manifesto.html

Instituto Brasileiro de Geografia e Estatística. (2008). *Educação melhora, mas ainda apresenta desafios* [Education is improving but still there are challenges to overcome]. Retrieved from http://www.ibge.gov.br/home/presidencia/noticias/noticia_visualiza.php?id_noticia=123

Instituto Brasileiro de Geografia e Estatística. (2009a). *IBGE: Desigualdade social cai, mas de forma lenta* [IBGE: Social inequalities drop, but still at slow pace]. Retrieved from http://www.ibge.gov.br/home/presidencia/noticias

Instituto Brasileiro de Geografia e Estatística. (2009b). *Síntese de indicadores sociais: Uma análise das condições de vida da população brasileira* [Synthesis of social indicators: Analysis of Brazilians' living conditions]. Rio de Janeiro, Brazil: Author.

Instituto Brasileiro de Geografia e Estatística. (2010). *Censo demográfico 2010* [Demographic census 2010]. Retrieved from http://www.ibge.gov.br/home/presidencia/noticias/noticia_visualiza.php?id_noticia=1766

Instituto de Pesquisa Econômica Aplicada. (2008). *PNDA 2007—Primeiras análises. educação, juventude, raça/cor* [PNDA 2007—First analysis: Education, youth, race/color] (Technical Report v. 4). Brasília, Brazil: Author.

Instituto Nacional de Estudos e Pesquisas Anísio Teixeira. (2007). *O que é PISA* [What is PISA]. Retrieved from http://www.inep.gov.br/imprensa/noticias/internacional/news07_05.htm

Instituto Nacional de Estudos e Pesquisas Anísio Teixeira. (2012). *ENEM 2012—Passo a passo* [ENEM 2012—Basic steps]. Retrieved from http://enem.inep.gov.br

International Association for the Evaluation of Educational Achievement. (2003). *Highlights from the trends in international mathematics and science study*. Washington, DC: National Center for Educational Statistics.

International Test Commission. (2000). *International guidelines for test use*. Retrieved from http://www.intestcom.org/guidelines/index.php

International Test Commission. (2005). *International guidelines on computer-based and Internet delivered testing*. Retrieved from http://www.intestcom.org/guidelines/index.php

Languages of the United States. (n.d.). Retrieved September 20, 2010, from http://en.wikipedia.org/wiki/Languages_of_the_United_States

Lawrence, G. (1982). *People types and tiger stripes: A practical guide to learning styles*. Gainesville, FL: Center for the Application of Psychological Types.

Leach, M., & Oakland, T. (2007). Ethics standards impacting test development and use: A review of 31 ethics codes impacting practices in 35

countries. *International Journal of Testing, 7*, 71–88. doi:10.1080/15305050709336859

Leach, M., & Oakland, T. (in press). Displaying ethical behaviors by psychologists when standards are unclear. In M. M. Leach, M. J. Stevens, A. Ferrero, & Y. Korkut (Eds.), *Oxford handbook of international psychological ethics*. New York, NY: Oxford University Press.

Maree, J. G. (2010a). Brief overview of the advancement of postmodern approaches to career counseling. *Journal of Psychology in Africa*, *20*, 361–368.

Maree, J. G. (2010b). *The Career Interest Profile (Version 3)*. Randburg, South Africa: Jopie van Rooyen.

Maree, J. G. (2010c). Career-story interviewing using the three anecdotes technique. *Journal of Psychology in Africa, 20*, 369–380.

Ministério da Educação e Cultura. (2008). *National report from Brazil* (UNESCO technical report). Brasilia, Brazil: Author.

Ministério da Educação e Cultura. (2010). *Políticas publicas pretendem transformar o ensino médio* [Public policies may change middle schools]. Retrieved from http://portal.mec.gov.br/index.php?option=com_content&view=article&id=15657

Motshekga, A. (2010, July). *Statement by the Minister of Basic Education, Mrs. Angie Motshekga, MP on the progress of the review of the National Curriculum Statement, Tuesday 06 July 2010.* Cape Town, South Africa: Ministry of Basic Education.

National Center for Education Statistics. (n.d.). *Program for International Student Assessment (PISA)*. Retrieved from http://nces.ed.gov/surveys/pisa

National Department of Education. (2008). *National strategy on screening, identification and support.* Pretoria, South Africa: Government Printers.

Nell, V. (1990). One world, one psychology: "Relevance" and ethnopsychology. *South African Journal of Psychology, 20*, 129–140.

Nell, V. (1994). Interpretation and misinterpretation of the South African Wechsler-Bellevue Adult Intelligence Scale: A history and a prospectus. *South African Journal of Psychology, 24,* 100–109.

Noronha, A. P. (2002). Os problemas mais graves e mais frequentes no uso dos testes psicológicos [The most serious and frequent problems on using psychological tests]. *Psicologia: Reflexão e Crítica, 15*, 135–142. doi:10.1590/S0102-79722002000100015

Oakland, T. (Ed.). (1977). *Psychological and educational assessment of minority children.* Larchmont, NY: Brunner/Mazel.

Oakland, T. (2004). Use of educational and psychological tests internationally. *Applied Psychology, 53*, 157–172. doi:10.1111/j.1464-0597.2004.00166.x

Oakland, T. (2005). What is multicultural school psychology? In C. Frisby & C. Reynolds (Eds.), *Comprehensive handbook of multicultural school psychology* (pp. 3–13). New York, NY: Wiley.

Oakland, T. (2009). How universal are test development and use? In E. Grigorenko (Ed.), *Assessment of abilities and competencies in an era of globalization* (pp. 1–40). New York, NY: Springer.

Oakland, T., & Gallegos, E. (2005). Legal issues associated with the education of children from multicultural settings. In C. Frisby & C. Reynolds (Eds.), *Comprehensive handbook of multicultural school psychology* (pp. 1048–1080). New York, NY: Wiley.

Oakland, T., & Hu, S. (1991). Professionals who administer tests with children and youth: An international survey. *Journal of Psychoeducational Assessment, 9*, 108–120. doi:10.1177/073428299100900201

Oakland, T., & Hu, S. (1992). The top ten tests used with children and youth worldwide. *Bulletin of the International Test Commission, 19*, 99–120.

Oakland, T., & Hu, S. (1993). International perspectives on tests used with children and youth. *Journal of School Psychology, 31*, 501–517. doi:10.1016/0022-4405(93)90034-G

Oakland, T., Leach, M. M., Bartram, D., Lindsay, G., Smedler, A.-C., & Zhang, H. (2012). An international perspective on ethics codes in psychology. In M. M. Leach, M. Stevens, A. Ferrero, & Y. Korkut (Eds.), *Oxford handbook of international psychological ethics* (pp. 19–27). New York, NY: Oxford University Press.

Owen, K. (1998). *The role of psychological tests in education in South Africa: Issues, controversies, and benefits.* Pretoria, South Africa: Human Sciences Research Council.

Owen, O. (1991). Test bias: The validity of the Junior Aptitude Tests for various population groups in South Africa regarding the constructs measured. *South African Journal of Psychology, 21*, 112–118.

Painter, D., & Terre Blanche, M. (2004). Critical psychology in South Africa: Looking back and looking ahead. *South African Journal of Psychology, 34*, 520–543.

Pasquali, L. (Org.). (2010). *Instrumentação psicológica: Fundamentos e práticas* [Psychological instruments: Foundations and practical issues]. Porto Alegre, Brazil: Artes Médicas.

Paterson, H., & Uys, K. (2005). Critical issues in psychological test use in the South African workplace. *South Africa Journal of Industrial Psychology, 31*(3), 12–22.

Penna, A. G. (2004). Breve contribuição à história da Psicologia aplicada ao trabalho [Brief contribution to the history of psychology applied to work]. *Mnemosine, 1*, 1–2.

Pessoti, I. (1988). Notas para uma historia da psicologia brasileira [Notes on the history of Brazilian psychology]. In Conselho Federal de Psicologia (Ed.), *Quem*

é opsicólogo brasileiro [Who is the Brazilian psychologist] (pp. 17–31). São Paulo, Brazil: Edicon.

Pfromm Netto, S. (1996). Pioneiros da psicologia escolar: Mira y Lópes (1886–1996) [School psychology pioneers: Myra y Lópes: 1886–1996]. *Psicologia Escolar e Educacional, 1*, 87–88. doi:10.1590/S1413-85571996000100015

Phelps, R. (Ed.). (2009). *Correcting fallacies about educational and psychological testing*. Washington, DC: American Psychological Association. doi:10.1037/11861-000

Phelps, R., & Gottfredson, L. (2009). Summary and discussion. In R. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 247–255). Washington, DC: American Psychological Association. doi:10.1037/11861-008

Prette, Z. A. P. (2008). *Psicologia escolar e educacional: Saúde e qualidade de vida* [School and educational psychology: Health and life quality]. Campinas, Brazil: Editora Alínea.

Professional Board for Psychology. (2006). *Rules of conduct pertaining specifically to the profession of psychology*. Pretoria, South Africa: Government Gazette.

Rainbow Nation. (n.d.). Retrieved from http://www.southafricaataglance.com/south-africa-rainbow-nation.html

Rossen, E., & Oakland, T. (2008). Graduate preparation in research methods: The current status of APA-accredited professional programs in psychology. *Training and Education in Professional Psychology, 2*, 42–49. doi:10.1037/1931-3918.2.1.42

Ruane, I. (2010). Obstacles to the utilization of psychological resources in a South African township community. *South African Journal of Psychology, 40*, 214–225.

Sehlapelo, M., & Terre Blanche, M. (1996). Psychometric testing in South Africa: Views from above and below. *Psychology in Society*, 21, 49–59.

Shinn, M. (1989). *Curriculum-based measurement: Assessing special children.* New York, NY: Guilford Press.

Sokal, M. M. (1981). The origins of the Psychological Corporation. *Journal of the History of the Behavioral Sciences, 17*, 54–67. doi:10.1002/1520-6696(198101)17:1<54::AID-JHBS2300170108>3.0.CO;2-R

South Africa Department of Health. (2010). *Regulations defining the scope of practice of practitioners of the profession of psychology*. Pretoria, South Africa: Government Printers.

Statistics South Africa. (2009). *Stats in brief.* Pretoria, South Africa: Author.

Stead, G. B. (2002). The transformation of psychology in post-apartheid South Africa: An overview. *International Journal of Group Tensions, 31*, 79–102. doi:10.1023/A:1014216801376

United Nations, Department of Economic and Social Affairs. (2009). World population prospects: The 2008 revision. *Population Newsletter, 87*, 1–4. Retrieved from http://www.un.org/esa/population/publications/popnews/Newsltr_87.pdf

U.S. Census Bureau. (2010). *American fact finder*. Retrieved from http://2010.census.gov/2010census/

Vandeyar, S. (2008). Shifting selves: The emergence of new identities in South African schools. *International Journal of Educational Development, 28*, 286–299. doi:10.1016/j.ijedudev.2007.05.001

Van Eeden, R. (1991). *Manual for the Senior South African Individual Scale—Revised (SSAIS–R): Part 1*. Pretoria, South Africa: Human Sciences Research Council.

Watson, M. B., & Fouche, P. (2007). Transforming a past into a future: Counseling psychology in South Africa. *Applied Psychology, 56*, 152–164.

Wechsler, S. M. (1996). Entrevista com a fundadora da Abrapee [Interview with Abrapee's founder]. *Psicologia Escolar e Educacional, 1*, 1–2.

Wechsler, S. M. (2001). Avaliação psicológica no Brasil; tendências e perspectivas no novo milênio [Psychological assessment in Brazil: Tendencies and perspectives for the new century]. In Conselho Regional de Psicologia (Ed.), *A diversidade da avaliação psicológica: Consideraçoes teóricas e práticas* (pp. 17–24). João Pessoa, Brazil: Editora Ideia.

Wechsler, S. M. (2009). Impact of test development movement in Brazil. *World*Go*Round, 36*(2), 7–8.

Wechsler, S. M. (2010, July). *The impact of cultural characteristics on test use.* Paper presented at the International Testing Commission Conference, Hong Kong, China.

CHAPTER 12

LEGAL ISSUES IN SCHOOL PSYCHOLOGICAL ASSESSMENTS

Matthew K. Burns, David C. Parker, and Susan Jacob

When Binet and Simon published their scale for measuring the intelligence of schoolchildren in 1905, they gave psychology new credibility within the science community and gave birth to psychoeducational assessment. School psychology traces its roots to before mental ability tests were developed (Fagan & Wise, 2007), but assessment has become a defining attribute of the field. The scientific method is the conceptual framework from which school psychology operates, and data-based decision making—an extension of the scientific method—is a foundational competency for school psychologists (Ysseldyke et al., 2006). Psychoeducational assessment of schoolchildren has changed considerably since Binet and Simon conducted their work, as have the potential uses of the assessment results.

The newly revised code of ethics of the National Association of School Psychologists (NASP; 2010) states, "School psychologists are committed to the application of their professional expertise for the purpose of promoting improvement in the quality of life for students, families, and school communities" (p. 2). School-based practitioners must be knowledgeable of legal and ethical issues associated with psychoeducational assessment and the uses of assessment data if they hope to promote supportive social and learning environments for all schoolchildren (NASP, 2010; Reschly & Bersoff, 1999). In this chapter, we discuss legal guidelines for assessment in school-based practice and the three-tier model (TTM) for delivery of comprehensive school psychological assessment and intervention services. We also discuss a contemporary assessment controversy, namely, differing approaches to the identification of students who have a specific learning disability (LD) as defined in the Individuals With Disabilities Education Improvement Act of 2004 (IDEIA) and who are therefore eligible for special education and related services.

LEGAL FRAMEWORK FOR SCHOOL PSYCHOLOGY ASSESSMENT PRACTICES

This chapter focuses on the legal regulation of school-based psychoeducational assessment practices. *School-based practice* is defined as "the provision of school psychological services under the authority of a state, regional, or local educational agency," whether the school psychologist "is an employee of the schools or contracted by the schools on a per-case or consultative basis" (NASP, 2010, p. 3). School-based assessment and intervention practices are highly regulated by law; for this reason, we begin with a brief overview of the legal underpinnings of school-based practice.

The 10th Amendment of the U.S. Constitution has been interpreted as prohibiting the federal government from establishing a nationalized educational system. State governments have assumed the duty and authority to educate youths, which is further delegated by state government to local school boards (Hubsch, 1989). When school psychologists employed by a school board make decisions in their official roles, such acts are seen as an extension of the authority of state government; in legal language, school-based practitioners are considered to be state

DOI: 10.1037/14049-012
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

actors (Jacob, Decker, & Hartshorne, 2011; NASP, 2010; Russo, 2006).

Because education is a state responsibility, for many years the federal government was reluctant to intervene in matters concerning the operation of the public schools. Beginning in the 1950s, however, the federal courts became increasingly involved in public education issues because of school actions that violated the legal rights of students and their parents under the U.S. Constitution. The entitlement to a public education created by state law is a property right protected by the 14th Amendment's equal protection clause. In *Brown v. Board of Education* (1954), the U.S. Supreme Court held that under the 14th Amendment, a state must provide equal educational opportunity to all of its citizens regardless of race. *Brown v. Board of Education* provided the legal reasoning for a subsequent series of "right-to-education" court cases that won access to a public education for students with disabilities (e.g., *Mills v. Board of Education of the District of Columbia*, 1972; *Pennsylvania Association for Retarded Children v. Commonwealth of Pennsylvania*, 1971, 1972). Pertinent to public school assessment practices, the Supreme Court also held that the due process clause of the 14th Amendment protects individuals from arbitrary or unwarranted stigmatization by the state that may interfere with the ability to acquire property (e.g., *Wisconsin v. Constantineau,* 1971). Consequently, a public school may not label a student as having a developmental or cognitive disability without a fair decision-making procedure that includes parental notice of the proposed label and the right to protest the classification.

In addition to the right-to-education court cases that required schools to offer all students with disabilities a free education, a series of court cases beginning in the 1970s challenged whether the assessment practices used by public schools to assign students to "unequal and inferior" classes for students with mental retardation were racially and culturally discriminatory (e.g., *Diana v. State Board of Education,* 1970). These court rulings, together with the right-to-education cases, identified multiple public school responsibilities to students with disabilities that were later incorporated in federal legislation (e.g., the Education for All Handicapped Children Act of 1975).

The U.S. Congress began to pass federal legislation designed to improve the nation's schools in 1965. Congress has the power to shape educational policy and practices by offering monies to states contingent on compliance with federal statutory law. The first major federal law that provided funds to states for education was the Elementary and Secondary Education Act of 1965, passed by Congress to ensure a basic floor of educational opportunity, particularly for students from disadvantaged backgrounds. A series of laws followed that provided funds to states for the education of students with disabilities, including the Education for All Handicapped Children Act of 1975, now known as IDEIA. IDEIA provides funds to states on the condition that states implement a plan to locate and offer a free, appropriate education to all children with disabilities within the state. This chapter focuses on assessment under IDEIA's Part B, the part of IDEIA that provides funds for students with disabilities ages 6 through 18 (or ages 3–21 as determined by state law). A portion of IDEIA Part B funds (15%) may be used to provide early intervention services to students who are struggling in the general education curriculum; the remaining funds are to provide special education and related services to students with a disability as defined by the law. *A child with a disability* means a student evaluated in accordance with the procedures outlined in the law who is found to qualify as having a disability in one of 13 categories and who, for that reason, needs special education and related services (34 C.F.R. § 300.8[a]; see also the Assessment of Learning Disability Eligibility section).

To receive funds, each state must have a plan that offers every child with a disability an opportunity to receive special education and related services in conformance with an Individualized Education Program (IEP). The IEP must be developed in accordance with the procedures outlined in the law and provide a special education program that is reasonably designed to confer benefit (*Board of Education of the Hendrick Hudson Central School District v. Rowley*, 1982). A child with a disability must be educated in the *least restrictive appropriate environment,* namely the educational setting selected from a continuum of alternative placements (ranging from a residential facility to the general education classroom) that is

closest to the general education classroom but also meets the child's individual special education needs. Determination of whether a child is eligible for special education under IDEIA is made by a group of people that includes a child's parent (or an adult acting in the place of a parent); if the child is found eligible, the IEP is developed by a team (the IEP team) that includes the parents. Under IDEIA, the parents of children with disabilities (and adult students) have multiple due process protections to safeguard against misclassification, inappropriate evaluation and placement, and failure of the school to provide an IEP reasonably designed to confer benefit. Parents may use administrative remedies outlined in IDEIA (e.g., impartial resolution meetings and due process hearings) to resolve disputes regarding their child's eligibility, classification, placement, or IEP, and they have the right to file a lawsuit against the school if they are not satisfied with the outcome of administrative remedies. Parents may recover the cost of their attorney's fees if they prevail in a court action on any significant issue. Because school–parent disputes under IDEIA are not uncommon, school psychological assessment practices must be legally defensible and documented with enough detail to withstand challenges in due process hearings and court (Jacob et al., 2011).

The U.S. Congress also enacted civil rights legislation that prohibits schools from discriminating against individuals on the basis of race, color, national origin, sex, or disability. Schools must comply with antidiscrimination legislation if they receive any federal funds for any purpose. The definition of *disability* under Section 504 of the Rehabilitation Act of 1973 is broader and more open-ended than IDEIA (Zirkel, 2009). Section 504 evaluation regulations generally require determination of the following:

1. Is there a physical or mental impairment?
2. Does that impairment substantially limit a major life activity?
3. What kind of accommodations would be needed so that the student will be able to enjoy the benefits of the school program? (Martin, 1992).

Section 504 does not require a specific categorical diagnosis, only the determination of a condition that substantially impairs one or more major life activities at school and requires special accommodation by the school. Under Section 504, schools are required to make accommodations to ensure that pupils with disabilities have equal opportunity to benefit from the schools' programs and activities as their peers without disability.

RELATIONSHIP BETWEEN ETHICAL AND LEGAL ASSESSMENT GUIDELINES

In 1969, NASP was formed to represent school-based psychologists better, particularly non–doctoral-level school psychologists. As described in the previous section, the legal landscape for schools and school psychologists was undergoing rapid change at that time. In 1974, a special issue of NASP's *School Psychology Digest* (now *School Psychology Review*) addressed emerging ethical and legal issues in school psychology (Kaplan, Crisci, & Farling, 1974). Contributors to the special edition recognized that school psychology practitioners needed additional guidance to navigate the legal changes confronting them, and they called for the development of a code of ethics specifically for school psychologists. The American Psychological Association's (APA's) 1963 code of ethics was seen as "either irrelevant or much too vague for operational clarity" for practitioners (Trachtman, 1974, p. 5). Some principles in APA's ethics code conflicted with changing education laws (Ackley, 1974; Bersoff, 1974; Trachtman, 1974). In addition, APA's ethics code did not address issues of growing importance to school-based practitioners such as balancing parent rights with the interests of children (Bersoff, 1974); involving students in decisions affecting their own welfare (Bersoff, 1974; Trachtman, 1974); ensuring fair and valid assessment of students from diverse linguistic and cultural backgrounds; and managing conflicts inherent in the dual roles of child advocate and school employee (Bersoff, 1974; Trachtman, 1974). In 1974, NASP adopted its own code of ethics, the *Principles for Professional Ethics* (PPE). The code was most recently revised in 2010 (see Armistead, Williams, & Jacob, 2011).

School psychology assessment practices are informed by NASP's (2010) PPE and APA's (2010)

Ethical Principles of Psychologists and Code of Conduct. The *Standards for Educational and Psychological Testing,* or *Standards* (American Educational Research Association, APA, & National Council on Measurement in Education, 1999), also provides criteria for acceptable assessment practices and has been cited as an authoritative source in court cases challenging assessment practices. Although legal requirements and ethical guidelines for school-based psychoeducational assessment are often similar, at times they result in ambiguity regarding how to address challenging situations. In challenging cases, we recommend that ethical guidelines be considered first because they typically recommend practices that are above and beyond those legally required; however, in situations in which ambiguity remains, legal requirements can be used to determine courses of action. What follows is an integration of ethical and legal standards for school psychological evaluations according to the temporal order of assessment activities.

Before Assessment

Before considering whether a comprehensive psychoeducational assessment of an individual student is needed, school psychologists ensure that appropriate behavioral and instructional practices have been implemented within the student's school environment (NASP PPE II.3.1). This step requires a systematic assessment of factors in the child's learning environment (Ysseldyke & Christenson, 1988). Ethically, a student should not be exposed to the risk of misdiagnosis unless deficiencies in instruction have first been ruled out (Messick, 1984).

At the outset of establishing a school psychologist–client relationship for the purpose of conducting a psychological assessment with individual students, it is ethically and legally necessary to engage parents and students in the informed consent process (NASP PPE I.1.2). This process ensures that the dignity and rights of the families and students working with school psychologists are respected. Both IDEIA and APA's and NASP's ethics codes provide guidance for how this process should occur. Generally, informed consent is obtained when assessment procedures go beyond normal educational activities (APA Ethical Standard 9.03) and in cases when school psychologists are involved in a student's education to an extensive degree (see NASP PPE I.1.1). When an initial assessment of whether a student has a disability under IDEIA or Section 504 of the Rehabilitation Act of 1973 is under consideration, informed consent is legally required (IDEIA; NASP PPE I.1.2; see also Section 504). An important part of the informed consent process is ensuring that consent is voluntary, ongoing, and informed. The individual providing consent (e.g., a parent or an individual acting in the place of a parent or an adult student) must be given sufficient information to make an informed choice about whether the psychoeducational assessment will be conducted and advised that he or she may revoke consent at any time. Soliciting informed consent involves discussion of the nature and purpose of the assessment, any potential consequences of the assessment results, who will receive information about the outcomes, and the limits of confidentiality (APA Ethical Standard 9.03a; IDEIA; NASP PPE I.1.3). It is ethically permissible to bypass a minor's assent if the service is considered to be of direct benefit to the student or is required by law; however, if a child's assent is not solicited, the school psychologist nevertheless ensures that the child is informed about the nature and purpose of the assessment (NASP PPE I.1.4).

Before beginning an assessment, school psychologists identify instruments and procedures that are technically adequate, valid for the purpose of the assessment, and appropriate for the student who is being assessed (APA Ethical Standard 9.02[b]; NASP PPE II.3.2). If a student is suspected of having a disability under IDEIA, the student must be assessed on the basis of procedures that are multifaceted (based on a variety of assessment tools and strategies), comprehensive (the child is assessed in all areas related to the suspected disability), technically adequate and valid for the purpose used, fair (nondiscriminatory), and useful (provide information that directly assists in determining educational needs; 34 C.F.R. § 300.304; see also Jacob et al., 2011, and Exhibit 12.1 and Table 12.1).

Exhibit 12.1
Excerpts From Individuals With Disabilities Education Improvement Act Regulations on Evaluation Procedures

§ 300.304 Evaluation procedures.

(b) *Conduct of evaluation.* In conducting the evaluation, the public agency must—

(1) Use a variety of assessment tools and strategies to gather relevant functional, developmental, and academic information about the child, including information provided by the parent, that may assist in determining—

(i) Whether the child is a child with a disability under § 300.8; and (ii) The content of the child's IEP [Individualized Education Program], including information related to enabling the child to be involved in and progress in the general education curriculum (or for a preschool child, to participate in appropriate activities);

(2) Not use any single procedure as the sole criterion for determining whether a child is a child with a disability and for determining an appropriate educational program for the child; and

(3) Use technically sound instruments that may assess the relative contribution of cognitive and behavioral factors, in addition to physical or developmental factors.

(c) *Other evaluation procedures.* Each public agency must ensure that—

(1) Assessments and other evaluation materials used to assess a child under this part—

(i) Are selected and administered so as not to be discriminatory on a racial or cultural basis;

(ii) Are provided and administered in the child's native language or other mode of communication and in the form most likely to yield accurate information on what the child knows and can do academically, developmentally, and functionally, unless it is clearly not feasible to so provide or administer;

(iii) Are used for the purposes for which the assessments or measures are valid and reliable;

(iv) Are administered by trained and knowledgeable personnel; and

(v) Are administered in accordance with any instructions provided by the producer of the assessments.

(2) Assessments and other evaluation materials include those tailored to assess specific areas of educational need and not merely those that are designed to provide a single general intelligence quotient.

(3) Assessments are selected and administered so as best to ensure that if an assessment is administered to a child with impaired sensory, manual, or speaking skills, the assessment results accurately reflect the child's aptitude or achievement level or whatever other factors the test purports to measure, rather than reflecting the child's impaired sensory, manual, or speaking skills (unless those skills are the factors that the test purports to measure).

(4) The child is assessed in all areas related to the suspected disability, including, if appropriate, health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative status, and motor abilities;

(5) Assessments of children with disabilities who transfer from one public agency to another public agency in the same academic year are coordinated with those children's prior and subsequent schools, as necessary and as expeditiously as possible, consistent with § 300.301(d)(2) and (e), to ensure prompt completion of full evaluations.

(6) In evaluating each child with a disability under §§ 300.304 through 300.306, the evaluation is sufficiently comprehensive to identify all of the child's special education and related services needs, whether or not commonly linked to the disability category in which the child has been classified.

(7) Assessment tools and strategies that provide relevant information that directly assists persons in determining the educational needs of the child are provided.

(Authority: 20 U.S.C. 1414[b][1–3], 1412[a][6][B])

Practitioners are obligated to choose instruments with adequate and up-to-date normative data to ensure appropriate comparative information between reference groups and the child or youth being assessed. They also consider individual student characteristics such as ethnicity, primary language, and disabilities when selecting assessment procedures so as to provide accurate, fair, and useful results (APA Ethical Standard 9.02c; NASP PPE I.3.2, II.3.5–3.6). Because an increasingly greater proportion of the children who attend U.S. schools come from ethnically and linguistically diverse backgrounds, school psychologists are frequently required to conduct assessments of students who come from backgrounds very different from their own. The IDEIA, APA's Ethics Code, NASP's PPE, and the *Standards* include multiple statements on valid and fair assessment of students with sensory or motor disabilities and those from culturally and linguistically diverse backgrounds.

TABLE 12.1

Five Elements of Ethically and Legally Appropriate Assessment Within Response to Intervention

Element of ethical assessment	Brief definition	Ethical practice in response to intervention
Multifaceted	Assessment must be based on different types of information from different sources.	Use multiple measures with validated outcomes. Do not rely too heavily on curriculum-based measurement alone.
Comprehensive	Directly measure all behaviors and domains that are relevant to the problem or suspected disability.	Select measures on the basis of their relationship with student outcomes and relevance to the referring questions. Could include measures of teachable skills, performance of skills, and instructional variables.
Fair	Assessment tools and procedures are selected in light of the child's age, gender, native language, disabilities, socioeconomic status, and ethnic background.	Assess student acculturation, language proficiency, and hearing, vision, and sensorimotor information before selecting measures and interventions.
Useful	Assessment procedures should be selected to provide a profile of the child's strengths and difficulties to aid in instructional planning.	Assessment data should be directly linked to goals and objectives and should inform the intervention process.
Valid	Select assessment procedures that have been validated for the purpose for which they are used.	In addition to using psychometric adequate measures, response-to-intervention protocols must be carefully crafted, and intervention planning should use a scientific problem-solving process that involves identifying the problem, generating solutions, and measuring outcomes.

Note. Information from Burns, Wagner, and Jacob (2008).

In accordance with the IDEIA (see Exhibit 12.1), tests and other assessment tools used in the evaluation of children with suspected disabilities should be

> provided and administered in the child's native language or other mode of communication and in the form most likely to yield accurate information on what the child knows and can do academically, developmentally, and functionally, unless it is not feasible to so provide or administer. (34 C.F.R. § 300.304[c][1][ii]; see also APA Ethical Standard 9.02; *Standards,* pp. 91–100; NASP PPE 11.3.5)

Furthermore, materials and procedures used to assess a child with limited English proficiency are selected and administered to ensure that they measure the extent to which the child has a disability and needs special education rather than the child's English language skills (34 C.F.R. § 300.304[c][3]; van de Vijver & Phalet, 2004).

Students from culturally and linguistically diverse backgrounds may have different degrees of acculturation to the mainstream culture, which is where normative data for standardized school psychological assessments are typically derived. Therefore, competent assessment of students from diverse backgrounds requires the practitioner to gather information about the family's degree of acculturation and to assess the child's language proficiency before selecting assessment tools (Dana, 2000; Ortiz, 2008; Paredes Scribner, 2002). Language proficiency information is needed to guide selection and interpretation of measures of aptitude, achievement, and adaptive behavior and in planning instruction and interventions (see Paredes Scribner, 2002). Even if a child from a culturally different background demonstrates some proficiency in spoken or written English, it is important to remember that commonly used intelligence tests (e.g., the Wechsler scales) tap the language, symbols, and knowledge children encounter in the dominant U.S. culture and schools

(Jacob et al., 2011). Ortiz (2008) provided useful guidance on how to conceptualize assessment practices for linguistically or ethnically diverse students. To minimize bias in data collection and interpretation, he suggested that the process of assessment "begin with the hypothesis that the examinee's difficulties are not intrinsic in nature, but rather that they are more likely attributable to external or environmental problems" (p. 664). Examiners must use their knowledge of a student's unique experiences and background to evaluate and interpret all information gathered. The hypothesis of normality is not rejected unless the data strongly suggest the contrary.

During Assessment

Consistent with IDEIA and APA's and NASP's codes of ethics, psychological assessments are conducted by qualified, knowledgeable personnel (APA Ethical Standard 9.07; NASP PPE II.5.2; see Table 12.1). Practitioner competence is necessary to ensure that assessment instruments are administered and interpreted appropriately (APA Ethical Standard 9.09.a, c) and that results are accurately communicated to students, parents, and educators (APA Ethical Standard 9.10; NASP PPE II.3.8). Furthermore, practitioners are obligated to "follow carefully the standardized procedures for administration and scoring specified by the test developer, unless the situation or test taker's disability dictates that an exception should be made" (*Standards* 5.1). If modifications are necessary, they are based on carefully considered professional judgment. Also, testing environments should be of "reasonable comfort and with minimal distractions" (*Standards* 5.4). If an assessment is not conducted under standard conditions, a description of the extent to which it varied from standard conditions should be included in the evaluation report (*Standards* 5.2; APA Ethical Standard 9.06; NASP PPE II.3.2).

A comprehensive assessment seeks to gather information from a variety of sources (NASP PPE II.3.3) for the purpose of making informed educational decisions that will benefit the student. However, in responsible psychological assessment, the practitioner also remains sensitive to pupil and family privacy (Matarazzo, 1986). During assessment, school psychologists are ethically obligated to respect privacy (APA Principle E; NASP PPE I.2). They do not seek or store personal information about the student, parents, or others that is not needed in the provision of services (APA Ethical Standard 4.04; NASP PPE I.2.2.).

After Assessment

A common reason for assessment of a student who is struggling academically or behaviorally is to determine whether the student is eligible to receive special education and related services under IDEIA or is eligible for accommodations under the Rehabilitation Act of 1973 (Section 504). If a student is found eligible under IDEIA or Section 504, assessment results are used to assist in identifying the most appropriate educational placement for the student that is closest to the general education classroom (i.e., the least restrictive appropriate environment), to inform development of an IEP that is reasonably designed to confer benefit, to identify school accommodations so that the student has educational opportunities equal to his or her peers without disabilities, or all of these. School psychologists must therefore "adequately interpret findings and present results in clear understandable terms so that the recipient can make informed choices" (NASP PPE II.3.8). School psychologists indicate any reservations that exist concerning validity of their findings (APA Ethical Standard 9.06; NASP PPE II.3.2). Moreover, educational plans that derive from a comprehensive assessment should be actively monitored to ensure the predicted benefits of the recommended program (NASP PPE II.2.2).

One of the primary ways by which school psychologists share assessment results is through the written report. In preparing reports, practitioners must consider their obligation to ensure their findings are understandable and useful to the intended recipient (NASP PPE II.3.8) as well as their obligation to safeguard the confidentiality of sensitive private information about the student and family (NASP PPE I.2). In accordance with the Family Educational Rights and Privacy Act of 1974, parents have the right to access all school psychological assessment findings for their child, but school-based practitioners should use discretion when choosing

the information to include in psychological reports prepared for different purposes. It may be ethically appropriate (and legally advisable) to exclude sensitive family and student information from a report written for the purpose of making special education decisions or identifying instructional needs. However, with parent permission, this information could be shared with others in the school setting or included in a referral to a professional outside the school. In all cases, school psychologists foster continued parental involvement through honest and forthright reporting of their findings within the promised timeframe (APA Ethical Standard 9.10; NASP PPE III; *Standards* 5.10).

It is also legally and ethically advisable practice for school psychologists to foster parent and student involvement in designing interventions on the basis of assessment results (IDEIA; NASP PPE II.3.10). School psychologists "discuss with parents the recommendations and plans for assisting their children. This discussion takes into account the ethnic/cultural values of the family and includes alternatives associated with each set of plans" (NASP PPE II.3.10). When possible, this discussion of the psychoeducational evaluation should include the child. Recommendations for program changes or additional services are discussed with the student, along with any alternatives that may be available (NASP PPE II.3.11). Consistent with ethical principles, students should be afforded opportunities to participate in decisions that affect them.

Finally, school psychologists are obligated to safeguard test security (*Standards* 11.7, 11.8; APA Ethical Standard 9.11; NASP PPE II.5.1). The development of valid assessment instruments is costly and requires extensive research. The disclosure of underlying principles or specific content of a test is likely to decrease its validity for future examinees. Disclosure of test content may also infringe on the intellectual property or copyright interests of the test producer (APA, 1996). In school-based practice, however, parents generally have a legal right to review their child's test answers on a school psychological test record form because the booklet on which an individual student's answers are recorded is an education record as defined by the Family Educational Rights and Privacy Act of 1974 (Rooker, 2008; Rosenfeld, 2010). NASP's code of ethics states, "School psychologists understand that, at times, parents' rights to examine their child's test answers may supersede the interests of test publishers" (NASP PPE II.5.3). Thus, it is ethically permissible for school-based practitioners to comply with education law and allow parents to review their child's answers on a school psychological test record form. School psychologists have no obligation to show parents test manuals or the testing materials (see Jacob et al., 2011).

In addition, school psychologists engage in professionally responsible record keeping practices, safeguarding the privacy and security of their school psychological education records (NASP PPE II.4). School psychologists who are employed by schools that receive federal funds are generally required to comply with the Family Educational Rights and Privacy Act of 1974 rather than the Health Insurance Portability and Accountability Act of 1996 (see U.S. Department of Health and Human Services & U.S. Department of Education, 2008). At the elementary or secondary school level, the Family Educational Rights and Privacy Act of 1974 does not make a distinction between student physical and mental health records and other types of student education records, which may pose special challenges in protecting the privacy of sensitive information included in a school psychologist's student records, especially in school districts in which the school psychologist's records are not under his or her own control (see Jacob et al., 2011).

School psychologists must engage in legal and ethical assessment practice before, during, and after assessments are conducted. However, the principles described here apply to all assessment activities. Assessment has changed dramatically since Binet and Simon published their first test. Thus, next we discuss the implications of legal and ethical assessments within a three-tiered model of service delivery.

ASSESSMENTS IN SCHOOL-BASED PRACTICE: THE THREE-TIER MODEL

The TTM has gained acceptance as a preferred model for the delivery of comprehensive school psychological assessment and intervention services

(Ysseldyke et al., 2006). The first tier (Tier 1—Universal) involves providing effective instruction for students in general education and monitoring student progress, the second (Tier 2—Targeted) involves providing small-group interventions to remediate a broad deficit (e.g., phonics, reading comprehension) while monitoring student progress, and the third (Tier 3—Intensive) involves providing individualized interventions to address a specific skill deficit while frequently monitoring student progress (Burns & Gibbons, 2008; Marston, 2003). In the following sections, we discuss specific assessment activities within the TTM and the legal and ethical implications of each.

Universal Screening

The first step in a TTM is to conduct universal screenings, in which school personnel collect data on important academic or social skills or possible mental health problems (e.g., reading skills, behavioral difficulties) for all students in a classroom, school, or district to identify those who may be at risk for developing further difficulties or who may potentially benefit from additional intervention (Glover & Albers, 2007). The school psychologist is often the school professional with the greatest expertise in measurement in the district. As such, practitioners may be asked to help administrators and teachers make decisions regarding whether a screening program is needed, select tests and assessment tools that are technically adequate for the intended purpose, and develop guidelines for appropriate use and interpretation of the results (Jacob et al., 2011).

The universal screening process attempts to identify students who may need intervention in general education programming. The screening of a student by a teacher or specialist to determine appropriate instructional strategies is not considered to be an evaluation requiring parental consent under IDEIA (34 C.F.R. § 300.302). However, consistent with the Protection of Pupil Rights Act of 1978, as amended in the No Child Left Behind Act of 2001, and NASP's code of ethics, parents should "be notified prior to the administration of school- or classroom-wide screenings for *mental health problems* [italics added] and given the opportunity to remove their child from participation" (NASP PPE I.1.1) because such screenings may be more intrusive on personal and family privacy than expected in the course of typical school activities (Jacob, 2009).

Progress Monitoring

According to the National Center on Student Progress Monitoring (n.d.), *progress monitoring* is the process of measuring an individual student's or group of students' performance on a regular basis (weekly or monthly) and comparing the rate of improvement toward a goal to evaluate the effectiveness of instruction or intervention. School psychologists should ensure that student response to the intervention is measured and documented with psychometrically adequate data and whenever modifications are made on the basis of the data (NASP PPE II.2.2).

School psychologists are ethically obligated to promote parental participation in school decisions that affect their child (NASP-PPE I.1.2, II.3.10), but informed parent consent is generally not needed unless a psychologist–client relationship will be established. Consistent with special education law (IDEIA), school-based practitioners are not ethically obligated to obtain parental consent to conduct classroom observations, assist in within-classroom interventions and progress monitoring, or participate in educational screenings conducted under the authority of the teacher within the scope of typical classroom functions unless these actions result in a significant intrusion on student or family privacy beyond what might be expected in the course of ordinary school activities (NASP PPE I.1.2). It is advisable for school districts to advise parents in their school district handbook that school psychologists routinely assist teachers in planning instruction and monitoring student progress and that parent consent is not sought for such activities (NASP PPE I.1.2). Schools should notify a parent before a student receives a Tier 2 or Tier 3 intervention that such interventions fall within the parameters of general education and will result in regular (e.g., weekly) progress monitoring by the school psychologist or other trained staff. As with universal screenings, the progress monitoring process should be clearly described to the students in language that

they will understand, and the results should be explained to the parents in a manner that they understand.

As noted previously, school personnel are obligated to ensure that progress monitoring is done in a manner that results in psychometrically adequate data. Assessments within a TTM run the gamut from informal measures such as curriculum-based measurement (Deno, 1985), curriculum-based assessment for instructional design (Gickling & Havertape, 1981), and interviews and structured observations to standardized norm-referenced measures of academic achievement and behavior. Thus, TTM may involve using measures other than those with established reliability and validity. As noted previously, psychoeducational evaluations should be multifaceted, comprehensive, fair, valid, and useful (Burns, Wagner, & Jacob, 2008; Jacob et al., 2011). Table 12.1 lists these five elements of school psychological assessment practices and briefly describes how to address each when implementing response to intervention (RtI). Next, we expand on the points that are most relevant to specific assessment activities within the TTM.

Curriculum-based measurement is frequently used to monitor student progress in the schools, and research has consistently demonstrated the reliability of curriculum-based measurement data (Wayman, Wallace, Wiley, Ticha, & Espin, 2007). However, progress monitoring data are interpreted by examining the slope of the data, which represents the rate of change in student performance per unit of time (Christ, 2006). For example, a student's reading growth could be measured weekly with curriculum-based measurement of oral reading fluency and result in an average increase of 1.80 words per minute per week. Traditional practices have suggested graphically displaying the data and visually analyzing the slope by drawing a line of best fit through the data (Fuchs, Fuchs, Hintze, & Lembke, 2006; Shinn, 1989). Recent research has found that decisions based on visual interpretations of slope data were unreliable, and the reliability of decisions based on numerical computations was significantly higher (Burns, Scholin, Kosciolik, & Livingston, 2010). However, numerical estimates of slope had standard errors of measure that were so large that they could not be interpreted until at least 8 to 10 weeks of data were collected using moderate assessment conditions and 4 to 7 weeks of data were collected using optimal assessment conditions (Christ, 2006).

Collecting sufficient data on the acquisition of academic skills or positive adaptive behaviors using quality assessment conditions (e.g., a well-constructed probe of academic skill or system for behavior coding, appropriate location or setting, individual trained to collect data, collection of reliability data) results in reliable decisions but does not ensure validity. The validity of decisions is dependent on several factors, including the purpose of the assessment (Murphy & Davidshofer, 2001; Salvia, Ysseldyke, & Bolt, 2010). Thus, data collected to monitor student progress should be used only to determine whether the intervention is effective. Moreover, the data collected should match the intervention target. For example, if a student is participating in an intervention that addresses phonetic skills, then phonics measures (e.g., nonsense word fluency) should be used to monitor student progress. It is acceptable to collect general outcome data (e.g., curriculum-based measurement oral reading fluency or multiskill math assessments) as well to indicate overall student progress, but the effectiveness of the intervention should be at least partially judged with data regarding change in that specific skill or target behavior.

Valid decision making with student progress monitoring data is also better ensured if both slope of progress and postintervention level are considered. Student progress is deemed to be sufficient when the slope of progress meets or exceeds a target rate, often defined by a comparison to a mean rate of progress for a given population (e.g., third-grade students in one school). However, practitioners should also consider whether the student's postintervention level met or exceeded benchmark expectations (i.e., scoring within a proficient or acceptable range, demonstrating the target behavior under specified conditions with frequencies similar to same-age, same-gender peers). An intervention would be judged as effective and successful if both the rate of positive change and the postintervention level met or exceeded expectations but would be effective and not

successful if the slope of progress was acceptable but the postintervention level remained below the desired level of proficiency (Riley-Tillman & Burns, 2009). The intervention would likely end in the former example but would continue in the latter. When both the slope and postintervention level score are below desired levels, then the student's performance exhibits a dual discrepancy (Fuchs, 2003) and the intervention is judged as ineffective. Previous research has found that a dual-discrepancy approach was superior to a single-discrepancy approach that only examines slope or postintervention level (Fuchs, 2003), and dual-discrepancy decisions regarding academic skill acquisition converged with the outcomes of norm-referenced achievement tests (Burns & Senesac, 2005; Speece & Case, 2001; Speece, Case, & Molloy, 2003). In other words, students who demonstrated a dual discrepancy scored significantly lower on measures of achievement than did students who were at risk for academic failure but who did not demonstrate a dual discrepancy.

Intervention Decisions

As stated earlier, school personnel need to document student response, but they also need to document that an intervention occurred to ensure ethically and legally defensible RtI practices. A recent due process hearing involved a parent seeking compensatory education for her child and a private evaluation, both at the school district's expense, because the school failed to evaluate her child for special education services despite extensive and ongoing behavioral difficulties. The school district stated that it was using RtI methods to determine special education eligibility, but it had no written record of any intervention plan or of progress monitoring data. The hearing officer found in favor of the parent (Delaware College Preparatory Academy and the Red Clay Consolidated School District Delaware State Educational Agency, 2009).

As noted previously, if interventions occur within general education under the authority of the teacher and do not result in a significant intrusion on student or family privacy, informed parental consent is not needed. However, as is the case with universal screening and progress monitoring, practitioners would be wise to notify parents about new or different interventions, including how often they will occur, for how long, who will implement them, and what the interventions will actually entail.

A legally and ethically defensible intervention plan will also incorporate research-based interventions. The phrase *scientific, research-based interventions* is not defined in IDEIA but is defined by the No Child Left Behind Act of 2001, the most recent reauthorization of the Elementary and Secondary Education Act of 1965, as research that

> (i) employs systematic, empirical methods that draw on observation or experiment;
> (ii) involves rigorous data analyses that are adequate to test the stated hypotheses and justify the general conclusions drawn;
> (iii) relies on measurements or observational methods that provide valid data across evaluators and observers and across multiple measurements and observations; and
> (iv) has been accepted by a peer-reviewed journal or approved by a panel of independent experts through a comparably rigorous, objective, and scientific review. (20 U.S.C. 6368)

Preference should be given to interventions described as effective in peer-reviewed professional literature. NASP's code of ethics requires school psychologists to use "assessment techniques and practices that the profession considers to be responsible, research-based practice" (NASP PPE II.3.2); similarly, the APA's ethics code states, "Psychologists' work is based upon established scientific and professional knowledge of the discipline" (APA Ethical Standard 2.04, see also the Preamble). A federal review panel indicated that strong evidence supports interventions based on intensive, systematic instruction for small groups of students (Gersten et al., 2009). Thus, a solid research base exists from which to select interventions, and providing them to a small group (i.e., Tier 2) is an effective practice.

Finally, interventions should be designed that address the student's problem because correctly

targeting the intervention is a basic requirement for research-based practice (Burns, VanDerHeyden, & Boice, 2008). As noted previously, assessments are designed for specific purposes. Similarly, interventions are geared toward a specific problem and should be used accordingly. The No Child Left Behind Act of 2001 identified five components of effective reading instruction, based on the findings of the National Reading Panel (2000): phonemic awareness, phonics, vocabulary development, reading fluency, and reading comprehension strategies. Thus, the five areas of reading instruction could serve as an intervention heuristic to identify appropriate interventions given a student's individual difficulties in the area of reading. Federal law provides little guidance regarding effective instruction in mathematics or other domains, but math and writing involve more easily identifiable and distinct subskill hierarchies (Baker, Gersten, & Graham, 2003; Shapiro, 2004). The U.S. Department of Education has, however, funded the development of several practice guides based on TTM that are available at the Institute of Education Sciences website (http://ies.ed.gov/ncee/wwc/publications/practiceguides). Examples include *Reducing Behavior Problems in the Elementary School Classroom* (Epstein, Atkins, Cullinan, Kutash, & Weaver, 2008) and *Assisting Students Struggling With Reading: Response to Intervention and Multi-Tier Intervention in the Primary Grades* (Gersten et al., 2009).

Targeted interventions are most likely to be effective if they are contextualized within and linked to quality core instruction (Fuchs et al., 2008; National Mathematics Advisory Panel, 2008). Thus, although practitioners give preference to interventions reported to be effective, they must also adapt those interventions to the setting and to the unique needs of the individual student. In other words, practitioners must strive for fidelity to the treatment program as it is described in the research literature while at the same time adapting the intervention to the characteristics of the school, general education curriculum, classroom, and student.

ASSESSMENT OF LEARNING DISABILITY ELIGIBILITY

Many aspects of school psychological assessment generate professional disagreements and lively debate. For example, questions may include how to conduct a fair and valid assessment of students who are English language learners (e.g., Dana, 2000; Ortiz, 2008; Paredes Scribner, 2002), the ethical and legal challenges associated with computer-assisted assessment (e.g., Harvey & Carlson, 2003; Naglieri et al., 2004), whether projective techniques provide incremental validity beyond available rating scales and other objective techniques in the assessment of students who may qualify as having a disability (e.g., Garb, Wood, Nezworski, Grove, & Stejskal, 2001; Hojnoski, Morrison, Brown, & Matthews, 2006; Lilienfeld, Wood, & Garb, 2000; Miller & Nickerson, 2007; Wood, Lilienfeld, Garb, & Nezworksi, 2000), and whether and how the broad definition of *disability* under Section 504 as amended by the Americans With Disabilities Amendments Act of 2008 will affect school practices (Zirkel, 2009). We have chosen to focus here on the contemporary debate regarding alternative methods of identifying whether a child has a specific LD and is eligible for special education and related services under IDEIA.

Identification of a student as having a disability as defined in IDEIA results in the student being eligible for special education and related services (i.e., a legal entitlement to special education services; Salvia et al., 2010). Several consequences of eligibility determination under IDEIA must be considered, including the dedication of fairly extensive educational resources to the student found eligible as well as the potential stigmatization of students as a result of being assigned a label such as LD. These considerations make eligibility determinations among the most important decisions that IEP teams make and require the most psychometrically rigorous data (Salvia et al., 2010). In the paragraphs that follow, we briefly explore using RtI as part of the LD eligibility evaluation procedure and the use of cognitive neuropsychological assessments to identify children as having an LD. These approaches are not presented as mutually exclusive, and each has strengths and shortcomings with regard to ensuring a valid, fair, and useful assessment.

IDEIA Eligibility Determination Criteria for Specific Learning Disability

As noted previously, to be eligible for special education under IDEIA Part B, a child must have

a disability as outlined in one of the 13 disability categories and he or she must need special education because of that disability. Furthermore, it is important to note that although a *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.; American Psychiatric Association, 2000) diagnosis may assist in determining eligibility under IDEIA (e.g., a diagnosis of developmental reading disorder), such a diagnosis is neither legally required nor sufficient under federal law to determine whether a student is eligible for special education under IDEIA Part B (Zirkel, 2009).

According to the IDEIA 2004 regulations,

> Specific learning disability means a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, that may manifest itself in an imperfect ability to listen, think, speak, read, write, spell, or to do mathematical calculations, including conditions such as perceptual disabilities, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia. (34 C.F.R. § 300.8[c][10])

Before the 2004 amendments, Individuals With Disabilities Education Act of 1990 regulations stated that the IEP team could determine that a child has a specific LD only if the child had a severe discrepancy between an area of academic achievement and intellectual ability. Under IDEIA, IEP teams are no longer required to take into consideration whether a child has a severe discrepancy between achievement and intellectual ability. Instead, state departments of education "must permit the use of a process based on the child's response to scientific, research-based intervention" (34 C.F.R. § 300.307[a][2]), which is commonly referred to as *RtI.* Thus, it is legally permissible to use data regarding student response to systematic research-based interventions as the LD identification evaluation.

The IDEIA regulations specify the team members who are to be involved in eligibility determination, and this team must include a minor child's parent or an adult acting in the place of the parent (34 C.F.R. § 300.308). The regulations go on to state that a team may determine that a child has a specific LD if the child does not achieve adequately for his or her age (or fails to meet state-approved grade-level standards) in one or more of the following areas when provided with appropriate learning experiences and instruction: oral expression, listening comprehension, written expression, basic reading skill, reading fluency skills, reading comprehension, mathematics calculation, and mathematics problem solving. The team must determine that the lack of adequate achievement is not primarily the result of a visual, hearing, or motor disability; mental retardation; emotional disturbance; cultural factors; environmental or economic disadvantage; or limited English proficiency (34 C.F.R. § 300.309).

The regulations also require the team to ensure that the underachievement by a child suspected of having a specific LD is not the result of a lack of appropriate instruction in reading or math and whether data demonstrate that before (or as a part of) the referral process the child was provided appropriate instruction in general education settings. As part of the evaluation, data-based documentation of repeated assessments (at reasonable intervals) of the student's progress during instruction must be considered (34 C.F.R. § 300.309). Furthermore, the regulations require an observation of the child's academic performance and behavior in the child's learning environment, including the general education classroom, or an age-appropriate setting if not in school (34 C.F.R. § 300.310[a], [c]). (See 34 C.F.R. § 300.311[b] for a list of the required components of the team report.)

Response to Intervention and Learning Disability Eligibility Determination

As noted previously, IDEIA regulations state that when identifying a child with a LD, a local educational agency may "use a process based on the child's response to scientific, research-based intervention" (34 C.F.R. § 300.307[a][2]). The process of using student RtI data to identify students with LD often coincides with a school district's implementation of the TTM.

School personnel do not need informed parental consent to implement interventions that are under the authority of the teacher and that are within the

scope of typical classroom practices, but consent for psychoeducational assessment is needed as soon as a disability is suspected. Thus, if at any point during the RtI process a student is suspected of having a disability under IDEIA or Section 504, the school is required to seek consent to conduct an individual evaluation in accordance with IDEIA (or Section 504) procedures and timelines. If parents request a special education eligibility evaluation during the RtI process and the school decides not to evaluate the child because the data do not suggest a disability, then the school must provide parents written notice of the refusal to evaluate and information describing their rights to challenge that decision (Burns, Wagner, et al., 2008). School districts may not require that RtI be implemented for a predetermined number of weeks before responding to a parent request for an evaluation under IDEIA or Section 504 (Acalanes [CA] Union High School District Office for Civil Rights, Western Division, San Francisco, 2009).

Using RtI data to identify a student as having an LD is an eligibility decision that must be done in a manner that ensures psychometric rigor. Conducting psychoeducational assessments in accordance with the five criteria previously discussed and summarized in Table 12.1 should help ensure valid decisions. IDEIA requires a full and individual initial evaluation before a child is classified as having a specific LD and states that students must be assessed in all areas related to the suspected disability (34 C.F.R. § 300.304; see Exhibit 12.1). The National Joint Committee on Learning Disabilities (2010) recommended that comprehensive evaluations be made up of multiple measures, both standardized and nonstandardized, and include other data sources, such as case history and interviews with parents, educators, related professionals, and the student if appropriate; direct observations in multiple settings; and continuous progress monitoring repeated during instruction and over time. All of those potential data sources should be considered, as should measures of motor, sensory, cognitive, communication, and behavior if believed to be relevant.

Simply adding additional data does not make the evaluation comprehensive unless those data are relevant, and all relevant areas are assessed. An IEP team could make an LD identification decision without administering any norm-referenced measures as long as the team directly measured achievement, behavior, instructional environment, and all other relevant factors (e.g., teachable skills, skill acquisition and performance, prior and current instructional opportunities, time allocated for instruction, academic learning time, pace of instruction, number of opportunities to respond, and sequencing of examples and nonexamples of skills, indicators of student progress over time; Burns, Wagner, et al., 2008). However, it is likely good practice to assess student academic skills with a standardized norm-referenced measure as part of the comprehensive evaluation (Fletcher, Lyon, Fuchs, & Barnes, 2007). Other norm-referenced psychological tests (e.g., intelligence) could be used only when necessary.

Although the TTM has been endorsed as the best manner to deliver school psychological services (Ysseldyke et al., 2006), implementing the model will require school psychologists to reconsider how they define valid assessments. For example, for LD eligibility decisions within a TTM to be valid, teachers and school psychologists must be trained to precisely implement interventions and to reliably measure the resulting changes in student performance. This change may mean relatively extensive training for some school personnel and increased support and resources necessary to implement a TTM.

Another threat to the validity of decisions made with a TTM could be the policy decision regarding what level of nonresponse is necessary to warrant LD identification. If the bar for determining failure to respond to interventions is set too low, it is likely that too many children will be referred for a comprehensive evaluation for suspected LDs, and a bar that is set too high may result in delayed IEPs for some children (Burns, Wagner, et al., 2008). The standard with which nonresponsiveness will be judged will be a policy decision much like previous LD diagnostic criteria (Ysseldyke, 2005). Empirical data exist to examine the validity of various nonresponsive criteria (Burns & Senesac, 2005; Fuchs, 2003), but little is known about the effect the chosen criteria would have on the frequency with which students would be referred for LD evaluation.

Use of Cognitive Neuropsychological Assessments to Identify Children With Learning Disabilities

Another school psychological assessment approach to identifying children with a specific LD involves the use of cognitive neuropsychological measures. This approach gained popularity in light of the 2004 amendments to the Individuals With Disabilities Education Act of 1992 that allow IEP teams to identify a child as having an LD if the child "exhibits a pattern of strengths and weaknesses in performance, achievement, or both, relative to age, State approved grade-level standards, or intellectual development, that is determined by the group to be relevant to the identification of a specific learning disability" (34 C.F.R. § 300.309 [a][2][ii]). This provision is often implemented by administering a battery of standardized norm-referenced tests, including intelligence measures and tests of achievement and conducting ipsative academic ability analyses, ipsative cognitive ability analyses, and integrated analyses to document that academic deficits are related to cognitive deficits and an evaluation of the degree to which cognitive deficits interfere with academic functioning (Kavale & Forness, 2003).

Neuropsychology is the study of the structure and function of the brain as it relates to specific psychological processes (Posner & DiGirolamo, 2000). Thus, cognitive neuropsychological assessment is the measurement of those psychological processes as indicators of brain functioning. Data from neuropsychological and cognitive assessments are interpreted through contemporary theories of intelligence such as the Cattell–Horn–Carroll model, which often requires data from multiple measures to implement (Flanagan, 2000). However, as stated earlier, simply adding additional measures does not ensure a multifaceted or comprehensive evaluation if the additional tests measure similar constructs. School psychologists could use cognitive neuropsychological assessment data as part of a comprehensive and multifaceted evaluation if they include data from other sources (e.g., parents, teachers) and also assess all areas related to the suspected disability.

Although providing a multifaceted and comprehensive evaluation is important, the data must also be fair, useful, and valid. The relationship between cognitive neuropsychological data and academic achievement is well established (Floyd, Evans, & McGrew, 2003; Hale, Fiorello, Bertin, & Sherman, 2003; Hale, Fiorello, Kavanagh, Hoeppner, & Gaither, 2001), and the interpretation framework is based on current theory regarding the nature of intelligence, both of which suggest evidence for validity. Moreover, intelligence tests and instruments used for cognitive neuropsychological evaluations often result in highly reliable data. However, advocates for a cognitive neuropsychological approach to identifying students with an LD advocate for doing so by identifying a discrepancy between two cognitive processes (Naglieri, 1999), which is a questionable practice given the concerns about basing identification decisions on data from discrepancies between scores on an intelligence test or between scores on cognitive and achievement measures (Aaron, 1997).

Additional research is needed before practitioners can conclude that using cognitive neuropsychological assessments as part of a comprehensive and multifaceted evaluation results in an assessment that is fair or useful. Using a cross-battery assessment that compiles data across measures to create a Cattell–Horn–Carroll profile has been recommended for students whose native language was not English (Rhodes, Ochoa, & Ortiz, 2005), but little research appeared to support that recommendation. Moreover, some of the cognitive processes associated with LD identification (e.g., difficulties with phonological processing, working memory, or rapid recall) are common among all children with learning difficulties, including those not identified as having an LD (Aaron, 1997; Dean & Burns, 2002). Finally, some have advocated for using cognitive neuropsychological data to design interventions (Fiorello, Hale, & Snyder, 2006), but previous meta-analytic research found only a small to moderate effect size ($d = 0.39$) for interventions based on cognitive processes data and a large effect size ($d = 0.84$) for direct instruction (Kavale & Forness, 2000). Additionally, research on academic interventions developed from cognitive neuropsychological data frequently have small sample sizes (e.g., 3–5 students), do not use true control groups, and report effect sizes by comparing pre- and postintervention

scores rather than comparing treatment and control group scores.

CONCLUSION

School psychology has undergone dramatic changes since its inception, many of which were the direct result of a changing legal landscape. For example, the enactment of the Education of All Handicapped Children Act of 1975 moved school psychologists into predominantly assessment (eligibility and classification) roles. Ironically, the 2004 amendments to the Individuals With Disabilities Education Act may move the field away from disability eligibility and classification evaluations into more instructionally relevant assessment paradigms. The emerging professional literature provides insight into strategies for implementing a TTM, but the concomitant role changes for school psychologists require new ways of thinking about established constructs and practices such as informed consent and validity. We recognize the potential value of a TTM to enhance educational outcomes for all students, but improved outcomes will only be achieved if TTM is implemented in a professionally sound manner. Finally, it is important to recognize that school psychologists and others are continually planning, evaluating, and debating approaches to identify children who have a specific LD as defined in IDEIA. The so-called "IDE[I]A Eligibility Mess" (Weber, 2009) is likely to continue to generate research and debate for years to come.

References

Aaron, P. G. (1997). The impending demise of the discrepancy formula. *Review of Educational Research, 67*, 461–502.

Acalanes (CA) Union High School District Office for Civil Rights, Western Division, San Francisco (California). (2009). *109 LRP 32284*. Retrieved from http://www.specialedconnection.com

Ackley, S. (1974). Psychologists and individual rights. *School Psychology Digest, 3*, 21–25.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

American Psychological Association. (1996). Statement on the disclosure of test data. *American Psychologist, 51*, 644–648. doi:10.1037/0003-066X.51.6.644

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2002, amended June 1, 2010)*. Retrieved from http://www.apa.org/ethics/code/index.aspx

Americans With Disabilities Amendments Act of 2008, Pub. L. 110–325. Retrieved from http://www.ada.gov/pubs/adastatute08mark.htm

Armistead, L., Williams, B. B., & Jacob, S. (2011). *Professional ethics for school psychologists: A problem-solving model casebook* (2nd ed.). Bethesda, MD: National Association of School Psychologists.

Baker, S., Gersten, R., & Graham, S. (2003). Teaching expressive writing to students with learning disabilities: Research-based applications and examples. *Journal of Learning Disabilities, 36*, 109–123. doi:10.1177/002221940303600204

Bersoff, D. N. (1974). The ethical practice of school psychology. *School Psychology Digest, 3*, 16–21.

Board of Education of the Hendrick Hudson Central School District v. Rowley, 458 U.S. 176, 102 S. Ct. 3034 (1982).

Brown v. Board of Education, 347 U.S. 483 (1954).

Burns, M. K., & Gibbons, K. (2008). *Response to intervention implementation in elementary and secondary schools: Procedures to assure scientific-based practices.* New York, NY: Routledge.

Burns, M. K., Scholin, S. E., Kosciolek, S., & Livingston, J. (2010). Reliability of decision-making frameworks for response to intervention for reading. *Journal of Psychoeducational Assessment, 28*, 102–114. doi:10.1177/0734282909342374

Burns, M. K., & Senesac, B. K. (2005). Comparison of dual discrepancy criteria for diagnosis of unresponsiveness to intervention. *Journal of School Psychology, 43*, 393–406. doi:10.1016/j.jsp.2005.09.003

Burns, M. K., VanDerHeyden, A. M., & Boice, C. H. (2008). Best practices in delivery of intensive academic interventions. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (5th ed., pp. 1151–1162). Bethesda, MD: National Association of School Psychologists.

Burns, M. K., Wagner, A., & Jacob, S. (2008). Ethical and legal issues associated with using responsiveness-to-intervention to assess learning disabilities. *Journal of School Psychology, 46*, 263–279. doi:10.1016/j.jsp.2007.06.001

Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading

fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review, 35*, 128–133.

Dana, R. H. (2000). Psychological assessment in the diagnosis and treatment of ethnic group members. In J. F. Aponte & J. Wohl (Eds.), *Psychological intervention and cultural diversity* (2nd ed., pp. 59–74). Boston, MA: Allyn & Bacon.

Dean, V. J., & Burns, M. K. (2002). A critical review of including intrinsic processing difficulties in learning disabilities diagnostic models. *Learning Disability Quarterly, 25*, 170–176. doi:10.2307/1511300

Delaware College Preparatory Academy and the Red Clay Consolidated School District Delaware State Educational Agency. (2009). *109 LRP 59893*. Retrieved from http://www.specialedconnection.com

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.

Diana v. State Board of Education, Civ. Act. No. C-70–37 (N.D. Cal., 1970, *further order*, 1973).

Education for All Handicapped Children Act of 1975, Pub. L. No. 94–142, 20 U.S.C. § 1400 *et seq.* renamed the.

Elementary and Secondary Education Act of 1965, Pub. L. No. 89–313.

Epstein, M., Atkins, M., Cullinan, D., Kutash, K., & Weaver, R. (2008). *Reducing behavior problems in the elementary school classroom: A practice guide* (NCEE 2008–012). Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides

Fagan, T. K., & Wise, P. S. (2007). *School psychology: Past, present, and future* (3rd ed.). Bethesda, MD: National Association of School Psychologists.

Family Educational Rights and Privacy Act of 1974, 20 U.S.C. § 1232g; 34 C.F.R. § Part 99.

Fiorello, C. A., Hale, J. B., & Snyder, L. E. (2006). Cognitive hypothesis testing and response to children with reading problems. *Psychology in the Schools, 43,* 835–853. doi:10.1002/pits.20192

Flanagan, D. P. (2000). Wechsler-based CHC cross-battery assessment and reading achievement: Strengthening the validity of interpretations drawn for Wechsler test scores. *School Psychology Quarterly, 15*, 295–329. doi:10.1037/h0088789

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2007). *Learning disabilities: From identification to intervention.* New York, NY: Guilford Press.

Floyd, R. G., Evans, J. J., & McGrew, K. S. (2003). Relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and mathematics achievement across school-age years. *Psychology in the Schools, 40*, 155–171. doi:10.1002/pits.10083

Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research and Practice, 18*, 172–186. doi:10.1111/1540-5826.00073

Fuchs, L. S., Fuchs, D., Craddock, C., Hollenbeck, K. N., Hamlett, C. L., & Schatschneider, C. (2008). Effects of small-group tutoring with and without validated classroom instruction on at-risk students' math problem solving: Are two tiers of prevention better than one? *Journal of Educational Psychology, 100*, 491–509. doi:10.1037/0022-0663.100.3.491

Fuchs, L., Fuchs, D., Hintze, J., & Lembke, E. (2006, July). *Progress monitoring in the context of responsiveness-to-intervention.* Paper presented at the Summer Institute on Progress Monitoring, Kansas City, Missouri.

Garb, H. N., Wood, J. M., Nezworski, M. T., Grove, W. M., & Stejskal, W. J. (2001). Toward a resolution of the Rorschach controversy. *Psychological Assessment, 13*, 433–448. doi:10.1037/1040-3590.13.4.433

Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W. D. (2008). *Assisting students struggling with reading: Response to intervention and multi-tier intervention for reading in the primary grades: A practice guide* (NCEE 2009–4045). Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides

Gickling, E. E., & Havertape, S. (1981). *Curriculum-based assessment (CBA)*. Minneapolis, MN: School Psychology Inservice Training Network.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 42*, 117–135. doi:10.1016/j.jsp.2006.05.005

Hale, J. B., Fiorello, C. A., Bertin, M., & Sherman, R. (2003). Predicting math competency through neuropsychological interpretation of WISC–III variance components. *Journal of Psychoeducational Assessment, 21*, 358–380. doi:10.1177/073428290302100404

Hale, J. B., Fiorello, C. A., Kavanagh, J. A., Hoeppner, J. B., & Gaither, R. A. (2001). WISC–III predictors of academic achievement for children with learning disabilities: Are global and factor scores comparable? *School Psychology Quarterly, 16*, 31–55. doi:10.1521/scpq.16.1.31.19158

Harvey, V. S., & Carlson, J. F. (2003). Ethical and professional issues with computer-related technology. *School Psychology Review, 32*, 92–107.

Health Insurance Portability and Accountability Act of 1996, Pub. L. 104–191, 26 U.S.C. § 294, 42 U.S.C. §§ 201, 1395b-5.

Hubsch, A. W. (1989). Education and self-government: The right to education under state constitutional law. *Journal of Law and Education, 18*, 93–133.

Individuals With Disabilities Education Act of 1990, Pub. L. 101–476, 20 U.S.C. Ch. 33.

Individuals With Disabilities Education Improvement Act of 2004, Pub. L. No. 108–446, 20 U.S.C. § 1400 *et seq.*

Jacob, S. (2009). Putting it all together: Implications for school psychology. *School Psychology Review, 38*, 239–243.

Jacob, S., Decker, D. M., & Hartshorne, T. S. (2011). *Ethics and law for school psychologists* (6th ed.). Hoboken, NJ: Wiley.

Kaplan, M. S., Crisci, P. E., & Farling, W. (1974). Editorial comment [Special issue on ethical and legal issues]. *School Psychology Digest, 3*(1).

Kavale, K. A., & Forness, S. R. (2003). Learning disability as a discipline. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 76–93). New York, NY: Guilford Press.

Marston, D. (2003, December). *Comments on three papers addressing the question: "How many tiers are needed within RtI to achieve acceptable prevention outcomes and to achieve acceptable patterns of LD identification?"* Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention Symposium, Kansas City, Missouri.

Martin, R. (1992). *Continuing challenges in special education law* [looseleaf notebook]. Urbana, IL: Carle Media.

Matarazzo, J. D. (1986). Computerized psychological test interpretations. *American Psychologist, 41*, 14–24. doi:10.1037/0003-066X.41.1.14

Messick, S. (1984). Assessment in context: Appraising student performance in relation to instructional quality. *Educational Researcher, 13*, 3–8.

Mills v. Board of Education of District of Columbia, 348 F. Supp. 866 (1972); *contempt proceedings*, 551 Educ. of the Handicapped L. Rep. 643 (D.D.C. 1980).

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: Principles and application* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Naglieri, J. A. (1999). *Essentials of CAS assessment*. New York, NY: Wiley.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, I., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist*, *59*, 150–162. doi:10.1037/0003-066X.59.3.150

National Association of School Psychologists. (2010). *Principles for professional ethics*. Bethesda, MD: Author. Retrieved from http://www.nasponline.org

National Center on Student Progress Monitoring. (n.d.). *What is progress monitoring?* Retrieved from http://www.studentprogress.org

National Joint Committee on Learning Disabilities. (2010). *Comprehensive assessment and evaluation of students with learning disabilities*. Retrieved from http://www.ldonline.org/about/partners/njcld#reports

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institutes of Child Health and Human Development.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Ortiz, S. O. (2008). Best practices in nondiscriminatory assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 661–678). Bethesda, MD: National Association of School Psychologists.

Paredes Scribner, A. (2002). Best assessment and intervention practices with second language learners. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 337–351). Bethesda, MD: National Association of School Psychologists.

Pennsylvania Association for Retarded Children (P.A.R.C.) v. Commonwealth of Pennsylvania, 334 F. Supp. 1257 (D.C.E.D. Pa. 1971), 343 F. Supp. 279 (E.D. Pa. 1972).

Posner, M. I., & DiGirolamo, G. J. (2000). Cognitive neuroscience: Origins and promise. *Psychological Bulletin, 126*, 873–889. doi:10.1037/0033-2909.126.6.873

Protection of Pupil Rights Act of 1978, 20 U.S.C. § 1232h.

Rehabilitation Act of 1973, Pub. L. No. 93–112, 29 U.S.C. § 701 *et seq.*

Reschly, D. J., & Bersoff, D. N. (1999). Law and school psychology. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of school psychology* (3rd ed., pp. 1077–1112). New York, NY: Wiley.

Rhodes, R. L., Ochoa, S. H., & Ortiz, S. O. (2005). *Assessing culturally and linguistically diverse students: A practical guide*. New York, NY: Guilford Press.

Riley-Tillman, T. C., & Burns, M. K. (2009). *Single case design for measuring response to educational intervention*. New York, NY: Guilford Press.

Rooker, L. S. (2008, June 30). *Response to "Letter to anonymous," 12 FAB 27, 109 LRP 7789, Family Policy Compliance Office*. Horsham, PA: LRP.

Rosenfeld, S. J. (2010). Must school districts provide test protocols to parents? *Communique, 38*(8), 1, 22–26.

Russo, C. J. (2006). *Reutter's The law of public education* (5th ed.). New York, NY: Foundation Press.

Salvia, J., Ysseldyke, J., & Bolt, S. (2010). *Assessment: In special and inclusive education* (11th ed.). Boston, MA: Houghton-Mifflin.

Shapiro, E. S. (2004). *Academic skills problems: Direct assessment and intervention* (3rd ed.). New York, NY: Guilford Press.

Shinn, M. R. (1989). *Curriculum-based assessment: Assessing special children.* New York, NY: Guilford Press.

Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology, 93*, 735–749. doi:10.1037/0022-0663.93.4.735

Speece, D. L., Case, L. P., & Molloy, D. E. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research and Practice, 18*, 147–156. doi:10.1111/1540-5826.00071

Trachtman, G. M. (1974). Ethical issues in school psychology. *School Psychology Digest, 3*, 4–5.

U.S. Department of Health and Human Services & U.S. Department of Education. (2008, November). *Joint guidance on the application of the Family Educational Rights and Privacy Act (FERPA) and the Health Insurance Portability and Accountability Act of 1996 (HIPAA) to student health records.* Retrieved from http://www.ed.gov/policy/gen/guid/fpco/doc/ferpa-hippa-guidance.pdf

Van de Vijver, F. J. R., & Phalet, K. (2004). Assessment in multicultural groups: The role of acculturation. *Applied Psychology, 53*, 215–236. doi:10.1111/j.1464-0597.2004.00169.x

Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41*, 85–120. doi:10.1177/00224669070410020401

Weber, M. C. (2009). The IDEA eligibility mess. *Buffalo Law Review, 57*, 83–160.

Wisconsin v. Constantineau, 400 U.S. 433 (1971).

Ysseldyke, J. (2005). Assessment and decision making for students with learning disabilities: What if this is as good as it gets? *Learning Disability Quarterly, 28*, 125–128. doi:10.2307/1593610

Ysseldyke, J., Burns, M., Dawson, P., Kelley, B., Morrison, D., Ortiz, S., . . . Telzrow, C. (2006). *School psychology: A blueprint for training and practice III.* Bethesda, MD: National Association of School Psychologists.

Ysseldyke, J. E., & Christenson, S. L. (1988). Linking assessment to intervention. In J. L. Graden, J. E. Zins, & M. J. Curtis (Eds.), *Alternative educational delivery systems* (pp. 91–109). Washington, DC: National Association of School Psychologists.

Zirkel, P. (2009). A step-by-step process for § 504/ ADA eligibility determinations: An update. *West's Education Law Reporter, 239*, 333–343.

PART II

EDUCATIONAL TESTING AND MEASUREMENT

CHAPTER 13

THE ASSESSMENT OF APTITUDE

Steven E. Stemler and Robert J. Sternberg

Anyone who has ever been a student, teacher, coach, or parent understands that individuals differ with regard to the speed with which they are capable of learning new information. Some people are "fast learners," others are "slow learners," and most are somewhere in between. Indeed, it is this simple observation that gave rise to efforts to systematically and scientifically measure individual differences in aptitude. Yet, the definition of aptitude itself has proven to be somewhat of a moving target over the years.

The construct of aptitude is often mentioned alongside ability and achievement. Many introductory texts on testing and measurement (e.g., Cohen & Swerdlik, 2005; Gregory, 2007; Kaplan & Saccuzzo, 2009) distinguish among these three terms in roughly the following way: Achievements represent *past* accomplishments or performance, abilities are skills that one can perform right here and now in the *present*, and aptitude reveals an individual's capacity for *future* performance. Although this simple heuristic is relatively useful at a general level, the specific definition of aptitude as well as procedures for assessing it remain hotly contested topics.

Thus, the aim of this chapter is twofold. First, it attempts to describe some of the most controversial elements that serve to define aptitude and highlights areas in the literature where researchers differ. Second, it explores current efforts to assess aptitude, particularly in the context of three domains: (a) college admissions testing, (b) talent identification and personnel selection in employment contexts, and (c) classroom assessment.

WHAT IS APTITUDE?

Disagreements about the definition of aptitude tend to revolve around three central themes. The first of these themes concerns the scope of aptitude. Specifically, is aptitude exclusively a cognitive ability or does it involve noncognitive components as well? The second theme has to do with whether aptitude is something that resides solely within the individual (and which therefore is domain general) or whether it is something that is necessarily the product of a person-by-situation interaction (and therefore is domain specific). Finally, the third major theme is whether aptitude is something that is fixed or modifiable. This section examines the arguments associated with each of these three themes.

Does Aptitude Encompass More Than Just Cognition?

Historical perspectives on aptitude. One of the main concerns of psychologists in the 20th century was the identification of individuals who would be most and least likely to succeed or profit from education. Indeed, the original charge handed down to Alfred Binet from the minister of public instruction in Paris, France, in 1904 was to develop a test of intelligence that would identify children who were not sufficiently profiting from their educational experience so that instructional modifications could be made to help these individuals reach greater levels of success (Birney & Stemler, 2007). The test battery he and his colleague, Theodore Simon, developed ultimately consisted of 30 items measuring everything from simple sensory input and memory

DOI: 10.1037/14049-013
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

to verbal abstractions and social comprehension (Binet, 1916/1905; Gregory, 2007).

As the test traveled across the Atlantic and was imported to the United States by Lewis Terman at Stanford University, an important element of the project was lost. The goal of Terman and his new Stanford–Binet test became the identification of individuals on a linear spectrum of intellectual ability, largely for the purposes of personnel selection rather than remediation or fit to instructional program. It is not coincidental that during roughly the same time period, Charles Spearman (1904) had proposed a general theory (*g* theory) of intelligence, which was based on his observation that levels of performance on many different tests tend to correlate positively—a phenomenon known as positive manifold. Thus, an individual's test scores from a variety of assessments were thought to be related to one another by the concept of an underlying general ability residing within the mind of the test taker. This general ability would manifest itself in an intelligence test score, which could then be used to predict potential for future success.

As a result of the pragmatic predictive success of this approach (Schmidt & Hunter, 1998), the concept of "aptitude became nothing more than the predictions made from conventional ability tests. General aptitude became synonymous with intelligence. Scholastic aptitude became synonymous with verbal and quantitative ability" (Snow, 1992, p. 7). Thus, historically speaking, "the picture of aptitude that most psychologists and educators carried around with them was an entity theory of a fixed, single rank order, general-purpose cognitive trait called intelligence" (Snow, 1992, p. 8; see also Volume 2, Chapter 8, this handbook; Chapter 3, this volume).

Factor analytic studies of aptitude. Formally, Spearman's theory was actually a two-factor theory of intelligence in which *g* was the main focus and explained the majority of observed variance, but a second factor *s* was posited to explain any remaining specific variance. The earliest empirical challenge to this theory came from Thurstone (1938), who approached the analysis of data from a different point of view. Thurstone's approach was to derive factors based on the concept of simple structure, which specified that items would load as highly as possible on only one factor and would have near-zero correlations with all other factors. In doing so, he arrived at a theory of intelligence that specified the existence of seven primary mental abilities: word fluency, verbal comprehension, spatial visualization, number facility, associative memory, reasoning, and perceptual speed (Thurstone, 1938).

Despite the potential incompatibility of the findings from Thurstone and Sperman's two approaches, they found resolution through the specification of a hierarchical factor structure, with a single *g* factor at the top and distinct but correlated subfactors underneath (Brody, 2000). Vernon (1950) simplified the model by proposing a general factor at the top and two broad group factors underneath (v:ed, a verbal-numerical-educational factor, and k:m, a practical-mechanical-spatial-physical factor).

In the 1960s, however, Guilford proposed a nonhierarchical model of intelligence that vastly extended the concept. Specifically, he proposed the existence of three major dimensions (operation, product, and content) on which he could classify any ability test. The operation dimension included five possibilities: cognition (knowing), memory, divergent production (generation of alternatives), convergent production, and evaluation. Each operation could be applied to four different types of content: figural, symbolic, semantic, and behavioral. The application of an operation to a content area could result in one of six products: units, classes, relations, systems, transformations, and implications. Thus, the different combinations could result in a potential of 120 distinct aspects of intelligence. Guilford's model was known as the structure of the intellect (Guilford, 1967), and it represents the most expanded view of intelligence in the field. However, empirical support for the structure of the intellect model has been weak (Brody, 1992).

One of the most important theoretical innovations came from Cattell and Horn's *gf–gc* theory (Horn & Cattell, 1966), which decomposed *g* into two factors: fluid and crystallized ability. Crystallized ability represents an individual's knowledge of information that has been learned, whereas fluid

ability represents an individual's ability to learn. Fluid intelligence is thought to be composed of working memory capacity, processing speed, and inductive reasoning (Kane & Engle, 2002). Fluid ability is very close to the concept of a domain-general aptitude.

Perhaps the most comprehensive contribution within the factor-analytic tradition has come from Carroll's (1993) reanalysis of more than 400 data sets containing cognitive ability test scores. As a result of this massive reanalysis, he arrived at a three-stratum model of intelligence, with a general factor at the top of the hierarchy (stratum III), eight broad factors at stratum II (fluid intelligence, crystallized intelligence, general memory and learning, broad visual perception, broad auditory perception, broad retrieval ability, broad cognitive speediness, and processing speed), and more specific factors at stratum I.

The relation of aptitude to noncognitive factors. Although widely accepted by researchers, the concept of aptitude as a single, cognitively oriented entity was not universally adopted. A second perspective on aptitude, most forcefully advocated by Snow (1977, 1978, 1992) was that the concept of aptitude is not limited to cognitive ability. Rather, he argued, other aspects, such as personality, motivation, and self-concept are also important components of aptitude. Consequently, according to Snow, aptitude consists not only of cognitive factors but also of affective and conative processes. To be clear, within this context, cognitive components refer to analysis and interpretation. Affective components refer to emotions and feelings. Conative components refer to goal setting and will.

Historically speaking, this broader definition of aptitude was consistently advocated by key figures in the field. Both Alfred Binet and David Wechsler, originators of the two most widely used intelligence tests in existence to date, supported this broader view that cognition was indelibly linked to feelings and attitudes (Corno, Cronbach, Kupermintz, & Lohman, 2001).

Thus, the idea that aptitude was composed of cognitive, affective, and conative aspects certainly had theoretical appeal even from the very earliest conceptions of aptitude testing. Researchers such as L. L. Thurstone attempted to develop tests of broader abilities that would yield profile scores. Unfortunately, however, "combining diverse scores into a prediction formula increased the power to predict grade average and other broad indices of success over the predictive power of a full-length 'general' test by only a discouragingly small amount" (Corno et al., 2001, p. 16). Questions remained, however, as to whether this disappointing result was attributable to a flawed theory of aptitude as broader than *g* or whether the result could be explained by technical limitations of the way the tests were operationalized.

The present era has witnessed a resurgence in interest in the measurement of so-called noncognitive factors for predictive purposes (Kyllonen, Roberts, & Stankov, 2008). Noncognitive factors include such constructs as personality dimensions, time management, self-concept, intercultural sensitivity, and motivation. These constructs are presently being investigated for possible operational use in employment and admissions testing by the Educational Testing Service (Kyllonen, 2005; Kyllonen et al., 2008; see also Chapters 14, 15, and 19, this volume). In addition, Silzer and Church (2009a) recently surveyed more than 100 professionals in organizations and consulting firms who had written on, presented on, or been involved with programs aimed at identifying high-potential employees (i.e., high aptitude) and found substantial overlap among organizations with regard to the key factors used in their approach to identifying potential. These factors included "cognitive skills, personality variables, learning variables, leadership skills, motivational variables, performance records, and other factors" (Silzer & Church, 2009a, p. 391).

Another perspective in the literature that lies somewhere between the *g*-based perspective on aptitude and the measurement of noncognitive abilities as additional components of aptitude may be found in Sternberg's (1985, 1997, 2005) theory of successful intelligence, which focuses on cognitive abilities but proposes a broader range of cognitive abilities than those measured by *g*. The theory argues that *successful intelligence* is a person's ability

to achieve his or her goals in life, within his or her sociocultural context, by capitalizing on strengths and correcting or compensating for weaknesses, to adapt to, shape, and select environments through a combination of analytical, creative, and practical skills (Sternberg, 2009).

Successful intelligence therefore conceptualizes cognitive ability in a broader way than *g* theory does and also acknowledges the important role of what some would call noncognitive factors. A variety of empirical studies have yielded multiple sources of validity evidence (e.g., content, construct, criterion) that support the theory (Stemler, Grigorenko, Jarvin, & Sternberg, 2006; Stemler, Sternberg, Grigorenko, Jarvin, & Sharpes, 2009; Sternberg, Ferrari, Clinkenbeard, & Grigorenko, 1996; Sternberg, Grigorenko, Ferrari, & Clinkenbeard, 1999). In addition, the theory has led to practical test results that have overcome the problems of incremental predictive validity encountered by previous efforts to measure scholastic aptitude in broader ways. More is said about this later in this chapter (see the section College Admissions Testing).

Is Aptitude a Personal Trait or a Person-by-Situation Variable?

Perhaps the most interesting of the three central themes in defining aptitude is the question of whether aptitude is a personal trait that resides within an individual or whether aptitude can be fully understood only by examining the interaction between the person and the situation.

The viewpoint that aptitude is a rather stable trait that resides within an individual carries with it the implication of an ability that is domain general. It is this point of view that gives rise to the idea of the "fast learner" versus the "slow learner." Simply by saying that some people are "fast learners," we are making the assumption that their rate of learning will be relatively stable regardless of what it is they are being asked to learn. Thus, individuals who are fast learners in mathematics will also be fast learners on the athletic field and will be quick to learn how to play musical instruments and so on. Regardless of the content they are learning, we are assuming that there is something inherent in their mind that represents a relatively fixed cognitive ability that allows them to quickly identify new patterns and rules in any domain of interest.

Corno et al. (2001) persuasively argued that the conception of aptitude as a fixed trait of an individual was the result of a misinterpretation of Darwin's theory of natural selection—a theory that had argued that adaptation is the result of a match between organism and environment (i.e., person and situation). Instead, Herbert Spencer, in coining the term *social Darwinism*, mistakenly interpreted Darwin's work and widely disseminated the notion that intelligence (and therefore, aptitude) was a domain-general trait residing within an individual. Thus, the argument proceeded, certain individuals were simply more likely than others to adapt to any set of circumstances they might encounter.

An alternative to the trait perspective is the point of view advocated by researchers who suggest that one cannot understand the concept of aptitude without understanding the context of what is being assessed (Cronbach, 1957; Silzer & Church, 2009a, 2009b; Snow, 1977, 1978). For example, an individual may exhibit a rapid rate of acquisition of knowledge in one domain (e.g., physics) but be a hopelessly slow learner in another domain (e.g., the piano). In such a case, the concept of aptitude only makes sense when discussed relative to the situational context (e.g., an aptitude for physics). This point of view assumes that there is no single cognitive trait inside the minds of individuals that will enable the prediction of rate of learning across all content domains and all time periods. Thus, advocates of the person-by-situation position are aligned with a position known as situated cognition. They believe that there are as many different aptitudes as there are situational contexts and therefore eschew the search for general structures of knowledge that hold across person, situation, and time.

In his classic American Psychological Association (APA) presidential address, Cronbach (1957) outlined the limitations inherent in what he called the two disciplines of psychology (experimental and differential). Specifically, he noted that experimental research on instructional situations often ignores variations in the aptitude of participants, whereas correlational studies of differential psychologists

tend to ignore important situational variation. Thus, he proposed the idea of uniting these two branches of psychology via the study of aptitude–treatment interactions (ATIs). Cronbach stated,

> An aptitude, in this context, is a complex of personal characteristics that accounts for an individual's end state after a particular educational treatment . . . [It] includes whatever promotes . . . survival in a particular educational environment, and it may have as much to do with styles of thought and personality variables as with the abilities covered in conventional tests . . . Such a theory deals with aptitude-treatment interactions." (Cronbach, 1967, pp. 23–24, 30, paragraph altered as cited in Corno et al., 2001, p. 20)

Although the concept of ATIs constituted a new theory of aptitude that challenged the conventional conception of aptitude as an individual trait, empirical support for ATI is still contested. Pashler, McDaniel, Rohrer, and Bjork (2008) have argued that there still is only weak evidence for ATIs; however, Sternberg, Grigorenko, and Zhang (2008a, 2008b) presented an alternative point of view. Some results that support the ATI concept are presented later in this chapter (see the section Classroom Assessment).

Is Aptitude Fixed or Modifiable?

A major debate with regard to the definition of aptitude relates to whether aptitude is something that is stable over time or whether it is modifiable, so that people are able to enhance their aptitude. The traditional psychometric view of aptitude is that it is a relatively fixed trait. According to Silzer and Church (2009a), this perspective is currently widely held by many leaders, managers, and human relations professionals who view the concept of potential as an innate individual capacity (e.g., one has a potential to a certain degree and that degree is not changeable).

An alternative perspective holds that aptitude is malleable. Two noteworthy theorists independently advocating this perspective were the Russian psychologist Lev Vygotsky and the Israeli psychologist Reuven Feuerstein. Vygotsky (1934/1978) introduced the concept of the zone of proximal development, which is the difference between the level of performance that is attainable by an individual on his or her own as compared with the performance that same individual can achieve when aided by someone more knowledgeable or experienced in the domain. The zone of proximal development varies between individuals, such that two people may profit differentially from outside help, and it also varies within individuals, such that the zone may be larger for a given individual in some domains (e.g., music) than in others (e.g., writing; Fabio, 2005). Thus, to Vygotsky, one could not fully understand an individual's potential by looking at scores on a static test of ability. A one-shot test can give us, at best, a snapshot of where an individual currently stands (i.e., his or her ability), but it tells us little about the comparative aptitude (i.e., potential for future performance) of two individuals with the same score. Rather, to assess aptitude more precisely, one must give individuals opportunities to demonstrate how quickly they can grasp new concepts with the aid of a more knowledgeable guide.

In a similar vein, Feuerstein and his colleagues advocated for the importance of what they have called *mediated learning experience* (Feuerstein, Rand, & Hoffman, 1979). Mediated learning experiences are conceptually similar to the zone of proximal development in that they require a more knowledgeable mentor to mediate between a performer and a task by guiding the performer along a scaffolded developmental path toward deeper understanding (Feuerstein, Klein, & Tannenbaum, 1991). What happens during the context of this mediated learning experience is that a qualitative change takes place in the individual's cognitive structure (Birney, 2003).

Silzer and Church (2009a) have proposed somewhat of a compromise position between the concept of aptitude as fixed versus aptitude as modifiable, suggesting that certain components of aptitude represent foundational dimensions, whereas other components represent growth dimensions. According to their proposed structure, foundational dimensions are relatively stable traits that include such components as strategic thinking, dealing with

complexity, and interpersonal skills, whereas growth dimensions include components that individuals can develop and expand, including openness to feedback, risk-taking, and achievement orientation.

Although the debate about whether aptitude is fixed or malleable has not yet reached an empirical conclusion, an important related question has been explored by Dweck (2006). In particular, her research focused on individuals' beliefs—what she calls mind-sets—about whether aptitudes are fixed or modifiable. A rather substantial and growing body of research demonstrates that this question has important practical relevance.

Individuals with a fixed mind-set tend to cling more readily to their first impressions of individuals and to believe that those impressions will accurately predict future behaviors (Erdley & Dweck, 1993). Furthermore, they tend to avoid or ignore subsequent information they receive that contradicts their initial impressions, creating a sort of self-fulfilling prophesy in terms of their beliefs about individuals (Gervey, Chiu, Hong, & Dweck, 1999; Plaks, Stroessner, Dweck, & Sherman, 2001). A fixed mind-set is particularly problematic when one is attempting to assess aptitude or potential because, as Heslin, Latham, and VanderWalle (2005) found, managers with fixed mind-sets are more likely to miss potential or even to misidentify those with low potential (i.e., false alarms) based on the rigidity of their initial impressions and their reluctance to take into account additional performance information.

The different points of view with regard to the three questions posed in this section lead directly to different approaches to the assessment of aptitude, as discussed in further detail in the next section of this chapter.

HOW IS APTITUDE ASSESSED?

Although many fields and professions have contributed to the definition and assessment of aptitude, three domains in particular have demonstrated a persistent concern with the construct. These domains are (a) college admissions testing; (b) employment testing, particularly in the context of talent identification and personnel selection; and (c) classroom assessment, particularly as it relates to ATIs and dynamic-assessment techniques. In this second section of the chapter, we illustrate the way in which the answers to the three key questions posed in the first section of this chapter have practical consequences for the assessment of aptitude.

College Admissions Testing

Philosophical paradigms. As Lemann (2000) has pointed out in his excellent history of the SAT, in the early 20th century a large-scale college admissions test could be based on basically four distinct paradigms. The first paradigm was associated with the philosophy of progressive education advocated by John Dewey (1916, 1938). The goal of individuals in this camp was to develop liberal-minded, free-thinking, and tolerant thinkers. They believed that the best route by which to accomplish this goal was to let schools set their own curricula. Thus, from an admissions perspective, what would be required is a test to determine which students across all of the schools had best developed a broad range of important intellectual abilities.

By contrast, the second paradigm came from individuals such as Ben Wood, who were concerned with a strict, standards-based approach to education. Advocates of this position wanted a standardized curriculum across all schools and felt that admissions tests ought to be based primarily on student achievement, which was of course to be aligned with the curriculum. The descendants of this philosophy are making a strong resurgence in the present day with the current push by the federal government for states to adopt "common core standards" on which all students may be tested and compared (U.S. Department of Education, 2010).

The third paradigm was found in the philosophy of educational expansionists, such as George Zook, who believed that the proper role of testing was to identify students in need of remediation. From this perspective, the goal was education for all and the ultimate goal of testing should be to determine the best fit between an individual and the kind of education that will allow that individual to progress up the developmental ladder. This conception aligns

well with the notion of aptitude as a situation-specific and modifiable trait.

The final paradigm, and the one that eventually won out, was drawn from those individuals who were believers in intelligence testing. The goal of individuals in this camp, which included Educational Testing Service founder Henry Chauncey, was to identify students who would be best able to profit from higher education by selecting those with the highest set of test scores on a new, scholastically oriented, intelligence-type test. The test that was developed based on this philosophy originally was known by its full title as the Scholastic Aptitude Test. The model of aptitude on which the SAT was based was that aptitude consists of only a narrow range of cognitive abilities (specifically verbal and quantitative reasoning), that it was a domain-general trait residing within a person, and that it was relatively fixed. Furthermore, the test was atheoretical and concerned primarily with its power to predict college grades.

Critics of the static testing procedures employed in this context were quick to point out that these tests emphasize previously acquired knowledge (e.g., vocabulary) and do not typically assess how an individual responds to changing circumstances or modifications to the test aimed at increasing levels of performance (Brown & French, 1979; Carlson & Wiedl, 1992; Fabio, 2005; Feuerstein et al., 1979). After some years of debate about whether the SAT could best be described as a measure of aptitude, ability, achievement, some combination thereof, or a subset thereof, the test developers have abandoned the concept of using the SAT as an acronym for its larger descriptive title (i.e., Scholastic Aptitude Test) and have instead settled on simply calling the test the SAT.

What is interesting to consider, however, is what a related type of test would look like that is based on a different conception of aptitude. For example, consider a test that is theoretically based, that conceives of aptitude as encompassing a broader set of cognitive or noncognitive skills beyond those measured in a *g*-type assessment, that assumes aptitude involves an interaction between the person and the environment, and that conceives of and measures aptitude as a modifiable quantity, rather than as a fixed entity.

Sternberg and colleagues (Sternberg & the Rainbow Project Collaborators, 2006) developed a novel assessment containing at least some of those features as a supplement to the standard SAT. Their aim was to develop an aptitude test for college admissions that was theoretically based and that measured a broader range of skills than is currently assessed by the SAT. It further would allow individuals to capitalize on their strengths and compensate for their weaknesses within the context of the test. Their test was developed as part of what was known as the Rainbow Project.

The Rainbow Project. The goal of the Rainbow Project was not to replace the SAT but rather to devise tests that would supplement the SAT, measuring cognitive skills that the SAT does not measure, as outlined by Sternberg's theory of successful intelligence (Sternberg, 1997). In addition to multiple-choice tests, the test used three additional measures of creative skills and three additional measures of practical skills.

Creative skills were measured by using a cartoon captioning task, written stories, and oral stories. On the cartoon task, participants were given five cartoons purchased from the archives of the *New Yorker* with the captions removed. The participant's task was to choose three cartoons and to provide a caption for each cartoon. Two trained judges rated all the cartoon captions for cleverness, humor, and originality. A combined creativity score was formed by summing the individual ratings on each dimension. Next, participants were asked to write two stories, spending about 15 min on each, choosing from the following titles: "A Fifth Chance," "2983," "Beyond the Edge," "The Octopus's Sneakers," "It's Moving Backwards," and "Not Enough Time."

A team of four judges was trained to rate the stories for originality, complexity, emotional evocativeness, and descriptiveness. Finally, participants were presented with five sheets of paper, each containing a set of pictures linked by a common theme. For example, participants might receive a sheet of paper with images of a musical theme, a money theme, or a travel theme. Each participant then chose one of the pages and was given 15 min to

formulate a short story and dictate it into a cassette recorder. The dictation period was not to be more than 5 min long. The process was then repeated with another sheet of images so that each participant dictated a total of two oral stories. Six judges were trained to rate the stories for originality, complexity, emotional evocativeness, and descriptiveness.

Practical skills were measured by using three different types of situational–judgment tests. On the first test, participants were shown a series of seven video-based vignettes designed to capture problems encountered in general, everyday life, such as determining what to do when one is asked to write a letter of recommendation for someone one does not know particularly well. On the second test, participants were given a written description of 15 vignettes designed to capture problems encountered in general business-related situations, such as managing tedious tasks or handling a competitive work situation. On the third test, a written inventory presented participants with 15 vignettes that captured problems encountered in general college-related situations, such as handling trips to the bursar's office or dealing with a difficult roommate. In all cases, the vignettes were followed by a variety of different options for how to handle the situation, and participants were asked to rate the quality of each potential response. Participant responses were then scored based on their distance from the group consensus as to the quality of each response.

A total of 1,015 students at 15 different institutions (13 colleges and two high schools) were tested with this new measure. The results showed that these tests significantly and substantially improved on the validity of the SAT for predicting 1st-year college grades (Sternberg & the Rainbow Project Collaborators, 2006), doubling prediction over the SAT alone, and increasing prediction by 50% over SAT and high school grade point average. The test also improved equity: Using the test to admit a class would result in greater ethnic diversity than would using just the SAT or just the SAT and grade point average. In addition, differences in achievement between White students and African American students were reduced on measures of creative skills, and differences in achievement between White and Latino students were greatly reduced on assessments that emphasized practical skills and creative skills.

One of the main contributions of the Rainbow Project is that it demonstrates that universities potentially can do a better job of predicting who is likely to succeed in college (i.e., who has "more" scholastic aptitude) when a broader range of skills are systematically assessed. Furthermore, universities could be in a better position to select an optimal mix of students of diverse skills, which can be particularly beneficial for ethnic-minority students who tend to perform better at these broader skills that traditionally have been undervalued in terms of assessment but that are highly valued in the university and workforce settings.

Further efforts to develop college-admissions tests that capture a broad range of cognitive and noncognitive skills are being vigorously pursued by several different research groups (Kyllonen et al., 2008; Schmitt, Oswald, & Gillespie, 2005; Schmitt et al., 2007; Stemler, 2012; Sternberg, Bonney, Gabora, Karelitz, & Coffin, 2010).

Employment Testing

A second area in which the assessment of aptitude has had a strong historical connection has been within the area of employment testing and personnel selection. There are two areas within this domain that are of particular interest. The first is within the context of the military and the second is within the context of private sector organizations.

Military testing. Some of the earliest systematic and large-scale standardized efforts to measure aptitude in the United States emerged within the context of the military. Although the relationship between large-scale standardized testing for military recruitment purposes and the development of the SAT is widely recognized (Gregory, 2007; Kaplan & Saccuzzo, 2009; Lemann, 2000), what is less commonly recognized is where those two programs diverge. One of the ways they began to diverge was with regard to the fundamental definition of aptitude that each adopted.

During World War II, a collection of aptitude tests was developed to select among men in the Army who applied for pilot training. Toward the

end of the war, DuBois (1947) evaluated test data from a 2-year period. What he found was that the aptitude test battery was highly predictive of graduation from pilot training school, with percentage passing correlating highly with a specific test score. Furthermore, he found that adding reading and mathematics tests to the composite failed to improve correlation with graduation from pilot training (Corno et al., 2001).

Thus, the concept of aptitude as a domain-specific construct that requires attention to the fit between the person and the demands of his or her occupation came to dominate the military definition of aptitude. Indeed, the Armed Services Vocational Aptitude Battery (ASVAB) is now touted by the military as "the most widely used multiple-aptitude test battery in the world" (Today's Military, 2012). Underlying this test is a different conception of aptitude than the large-scale tests historically used for college-admissions purposes, which tend to consider aptitude as a domain-general capacity of an individual.

The ASVAB includes 10 subtests. These subtests include General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), Electronics Information (EI), and Sum of Word Knowledge and Paragraph Comprehension (VE). Scores from each of these subtests are combined in unique ways to determine the potential of the applicant to succeed in a variety of job positions within each of the branches of the military. The Army creates different subscale scores (line scores) for different professions. For example, positions classified as Clerical require a combination of three tests (VE + AR + MK). Combat positions require a minimum score for the combination of AR + CS + AS + MC. Electronics positions require a minimum score for GS + AR + MK + EI, and so on. More specifically, different positions have different required cut scores. For example, a Special Forces Weapons Sergeant requires a minimum score of 110 on the General Technical subscale (VE + AR) and 100 on the Combat subscale.

Recent research with regard to personnel selection among higher level officers in the military has focused on the development of tests that measure a broad range of cognitive and noncognitive factors as well, including capacities such as adaptability and mental flexibility (Matthew, Beckmann, & Sternberg, 2008; Matthew & Stemler, 2008; Pulakos et al., 2002; Mueller-Hanson, Swartout, Hilton, & Nelson, 2009; Stemler, 2009). Thus, just as the trend in college admissions testing has been to expand the conception of aptitude, it appears that so too in the military context, the definition of aptitude as inclusive of both cognitive and noncognitive factors may be gaining momentum.

Talent identification in the private sector. One interesting question to consider is the way in which the answer to the question of "potential for what?" changes when one views the organization, rather than the individual, as the relevant unit of analysis. As Yost and Chang (2009) pointed out, human resource professionals often think of assessing aptitude within the context of a hierarchical organizational structure in which the aim is to identify potential along a particular dimension (e.g., leadership potential). When one steps back and looks at it from an organizational perspective, however, the main purpose of talent identification is to identify unique profiles of strengths and weaknesses among employees to fulfill a variety of different job roles in an organization. Thus, the aim of the organization is to increase the overall potential of the system.

Indeed, as Yost and Chang (2009) wrote,

> In today's business landscape, the environments in which organizations operate are often so dynamic that investing in only a few people for a limited set of roles is risky. As the last two decades have shown, organizational strategies can change quickly and dramatically, requiring a completely different talent mix to meet future challenges. In dynamic markets, organizations can't always anticipate the challenges they will face and the people they will need in order to compete. (p. 442)

Thus, rather than simply using aptitude assessment to identify leadership potential within an organization, a more comprehensive approach to

aptitude assessment involves identifying and capitalizing on the unique blend of strengths and weaknesses of the organization's employees.

Mone, Acritani, and Eisinger (2009) have cautioned, however, that managers sometimes confuse the assessment of current skills that foster immediate promotability with the kinds of traits that are important for long-term potential in a future role. They suggested separating out these two assessments: "More specifically, stating that exclusively looking at current performance over time does not predict success in advanced roles has helped managers broaden their scope and more accurately assess potential" (p. 427).

In their recent focal article in the journal *Industrial and Organizational Psychology*, Silzer and Church (2009a) illustrated that many organizations take the perspective that the search for potential and aptitude involves the observation of a person-by-situation interaction. These authors suggested that, in practice, employees may need to be given a range of bosses and jobs for their underlying potential to be fully assessed. Doing so will help to address questions such as how quickly an employee can adapt to a particular set of demands, a particular kind of task, or a particular type of supervisor.

Certainly one does not have to dig deeply into the literature on social psychology to understand the profound influence the situation can have on the performance of an individual. In a classic study of the Pygmalion effect, Rosenthal and Jacobson (1968/1992) randomly assigned students to one of two classrooms. The students showed similar levels of achievement before instruction; however, the teacher in one classroom was told that the students had been identified by the assessments as "gifted," whereas the teacher in the control classroom had not been told anything. Stunningly, the results showed that upon posttest, the students in the so-called gifted classroom actually significantly outperformed students in the control classroom. The differences were largely attributed to the way in which teachers interacted with students whom they thought had been identified as high potential.

Although the Rosenthal and Jacobson (1968/1992) study has been criticized on methodological grounds, the general findings have been replicated by researchers employing more rigorous techniques (Weinstein, 2002). For example, a study by Eden and Shani (1982) replicated this finding in the business world and found that trainees who were labeled as having great potential for high performance before a simulated training session outperformed those who were not labeled as high potential. As Heslin (2009) has pointed out, however, it is still not entirely clear how this mechanism works. Studies such as those by Silzer and Church (2009a) do not elaborate on the kinds of bosses that are likely to recognize or to overlook their employee's potential. Nevertheless, these studies and others do demonstrate the important influence of situational characteristics on performance. Therefore, it is perhaps not surprising that many individuals in the private sector appear committed to the notion that aptitude involves a person-by-situation interaction.

Classroom Assessment

A third domain in which the assessment of aptitude has been of keen interest has been within the area of classroom assessment. Two lines of research that are particularly relevant to the discussion of aptitude within this domain are ATI studies and dynamic-assessment research.

ATIs. According to the theory of successful intelligence (Sternberg, 1997), different students have different combinations of cognitive skills (e.g., analytic, creative, and practical). Furthermore, the theory is based on the notion that students learn in different ways—that they have different styles of learning (Sternberg, Grigorenko, & Zhang, 2008a, 2008b), just as teachers have different styles of teaching (Spear & Sternberg, 1987).

Teaching for analytical thinking means encouraging students to (a) analyze, (b) critique, (c) judge, (d) compare and contrast, (e) evaluate, and (f) assess. When teachers refer to teaching for "critical thinking," some of them may mean teaching for analytical thinking. An example of an exercise developing such skills would be to ask students to

compare and contrast two works of literature, to evaluate the conclusions drawn from a scientific experiment, or to critique a work of art.

Teaching for creative thinking means encouraging students to (a) create, (b) invent, (c) discover, (d) imagine if . . . , (e) suppose that . . . , and (f) predict. Teaching for creative thinking requires teachers not only to support and encourage creativity but also to act as a role model and to reward creativity when it is displayed (Sternberg & Lubart, 1995; Sternberg & Williams, 1996). Examples of teaching activities might include asking students to design a psychological experiment to test an hypothesis, to invent an alternative ending for a story they have read, or to create a mathematics problem.

Teaching for practical thinking means encouraging students to (a) apply, (b) use, (c) put into practice, (d) implement, (e) employ, and (f) render practical what they know. Such teaching must relate to the real practical needs of the students, not just to what would be practical for individuals other than the students (Sternberg et al., 2000). Examples might include asking students to apply what they have read in a story to their life, use their knowledge of mathematics to balance a checkbook, or persuade someone that an argument they are employing is sound.

To validate the relevance of the theory of successful intelligence in the classroom, researchers have carried out a number of instructional studies with different age-groups and subject matters. (Other kinds of research support are summarized in Sternberg, 1985, 1997, 2003b.)

In one study (Sternberg et al., 1999), the investigators used the Sternberg Triarchic Abilities Test (Sternberg, 2003a), which assesses analytical, creative, and practical skills through multiple-choice and essay items. The test was administered to 326 children across the United States and in some other countries who were identified by their schools as gifted by any standard whatsoever. Children were selected for a summer program in (college-level) psychology if they fell into one of five ability groupings: high analytical, high creative, high practical, high balanced (high in all three abilities), or low balanced (low in all three abilities).

The high school students who came to Yale were then divided into four instructional groups. Students in all four instructional groups used the same introductory psychology textbook (a preliminary version of Sternberg, 1995) and listened to the same psychology lectures. What differed among them was the type of afternoon discussion section to which they were assigned. They were assigned to an instructional condition that emphasized either memory, analytical, creative, or practical instruction. For example, in the memory condition, they might be asked to describe the main tenets of a major theory of depression. In the analytical condition, they might be asked to compare and contrast two theories of depression. In the creative condition, they might be asked to formulate their own theory of depression. In the practical condition, they might be asked how they could use what they had learned about depression to help a friend who was depressed.

Students in all four instructional conditions were evaluated in terms of their performance on homework, a midterm exam, a final exam, and an independent project. Each type of work was evaluated for memory, analytical, creative, and practical quality. Thus, all students were evaluated in exactly the same way.

The results showed that there was an aptitude-treatment interaction whereby students who were placed in instructional conditions that better matched their pattern of abilities outperformed students who were mismatched. In other words, when students are taught at least some of the time in a way that fits how they think, they do better in school. These results suggest that the negative Cronbach and Snow (1977) results for ATIs may have been due to lack of theoretical basis for instruction or theoretical match between instruction and assessment.

Dynamic assessment. Although the concept of dynamic assessment predates the concept of ATI, it has only been much more recently that scholars have attempted to empirically evaluate dynamic assessment procedures as a method for assessing aptitude (Grigorenko & Sternberg, 1998; Haywood & Lidz, 2007; Lidz & Elliott, 2000;

Sternberg & Grigorenko, 2002; see also Chapter 7, this volume).

The basic premise of dynamic assessment is that one cannot truly understand or assess the aptitude of an individual simply by administering a test at one time point and interpreting that test score. Rather, dynamic assessment is based on the theoretical work of Vygotsky (1934/1978) and Feuerstein and Feuerstein (1994), mentioned earlier, both of whom noted the importance of assessing an individual at more than one point in time and comparing the performance of individuals when they are alone with the performance of those same individuals when they are guided by more knowledgeable others. The chief interest of psychologists and educators engaging in dynamic assessment is not where the test takers are now, given their previous educational experience, but where they can be tomorrow, assuming that they are given adequate educational intervention from now on (Grigorenko, 2009).

As Elliott (2003) has noted, *dynamic assessment* is an "umbrella term used to describe a heterogeneous range of approaches" (p. 16). Some advocates of dynamic assessment conceive of aptitude as domain general (Feuerstein et al., 1979), whereas others believe that aptitude is domain specific (Camilleri, 2005; Guthke, 1992). In practice, there are four main approaches to dynamic assessment (Jeltova et al., 2007).

The first approach is referred to as the test–teach–retest approach. This approach is not so far removed from the procedures invoked by many classroom teachers in the 21st century. The difference is often in the level of detail that occurs at the instructional level. Strictly speaking, the test–teach–retest approach, associated primarily with the work of Budoff (1987), involves protocols for pointing out errors that can be developed that are standardized and even automated.

A second approach to dynamic testing has been dubbed the learning-test approach (Beckmann & Guthke, 1995, 1999; Guthke, 1992). Under this model, the participants are given a pre- and posttest with an intervention in between; however, the procedure extends the previous approach by offering a sequential construction of what information and skills are needed to ensure a successful solution. Furthermore, qualitative analyses of errors are used to diagnose learning processes. Although Guthke, Beckmann, and Dobat (1997) found the results from the learning test to be better predictors of knowledge acquisition and knowledge application in the context of complex performance, Hessles and colleagues (Hessles & Hamers, 1993; Hamers, Hessles, & Pennings, 1996) found no increase in the predictive power of learning tests over traditional tests in their sample of test takers.

A third approach to dynamic testing is called the graduated-prompt approach (Campione & Brown, 1987). The idea behind this model is to give the participant a pretest, a hinted stage, a posttest, and a hint-assisted posttest. This procedure has been shown to be predictive of school readiness for students who are ready to respond to intervention in language production (Olswang & Bain, 1996). As Jeltova et al. (2007) pointed out, however, this approach has a few technical problems. One main criticism is that hints differ in helpfulness across students, so there is some difficulty inherent in trying to standardize the utility of different types of hints.

Finally, the fourth main approach to dynamic testing is called testing the limits (Carlson & Wiedl, 1992). The key objective under this model is to find the best match between the individual and the test situation that will evoke the best possible performance. Thus, this approach is a highly person-by-situation-oriented perspective on the concept of aptitude.

Each of the main approaches to dynamic assessment described here share in common the belief that aptitude cannot be assessed as a fixed, latent trait of an individual that is revealed within the context of a one-shot static test. In general, researchers in this area tend to conceptualize aptitude as largely a cognitive capacity that is situation specific and that is malleable.

DISCUSSION

This chapter has outlined three major debates regarding the definition of aptitude. The first of these asks whether aptitude consists of only cognitive elements or whether it also includes noncognitive elements. Debates around this issue have been

played out largely within the domain of college admissions testing. Although there remains some disagreement about this matter, many major theorists in the field in the 21st century suggest that aptitude includes cognition as well as other noncognitive components. Where theorists tend to diverge is largely with regard to how many other components they include under this umbrella.

The second question, which relates to whether aptitude itself is a rather domain-general trait of individuals or whether it is a domain-specific product of a person-by-situation interaction, has largely been emphasized within the field of employment testing. Disagreements on this question can be viewed with respect to a lock-and-key metaphor of aptitude. Theorists advocating a person-by-situation perspective of aptitude view situational characteristics as a lock and personal attributes as the key that opens the door. By contrast, theorists who view aptitude as a domain-general construct that can be found within an individual without much regard to situational factors view personal aptitude as a skeleton key that will open any situational door.

The third major distinction with regard to aptitude is whether it is fixed or modifiable. Recently, the question as to whether aptitude is fixed or malleable has been most heavily emphasized within the context of classroom assessment. Although some traditional theories have tended to conceive of aptitude as a fixed entity, new advances in technology are beginning to enable assessments that allow for more dynamic testing of individuals and that will open new avenues to those who view aptitude as modifiable and something that is best assessed dynamically. Furthermore, recent research by Dweck (2006) and her colleagues has demonstrated the powerful influence that even the mind-set one has with regard to the question of modifiability of aptitude can have profound consequences across a variety of situational contexts, including teaching, employment, and sports.

Different points of view on these fundamental debates in the definition of aptitude will lead to (and have led to) very different approaches to the assessment of aptitude. And perhaps not surprisingly, each approach contains both advantages and disadvantages. Some triumphs of the different approaches include their success at predicting desired outcomes with relative accuracy, their ease of administration, and the information they yield for the user. Some of the limitations of these procedures include the problem of upscaling dynamic assessments to make them group administered and the problems associated with labeling individuals who are identified or misidentified as high potential. With the range of different approaches to measuring aptitude that currently exist and the likelihood that further techniques will be developed that align with different perspectives on the nature of aptitude, there is good reason to be optimistic about the future for aptitude assessment.

References

Beckmann, J. F., & Guthke, J. (1995). Complex problem solving, intelligence status, and learning ability. In J. Funke & P. Frensch (Eds.), *Complex problem solving: The European perspective* (pp. 177–200). Hillsdale, NJ: Erlbaum.

Beckmann, J. F., & Guthke, J. (1999). *Assessing reasoning ability*. Gottingen, Germany: Hogrefe.

Binet, A. (1916/1905). *New methods for the diagnosis of the intellectual level of subnormals. Classics in the history of psychology*. Retrieved from http://psychclassics.yorku.ca/Binet/binet1.htm

Birney, D. (2003). Mediating the impact of mediated learning: Review of "Experience of mediated learning: An impact of Feuerstein's theory in education and psychology" by A. Kozulin and Y. Rand (Eds.). *Contemporary Psychology: APA Review of Books, 48*, 677–679.

Birney, D., & Stemler, S. E. (2007). Intelligence quotient. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics* (Vol. 2, pp. 473–476). Thousand Oaks, CA: Sage.

Brody, N. (1992). *Intelligence* (2nd ed.). San Diego, CA: Academic Press.

Brody, N. (2000). History of theories and measurements of intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 16–33). New York, NY: Cambridge University Press.

Brown, A. L., & French, L. A. (1979). The zone of potential development: Implications for intelligence testing in the year 2000. *Intelligence, 3*, 255–273. doi:10.1016/0160-2896(79)90021-7

Budoff, M. (1987). The validity of learning potential assessment. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 54–81). New York, NY: Guilford Press.

Camilleri, B. (2005). Dynamic assessment and intervention: Improving children's narrative abilities. *International Journal of Language and Communication Disorders, 40*, 240–242.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic testing with school achievement. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 82–115). New York, NY: Guilford Press.

Carlson, J., & Wiedl, K. H. (1992). The dynamic assessment of intelligence. In H. C. Haywood & D. Tzuriel (Eds.), *Interactive assessment* (pp. 167–186). New York, NY: Springer-Verlag.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytical studies*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511571312

Cohen, R. J., & Swerdlik, M. E. (2005). *Psychological testing and assessment: An introduction to tests and measurements* (6th ed.). Boston, MA: McGraw-Hill.

Corno, L., Cronbach, L. J., Kupermintz, H., & Lohman, D. F. (2001). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. New York, NY: Routledge.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671–684. doi:10.1037/h0043943

Cronbach, L. J. (1967). How can instruction be adapted to individual differences? In R. M. Gagne (Ed.), *Learning and individual differences* (pp. 23–39). Columbus, OH: Merrill.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods*. New York, NY: Irvington.

Dewey, J. (1916). *Democracy and education*. New York, NY: Macmillan.

Dewey, J. (1938). *Experience and education*. New York, NY: Macmillan.

DuBois, P. H. (Ed.). (1947). *The classification program* (Army Air Forces Psychology Program Research Reports, No. 2). Washington, DC: U.S. Government Printing Office.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York, NY: Random House.

Eden, D., & Shani, A. (1982). Pygmalion goes to boot camp: Expectancy, leadership, and trainee performance. *Journal of Applied Psychology, 67*, 194–199. doi:10.1037/0021-9010.67.2.194

Elliott, J. G. (2003). Dynamic assessment in educational settings: Realising potential. *Educational Review, 55*, 15–32. doi:10.1080/00131910303253

Erdley, C. A., & Dweck, C. S. (1993). Children's implicit personality theories as predictors of their social judgments. *Child Development, 64*, 863–878. doi:10.2307/1131223

Fabio, R. A. (2005). Dynamic assessment of intelligence is a better reply to adaptive behavior and cognitive plasticity. *Journal of General Psychology*, *132*, 41–66. doi:10.3200/GENP.132.1.41-66

Feuerstein, R., & Feuerstein, S. (1994). Mediated learning experience: A theoretical review. In R. Feuerstein, P. S. Klein, & A. J. Tannebaum (Eds.), *Mediated learning experience (MLE): Theoretical, psychosocial, and learning implications* (pp. 3–51). London, England: Freund.

Feuerstein, R., Klein, P. S., & Tannenbaum, A. J. (Eds.). (1991). *Mediated learning experience (MLE): Theoretical, psychosocial, and learning implications*. London, England: Freund.

Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device-theory, instruments, and techniques*. Baltimore, MD: University Park Press.

Gervey, B. M., Chiu, C., Hong, Y., & Dweck, C. S. (1999). Differential use of person information in decisions about guild versus innocence: The role of implicit theories. *Personality and Social Psychology Bulletin*, *25*, 17–27. doi:10.1177/0146167299025001002

Gregory, R. J. (2007). *Psychological testing: History, principles, and applications* (5th ed.). Boston, MA: Pearson.

Grigorenko, E. L. (2009). Dynamic assessment and response to intervention: Two sides of one coin? *Journal of Learning Disabilities*, *42*, 111–132. doi:10.1177/0022219408326207

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, *124*, 75–111. doi:10.1037/0033-2909.124.1.75

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Guthke, J. (1992). Learning tests: The concept, main research findings, problems, and trends. *Learning and Individual Differences, 4*, 137–151. doi:10.1016/1041-6080(92)90010-C

Guthke, J., Beckmann, J. F., & Dobat, H. (1997). Dynamic testing-problems, uses, trends, and evidence of validity. *Educational and Child Psychology, 14*, 17–32.

Hamers, J. H. M., Hessles, M. G. P., & Pennings, A. H. (1996). Learning potential in ethnic minority children. *European Journal of Psychological Assessment, 12*, 183–192. doi:10.1027/1015-5759.12.3.183

Haywood, H. C., & Lidz, C. (2007). *Dynamic assessment in practice*. New York, NY: Cambridge University Press.

Heslin, P. A. (2009). "Potential" in the eye of the beholder: The role of managers who spot rising stars. *Industrial and Organizational Psychology*, *2*, 420–424. doi:10.1111/j.1754-9434.2009.01166.x

Heslin, P. A., Latham, G. P., & VanderWalle, D. (2005). The effect of implicit person theory on performance appraisals. *Journal of Applied Psychology, 90*, 842–856. doi:10.1037/0021-9010.90.5.842

Hessles, M. G. P., & Hamers, J. H. M. (1993). A learning potential test for ethnic minorities. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical methodological, and practical issues* (pp. 285–311). Lisse, the Netherlands: Swets & Zeitlinger.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*, 253–270. doi:10.1037/h0023816

Jeltova, I., Birney, D., Fredine, N., Jarvin, L., Sternberg, R. J., & Grigorenko, E. L. (2007). Dynamic assessment as a process-oriented assessment in educational settings. *Advances in Speech Language Pathology*, *9*, 273–285. doi:10.1080/14417040701460390

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin and Review*, *9*, 637–671. doi:10.3758/BF03196323

Kaplan, R. M., & Saccuzzo, D. P. (2009). *Psychological testing: Principles, applications, and issues* (7th ed.). Belmont, CA: Wadsworth.

Kyllonen, P. C. (2005, September). The case for noncognitive assessments. *ETS R&D Connections,* 1–7.

Kyllonen, P. C., Roberts, R. D., & Stankov, L. (Eds.). (2008). *Extending intelligence: Enhancement and new constructs.* New York, NY: Erlbaum.

Lemann, N. (2000). *The big test: The secret history of the American meritocracy*. New York, NY: Farrar, Straus, & Giroux.

Lidz, C., & Elliott, J. G. (2000). *Dynamic assessment: Prevailing models and applications.* Greenwich, CT: Elsevier.

Matthew, C. T., Beckmann, J. F., & Sternberg, R. J. (2008). *Development of a test battery to assess mental flexibility based on Sternberg's theory of successful intelligence.* (Tech. Rep. No. 1222). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Matthew, C. T., & Stemler, S. E. (2008). *Exploring pattern recognition as a predictor of mental flexibility* (Final report submitted to the Army Research Institute for the Behavioral and Social Sciences for Project Award #W91WAW-07-C-00). Middletown, CT: Wesleyan University.

Mone, E. M., Acritani, K., & Eisinger, C. (2009). Take it to the roundtable. *Industrial and Organizational Psychology*, *2*, 425–429. doi:10.1111/j.1754-9434.2009.01167.x

Mueller-Hanson, R. A., Swartout, E., Hilton, R., & Nelson, J. (2009, June). *Proof of concept research for developing adaptive performance: Validation plan.* Arlington, VA: Personnel Decisions Research Institutes.

Olswang, L. B., & Bain, B. A. (1996). Assessment information for predicting upcoming change in language production. *Journal of Speech and Hearing Research, 39*, 414–423.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest, 9*, 105–119.

Plaks, J. E., Stroessner, S. J., Dweck, C. S., & Sherman, J. W. (2001). Person theories and attention allocation: Preferences for stereotypic versus counterstereotypic information. *Journal of Personality and Social Psychology, 80*, 876–893. doi:10.1037/0022-3514.80.6.876

Pulakos, E. D., Schmitt, N., Dorsey, D. W., Arad, S., Hedge, J. W., & Borman, W. C. (2002). Predicting adaptive performance: Further tests of a model of adaptability. *Human Performance*, *15*, 299–324. doi:10.1207/S15327043HUP1504_01

Rosenthal, R., & Jacobson, L. (1992). *Pygmalion in the classroom* (Expanded ed.). New York, NY: Irvington. (Original work published 1968)

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274. doi:10.1037/0033-2909.124.2.262

Schmitt, N., Oswald, F. L., & Gillespie, M. A. (2005). Implications of broadening the performance domain for the prediction of academic success. In W. F. Camara & E. Kimmel (Eds.), *Choosing students: Higher education admission tools for the 21st century* (pp. 195–214). Mahwah, NJ: Erlbaum.

Schmitt, N., Oswald, F. L., Kim, B. H., Imus, A., Drzakowski, S., Friede, A., & Shivpuri, S. (2007). The use of background and ability profiles to predict college student outcomes. *Journal of Applied Psychology, 92*, 165–179. doi:10.1037/0021-9010.92.1.165

Silzer, R., & Church, A. H. (2009a). The pearls and perils of identifying potential. *Industrial and Organizational Psychology*, *2*, 377–412. doi:10.1111/j.1754-9434.2009.01163.x

Silzer, R., & Church, A. H. (2009b). The potential for potential. *Industrial and Organizational Psychology*, *2*, 446–452. doi:10.1111/j.1754-9434.2009.01172.x

Snow, R. E. (1977). Individual differences and instructional theory. *Educational Researcher*, *6*(10), 11–15.

Snow, R. E. (1978). Theory and method for research on aptitude processes. *Intelligence*, *2*, 225–278. doi:10.1016/0160-2896(78)90019-3

Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist*, *27*, 5–32. doi:10.1207/s15326985ep2701_3

Spear, L. C., & Sternberg, R. J. (1987). Teaching styles: Staff development for teaching thinking. *Journal of Staff Development*, *8*(3), 35–39.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201–293. doi:10.2307/1412107

Stemler, S. E. (2009). *The measurement of adaptability in military personnel.* Orlando: Institute of Simulation and Training, University of Central Florida.

Stemler, S. E. (2012). What should university admissions tests predict? *Educational Psychologist, 47*, 5–17. doi: 10.1080/00461520.2011.611444

Stemler, S. E., Grigorenko, E. L., Jarvin, L., & Sternberg, R. J. (2006). Using the theory of successful intelligence as a basis for augmenting AP exams in psychology and statistics. *Contemporary Educational Psychology*, *31*, 75–108.

Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., Jarvin, L., & Sharpes, K. (2009). Using the theory of successful intelligence as a framework for developing assessments in AP physics. *Contemporary Educational Psychology, 34*, 195–209. doi:10.1016/j.cedpsych.2009.04.001

Sternberg, R. J. (1985). *Beyond IQ.* New York, NY: Cambridge University Press.

Sternberg, R. J. (1995). *In search of the human mind.* Orlando, FL: Harcourt Brace.

Sternberg, R. J. (1997). *Successful intelligence.* New York, NY: Plume.

Sternberg, R. J. (2003a). *Sternberg triarchic abilities test.* Unpublished test.

Sternberg, R. J. (2003b). *Wisdom, intelligence, and creativity synthesized.* New York, NY: Cambridge University Press. doi:10.1017/CBO9780511509612

Sternberg, R. J. (2005). The theory of successful intelligence. *Interamerican Journal of Psychology*, *39*, 189–202.

Sternberg, R. J. (2009, Fall). WICS: A new model for liberal education. *Liberal Education*, *95*(4), 20–25.

Sternberg, R. J., Bonney, C. R., Gabora, L., Karelitz, T., & Coffin, L. (2010). Broadening the spectrum of undergraduate admissions. *College and University*, *86*, 2–17.

Sternberg, R. J., Ferrari, M., Clinkenbeard, P., & Grigorenko, E. L. (1996). Identification, instruction, and assessment of gifted children: A construct validation of a triarchic model. *Gifted Child Quarterly*, *40*, 129–137. doi:10.1177/001698629604000303

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J., Snook, S., Williams, W. M., . . . Grigorenko, E. L. (2000). *Practical intelligence in everyday life.* New York, NY: Cambridge University Press.

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing.* New York, NY: Cambridge University Press.

Sternberg, R. J., Grigorenko, E. L., Ferrari, M., & Clinkenbeard, P. (1999). A triarchic analysis of an aptitude–treatment interaction. *European Journal of Psychological Assessment*, *15*, 3–13. doi:10.1027//1015-5759.15.1.3

Sternberg, R. J., Grigorenko, E. L., & Zhang, L.-F. (2008a). A reply to two stylish critiques. *Perspectives on Psychological Science*, *3*, 516–517. doi:10.1111/j.1745-6924.2008.00092.x

Sternberg, R. J., Grigorenko, E. L., & Zhang, L.-F. (2008b). Styles of learning and thinking matter in instruction and assessment. *Perspectives on Psychological Science*, *3*, 486–506. doi:10.1111/j.1745-6924.2008.00095.x

Sternberg, R. J., & Lubart, T. I. (1995). *Defying the crowd: Cultivating creativity in a culture of conformity*. New York, NY: Free Press.

Sternberg, R. J., & the Rainbow Project Collaborators. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence, 34*, 321–350. doi:10.1016/j.intell.2006.01.002

Sternberg, R. J., & Williams, W. M. (1996). *How to develop student creativity*. Alexandria, VA: Association for Supervision and Curriculum Development.

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago, IL: University of Chicago Press.

Today's Military. (2012). *ASVAB test*. Retrieved from http://www.todaysmilitary.com/before-serving-in-the-military/asvab-test

U.S. Department of Education. (2010). *Statement on National Governors Association and State Education Chiefs Common Core Standards.* Retrieved from http://www.ed.gov/news/press-releases/statement-national-governors-association-and-state-education-chiefs-common-core

Vernon, P. E. (1950). *The structure of human abilities.* London, England: Methuen.

Vygotsky, L. W. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press. (Original work published 1934)

Weinstein, R. S. (2002). *Reaching higher: The power of expectations in schooling.* Cambridge, MA: Harvard University Press.

Yost, P. R., & Chang, G. (2009). Everyone is equal, but some are more equal than others. *Industrial and Organizational Psychology*, *2*, 442–445. doi:10.1111/j.1754-9434.2009.01171.x

CHAPTER 14

COLLEGE, GRADUATE, AND PROFESSIONAL SCHOOL ADMISSIONS TESTING

Wayne Camara, Sheryl Packman, and Andrew Wiley

In the United States, admissions into colleges, universities, and professional programs has long been the subject of much discussion and controversy (Bowen & Bok, 1998; Fisher & Resnick, 1990; Shelton, 1997; Zwick, 2002). The value of a degree from both a financial and individual standpoint has been well established. Baum and Ma (2007) demonstrated that a full-time employee without a college degree would earn, on average, approximately $31,500. This amount is considerably lower than those with a college degree who earn, on average, approximately $50,900 each year. College degree recipients also engage in more prosocial behaviors, such as volunteering, voting, and participating in political activities (Bowen & Bok, 1998; Goldberg & Smith, 2008).

The numbers are even more dramatic when recipients of advanced degrees are investigated. Lacey and Crosby (2004) estimated that obtaining a master's degree increased earnings for employees by approximately 21% compared with employees completing similar work who had obtained only a bachelor's degree. Another report estimated that the average annual earnings for full-time employees with a bachelor's degree was approximately $45,400 compared with full-time professional degree holders (MD, JD, DDS, or DVM), who had an average annual salary of $99,300 (U.S. Census Bureau, 2002). The same report estimated that employees with a bachelor's degree who worked full time throughout adulthood would earn approximately $2.1 million dollars. In contrast, employees holding a master's degree were estimated to earn $2.5 million, whereas those with a doctorate ($3.4 million) and professional degrees ($4.4 million) were estimated to earn even more.

Given these benefits, it is not surprising that the competition to enter into higher education institutions and professional programs can be extremely high. Colleges and universities have observed an increase in applications. For the fourth consecutive year, more than 75% of 4-year institutions have had an increase in applicants according to the National Association for College Admission Counseling (NACAC; Clinedinst & Hawkins, 2009). In addition, students in general are sending applications to an increasing number of schools; 22% of applicants for the fall 2008 entering class applied to seven or more colleges compared with 19% in 2007. This increase has also been observed in graduate schools, with the average number of law school applications per student increasing from 5.0 in 2002 to 6.5 in 2008 (Handwerk, 2009). In 2009, medical school applicants completed an average of 13 applications, resulting in 562,694 applications being sent with only 18,390 eventual matriculants (Association of American Medical Colleges, 2010).

In the competition for openings within undergraduate institutions, students are typically required to submit scores from either the SAT or the American College Test (ACT). Although there has been an increase in the number of colleges that do not require admissions test scores from applicants, the vast majority of competitive 4-year institutions still require a test score and test optional institutions have reported that approximately 65% to 85% of applicants submit test scores (Camara, 2009). Unlike undergraduate

DOI: 10.1037/14049-014
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

admissions, virtually all accredited full-time graduate and professional degree programs require admissions tests for consideration. Within the graduate school arena, the LSAT is used for law school admissions, the MCAT for medical school admissions, the GMAT for business school admissions, and the GRE for almost all other graduate school program admissions. Recently, more than 500 business schools have started accepting GMAT or GRE scores for admission (Burnsed, 2010). This chapter provides a brief overview of each of these examinations, along with key test features (see Tables 14.1 and 14.2 for a summary), and the research that has been produced to support its use in admissions.

UNDERGRADUATE ADMISSIONS TESTS

Many researchers have noted the similarities between the ACT and the SAT, the two standardized, undergraduate admissions testing programs (e.g., Beatty, Greenwood, & Linn, 1999; Camara, 2009). Despite their similarities, each test was designed with somewhat different purposes and retains some important distinctions in content and structure. The SAT was originally developed for admissions decisions at competitive institutions and measured general verbal and mathematical reasoning to provide "a standard way of measuring a student's ability to do college-level work" (as quoted in Wightman & Jaeger, 1998, pp. 5–6). In contrast, the ACT was designed for Midwestern institutions that generally admitted all qualified applicants. The ACT was intended not only to assist these colleges in admissions and recruitment but also with course placement and academic planning. It had the additional purpose of helping students to "identify and develop realistic plans for accomplishing their educational and career goals" (as quoted in Wightman & Jaeger, 1998, p. 3). According to Beatty et al. (1999),

> Although the distinction between the coastal and Midwestern institutions that accounted for these differences has faded,

TABLE 14.1

Characteristics of Undergraduate Admissions Tests

Characteristics	ACT	SAT	SAT subject tests
First administration	1959	1926	1901
Volumes	1,480,469 examinees from 2009 cohort	1,530,128 examinees from 2009 cohort	294,893 examinees from 2009 cohort
Test sections (no. of items)	English (75), math (60), reading (40), science (40), writing (optional; 1 essay)	critical reading (67), math (54), writing (49 + 1 essay)	20 tests in English (60), foreign languages (70–85), history (90–95), math (50) & science (75–85)
Item types	SR + 1 essay	SR, grid-in + 1 essay	SR
Delivery	paper based	paper based	paper based
Format	linear	linear	linear
SR scoring	computer; rights-only scoring	computer; formula scoring	computer; formula scoring
Essay scoring	human	human	n/a
Score scale	1–36 (composite) 1–18 (each subscore)	200–800 (each section)	200–800
Essay/score scale	2–12	2–12	n/a
Total test time	175 min + 30 min for writing	225 min	60 min
No. of administrations	6 times per year	7 times per year	6 times per year; maximum of 3 tests per administration
Cost	$33; $48 with writing	$47	$21 (registration) + $10 (per test); $21 for tests with listening

Note. ACT = American College Test; SR = selected response.

TABLE 14.2

Characteristics of Graduate Admissions Tests

Characteristics	Revised GRE general test	GRE subject tests	GMAT	LSAT	MCAT
First administration	1949	varied	1954	1948	1928
Volumes	1,489,162 (verbal) & 1,489,044 (quant) examinees from July 1, 2006 to June 30, 2009	Ranged from 3,837 to 22,683 examinees by test from July 1, 2006 to June 30, 2009	265,613 examinees in 2008–2009	171,514 tests administered in 2009–2010	79,244 tests administered in 2009
Test sections (no. of items)	analytical writing (2 essays), verbal reasoning (~40), quantitative reasoning (~40),	8 tests in science (100–200), computer science (70), English (230), math (66) & psychology (205)	analytical writing (2 essays), quantitative (37), verbal (41)	analytical reasoning (23–24), logical reasoning (48–52), reading comprehension (26–28), variable (24–28) writing (unscored; 1 essay)	biological sciences (52), physical sciences (52), verbal reasoning (40), writing (2 essays)
Item types	SR, numeric entry, + 2 essays	SR	SR + 2 essays	SR + 1 essay	SR + 2 essays
Delivery	computer based; paper-based option	paper based	computer based	paper based	computer based
Format	multistage adaptive (SR, numeric entry only)	linear	adaptive (SR only)	linear	Linear
SR scoring	multistage adaptive; paper based: rights-only scoring	computer; formula scoring	computer adaptive	computer; rights-only scoring	computer; rights-only scoring
Essay scoring	human + AI	n/a	human	n/a	human + AI
Score scale	130–170	200–990	200–800	120–180	1–45
Essay/score scale	0–6	n/a	0–12	unscored	J (2) – T (12)[a]
Total test time	190 min (+ any variable sections); paper based = 210 min	170 min	210 min	210 min	260 min
No. of administrations	year-round administration; maximum of 1 administration per calendar month and 5 times per year	3 times per year	year-round administration; maximum of 1 administration per calendar month and 5 times per year	4 times per year	> 25 times per year
Cost	$190	$140	$250	$136	$230

Note. GRE = Graduate Record Examination; GMAT = Graduate Management Admission Test; LSAT = Law School Admission Test; MCAT = Medical College Admission Test; SR = selected response; AI = artificial intelligence.
[a]J and T are the lowest and highest MCAT essay letter grades, respectively.

> the SAT and the ACT have retained their distinct goals (despite the fact that in many institutions the two tests are used almost interchangeably). (p. 5)

The ACT

The ACT (2007, 2010) test, created and maintained by ACT, Inc., is a test of high school educational achievement and college readiness taken by

college-bound high school students. Almost 3,000 colleges and universities use the ACT for admissions and placement decisions. In addition, some states include the ACT in their statewide assessment programs for accountability and high school graduation. The test is also used by various organizations and agencies to award financial assistance and scholarships for postsecondary education.

The first ACT administration was in the fall of 1959. In the high school class of 2009, 1,480,469 graduating seniors took the ACT during their high school careers compared with 1,171,460 in 2004. There are four required tests in English, mathematics, reading, and science, and one optional writing test. Each of the required tests is composed of four-option multiple-choice questions and the writing test includes a single essay. Currently, the ACT is a paper-based test administered six times a year. Administration is timed and takes 2 hours and 55 min without the optional writing test. Taking the writing test increases testing time by 30 min. Skills that are acquired in high school and important for postsecondary success are assessed, including reasoning, problem solving, analysis, evaluation, interpretation, integration, and application.

The English test includes five prose passages accompanied by 75 selected-response items. Of these items, 40 assess conventions of the English language (usage/mechanics) and 35 assess rhetorical skills. Usage/mechanics includes punctuation (13% of items), grammar and usage (16%), and sentence structure (24%). The rhetorical skills section is composed of strategy (16%), organization (15%), and style (16%). Test takers receive a total score on the English test as well as subscores on usage/mechanics and rhetorical skills. This test is administered in 45 min.

The mathematics test consists of 60 selected-response items with 24 items on pre-algebra (23% of items)/elementary algebra (17%), 18 items on intermediate algebra (15%)/coordinate geometry (15%), and 18 items on plane geometry (23%)/trigonometry (7%). A total score is reported along with subscores on the three content sections. Content is integrated with skill to assess the ability to use knowledge, facts, and formulas to solve problems in mathematical and real-world situations as well as knowledge of and the ability to integrate major concepts. Certain calculators are permitted and 60 min is allotted for this test.

The reading test is composed of four passages in social studies (25% of items), natural sciences (25%), prose fiction (25%), and humanities (25%). The accompanying 40 selected-response items assess reading comprehension through the skills of referring to explicit content and reasoning to determine implicit content. In addition to a total score on all items, two subscores based on the 20 items each that assess social studies/sciences and art/literature reading skills are reported. Test takers are given 35 min to complete this test.

The science test includes seven sets of scientific information in three formats, data representation (38% of items), research summaries (45%), and conflicting viewpoints (17%) along with 40 selected-response items. Content in biology, chemistry, physics, and Earth/space sciences are assessed through interpretation, analysis, evaluation, reasoning, and problem solving. Test takers are assumed to have completed 1 year of biology and 1 year of a physical or Earth science course. Only one total score is reported for this test. This test is administered in 35 min, and calculators are not permitted.

The optional writing test was added to the ACT in 2005 and is composed of a 30-min essay. One prompt that presents two points of view on an issue is provided and the test taker needs to take a position on that issue and respond to a related question. The writing test essay is scored holistically by two human readers on a scale of 1 to 6 for a sum total score ranging from 2 to 12. If the readers disagree by more than one point, a third reader resolves the discrepancy. A subscore on this test is reported, which reflects a student's performance on the essay, along with a combined score on the writing test and the English test. In 2009, 45% of college-bound seniors took the ACT without the optional essay (ACT, 2009).

The ACT test is scored with rights-only scoring, which means that 1 point is awarded for each correct answer and there is no deduction for incorrect responses. One composite score and the four total scores on each test are reported on a scale from 1 to 36. All subscores (e.g., rhetorical skills, plane geometry/trigonometry) are reported on a scale of 1 to 18. Finally, students preparing for the ACT

often take PLAN in the 10th grade and EXPLORE in the eighth grade.

In the last decade, several states administered the ACT to all students. In states such as Illinois and Michigan, the ACT is used as part of the state and federal accountability system under No Child Left Behind (NCLB), whereas in other states like Colorado and Kentucky it is administered to all students as a measure of college readiness to aid schools and students in gaining greater understanding of students' preparedness for postsecondary education. In 2005, ACT released a report that established cut scores that predict college readiness. The cut scores were set at the point at which students have a 50% probability of attaining a B or higher and a 75% probability of attaining a C or higher on freshmen courses in each subject (Allen & Sconing, 2005). This report, and each year's annual release, identifies the number and percent of college-bound seniors who are considered to be college ready. In 2009, 23% of students were considered college ready across all benchmarks (ACT, 2009).

The SAT

The SAT (College Board, 2010), which launched in 1926, is owned and managed by the College Board. The SAT is a standardized college admissions test that determines college readiness and is also used for awarding scholarships and financial aid based on academic potential. Almost every postsecondary institution in the United States uses SAT scores to make admissions decisions.

Three required tests are administered for the SAT: critical reading, mathematics, and writing. The three tests are administered in 10 separately timed sections, three each for writing, critical reading, and mathematics, and one unscored variable section used for pretesting new test items or equating test forms. The order of administration of the test sections varies, except that the essay, which is one of three sections for the writing test, is always the first section administered. Total testing time for the SAT is 3 hours and 45 min.

The critical reading test consists of 48 passage-based and 19 sentence-completion items. Content wise, the items are divided up into 42 to 50 extended-reasoning items, 4 to 6 literal comprehension, and 12 to 16 vocabulary-in-context selected-response items. The skills assessed on these questions include determining the meaning of words, reading comprehension, analyzing information, making inference, and evaluation.

The mathematics text includes 44 selected-response and 10 student-produced items. The latter item types were first introduced with the 1994 revision of the SAT and also are referred to as "grid-in" items because students must determine the correct response and grid-in the numerals and math symbols (e.g., fraction sign, decimal point, negative sign) using a standard bubble answer sheet. From a content perspective, the mathematics test includes 11 to 13 items on numbers and operations, 19 to 21 on algebra and functions, 14 to 16 on geometry and measurement, and six to seven on data analysis, statistics, and probability. These items assess knowledge and application of mathematical concepts, data interpretation, and problem solving.

The writing test has one section composed of a 25-min essay and two additional sections composed of selected-response items. During the essay, test takers respond to a provided prompt that contains a quote or statement on a general issue. The other two writing sections include 25 improving sentences items, 18 identifying sentence errors items, and 6 improving paragraph items. The skills assessed through the writing section include the ability to develop and support a point of view, editing and revising, organization, and knowledge of correct grammar usage, sentence structure, and effective sentences.

The SAT is paper based and formula scored, which means that test takers receive one point for every correct answer and lose one quarter of one point for every incorrect answer. SAT scores on all three tests range from 200 to 800, in 10-point increments. For the writing test, each essay is graded on a scale from 1 to 6 by two human readers. A third reader grades the essay if the two readers differ by more than one point. These scores are aggregated to produce an essay subscore that ranges from 2 to 12. The essay subscore counts for approximately 30% and the selected-response subscore counts for 70% of the writing composite score.

Of the students in the graduating class of 2009, 1,530,128 took the SAT (College Board, 2009b) at some point in their high school career. As part of their preparation, students can take the PSAT/NMSQT during 10th or 11th grade to prepare for and predict their scores on the SAT. In 2009, 1,545,856 students in the 11th grade and 1,517,231 students in the 10th grade elected to take the PSAT/NMSQT (College Board, 2009a). The College Board recently introduced the ReadiStep examination, designed for eighth-grade students, with the purpose of presenting an early snapshot of student progression and development toward college readiness.

Wiley, Wyatt, and Camara (2010) developed a college readiness index designed to estimate the percentage of SAT students considered to be college ready. Student SAT scores, along with high school grade point average (GPA), and an index of academic rigor derived from the SAT Questionnaire were combined to develop a single estimate of student college readiness. Wiley et al. estimated that in 2009, 32% of SAT test takers should be considered college ready.

The SAT Subject Tests

The SAT Subject Tests are a set of college admissions tests produced by the College Board. The purpose of these tests is for college-bound students to demonstrate acquisition of subject-specific knowledge and skills. There are 20 Subject Tests that cover English literature, U.S. history, world history, mathematics Level 1 and Level 2, biology, chemistry, physics, Chinese, French, French with listening, German, German with listening, modern Hebrew, Italian, Japanese with listening, Korean with listening, Latin, Spanish, and Spanish with listening. Scores on the Subject Tests are particularly useful for students seeking admission into a particular program of study or school within a college or university who want to distinguish their ability from other applicants. Test scores are also used for placement into college courses. Scores are currently required or recommended by approximately 160 higher education institutions.

The Subject Tests are paper based and contain only selected-response items. Tests are formula scored, whereby one point is awarded for each correct response and one quarter, one third, and one half of a point is deducted for each incorrect answer that has five, four, and three response options, respectively. Each test is administered in 1 hour. During a single administration, students can take between one and three tests. Each year, the tests are administered six times in the United States and internationally. In 2009, there were 294,893 graduating seniors who took at least one SAT Subject Test during their high school career.

ADMISSIONS TESTS FOR GRADUATE SCHOOLS

Although there is a fair amount of similarity among the admissions tests used at the undergraduate level, there is notably more variety to be found when reviewing admissions exams used at the graduate level. These examinations are designed to assist in admissions decisions to institutions as wide-ranging as law schools, medical schools, and psychology programs.

The Law School Admission Test

The Law School Admission Test (LSAT; Law School Admissions Council, 2010), first administered in 1948, is maintained by the Law School Admissions Council (LSAC). All applicants to law schools approved by the American Bar Association are required to take this test. The LSAT is designed to help law school admissions officers make admissions decisions based on the reasoning skills of their applicants.

The LSAT is a paper-based test composed of six 35-min sections that include one section on reading comprehension, one section on analytical reasoning, two sections on logical reasoning, one unscored writing sample, and one variable section used for equating and pretesting purposes. All items, aside from the writing sample, are selected response. Four passages in the reading comprehension section are followed by five to eight associated items that measure the ability to determine the author's main idea, draw inferences, find information, or describe the structure. The analytical section is composed of four different logic games that involve organization of elements based on a set of statements. The logical

reasoning section provides an argument or set of facts and requires the test taker to identify assumptions, conclusions, errors in logic, similar lines of reasoning, or statements that would weaken or strengthen the argument. Finally, the writing sample presents a problem and two solutions, and test takers must feature one solution in a carefully constructed essay.

The LSAT is rights-only scored and scores are reported on a scale from 120 to 180. The last section of each test administration is the writing sample, which is scanned and sent directly to each law school a test taker applies. The variable section is always one of the first three sections administered. Total testing time is 3.5 hours. In 2009–2010, 171,514 tests were administered compared with 145,258 in 2004–2005.

The Medical College Admission Test

The Medical College Admission Test (MCAT; Association of American Medical Colleges, 2010), first administered in 1928, is managed by the Association of American Medical Colleges and is designed to assist medical schools with their admissions decisions. The MCAT is a required component of admissions applications for almost all medical schools in the United States.

The MCAT measures problem-solving, critical-thinking, and writing skills as well as knowledge of the scientific concepts necessary for the successful study of medicine. Four sections on the test are administered in the following order: physical sciences (52 items in 70 min), verbal reasoning (40 items in 60 min), writing (two essays in 60 min), and biological sciences (52 items in 70 min). The writing sample is the only section that is not selected response, whereby test takers type two short essays from a given topic statement that demonstrates their ability to develop a central idea and to present a clear argument that supports their central idea. Total testing time is 4 hours and 20 min.

The MCAT is scored with rights-only scoring. The selected-response sections are scored on a 1- to 15-point scale. Each essay is scored on a 1- to 6-point scale by one human reader and one automated scoring system. The four scores received on both essays are combined and converted into a letter grade from J (equal to a numerical 2) to T (equal to a numerical 12). All scores are aggregated into one composite score reported with the essay score for a maximum score of 45T. In January of 2007, the MCAT changed from a paper-based delivery system to a computer-based model. The test is administered at least 25 times per year at testing centers located across the United States and internationally.

The Graduate Record Examination General Test

The Graduate Record Examination (GRE; Educational Testing Service [ETS], 2010a) General Test is a broad graduate school admissions test produced by ETS. The first administration was in 1949. Currently, more than 3,200 graduate programs and business schools use the GRE to evaluate readiness for graduate-level work and to award scholarships and fellowships. The GRE General Test is administered to more than 600,000 potential applicants each year at computer-based testing centers in the United States and abroad.

ETS revised the GRE General Test to better reflect the content and higher level cognitive skills required to succeed in 21st-century graduate and business programs. The revised test launched in August 2011. Similar to the previous version, the revised version is composed of three sections: analytical writing, verbal reasoning, and quantitative reasoning.

The analytical writing section, which is always administered first, includes one 30-min "analyze an issue" task in which the test taker must justify their position on a provided critical issue. There is also a 30-min "analyze an argument" task in which the test taker must evaluate a logical argument. Each essay within the analytical writing section is scored holistically by at least one trained reader using a 6-point scale and that score is checked by e-rater, an artificial intelligence scoring program developed by ETS. If the human rater and e-rater scores are discrepant, a second human reader rates the essay and the final score is calculated by averaging the two human reader scores, and in the rare cases in which these scores differ by more than a point, an additional human score is obtained and a final adjudicated score is produced. Scores are reported on a 0 to 6 scale in half-point increments.

The verbal reasoning section includes text completion, sentence equivalence, and reading comprehension items presented in two sections with approximately 20 items per section. Thirty minutes are allotted for each section. Item types in this section include multiple-choice items and select-in-passage items in which a sentence within a passage is selected as the response. Antonyms and analogies were removed from this section of the revised test. The quantitative reasoning section assesses arithmetic, algebra, geometry, and data analysis through approximately 40 selected-response, numeric-entry, and quantitative-comparison items administered in two sections. Thirty-five minutes are allotted for each section.

The prior version of the GRE was fully adaptive at the individual item level; the current GRE is a two-stage adaptive computer-based test where answers to the first verbal and quantitative reasoning sections determine the items administered in the second section for each content area, respectively. Item selection for the second section is based on the statistical characteristics (e.g., difficulty) of the preceding items answered correctly, and the required variety of item types and content coverage. Scoring takes into account the number of items viewed, the number of correct answers provided, and the statistical properties of the items taken. Additional changes to the test include the ability to edit responses and skip and return to questions within each reasoning section as well as the use of an on-screen calculator for the quantitative reasoning section. Based on all of these changes, a new score scale was also created for the reasoning sections, resulting in a 130- to 170-score scale, reported in one-point increments.

A paper-based version is administered in areas without access to computer-based testing. This version is a linear test that also incorporates the new items types on the revised computer-based test. Test takers mark their answers in a test book rather than a traditional answer sheet and a calculator is provided whenever access is also provided for computer-based test takers. The analytical writing section is identical in structure and format to the computer-based test. The scoring, however, does not utilize e-rater. Instead, two trained readers score each essay and their scores are averaged, and adjudicated with an additional reader if discrepant by more than one point. The verbal reasoning and quantitative reasoning sections on the paper-based test each contain 50 items equally split into two sections to be completed in 35 and 40 min per section, respectively. These two sections are scored using rights only scoring.

In addition to the GRE General Tests, there are also eight GRE Subject Tests, which cover college-level biochemistry, cell and molecular biology, biology, chemistry, computer science, literature in English, mathematics, physics, and psychology. Students take these tests to show additional competency in a specific subject area and readiness for a specialized graduate program or school. Scores on these tests are one tool that admissions officers can use in the selection and placement process.

Each test is composed of between 66 and 230 selected-response items. The GRE Subject Tests are formula scored such that wrong answers are penalized more than omits (score = number right minus one quarter of number wrong). Scores range from 200 to 990, reported in 10-point increments. The tests are administered three times per year at paper-based test centers.

The Graduate Management Admission Test

The Graduate Management Admission Test (GMAT; Graduate Management Admission Council, 2010), which was first administered in 1954, is maintained by the Graduate Management Admission Council. The purpose of this test is to assist business schools in selecting applicants for admission, provide graduate school counseling, and award financial aid based on academic potential. The GMAT is used by more than 1,900 schools worldwide.

The GMAT consists of an analytical writing assessment (AWA), a quantitative section, and a verbal section. The AWA includes two 30-min writing tasks, the first of which requires test takers to analyze a given issue and explain and support their opinion on the subject. The second task for the AWA requires that an evaluation and critique of an argument presented about a topic. There are 37 selected-response questions on the quantitative section, which are completed in 75 min. Each item tests problem-solving or data-sufficiency skills within the content areas of arithmetic, elementary algebra, or geometry. The 75-min

verbal section includes 41 selected-response items that cover reading comprehension, critical reasoning, and sentence correction. Total testing time is 3.5 hours.

Like the GRE General Test, the GMAT is delivered on the computer, with the quantitative and verbal sections of the GMAT being computer-adaptive tests. GMAT scores are reported on a scale from 200 to 800 in 10-point increments. Two human readers independently score both components of the AWA. Scores are then averaged and can range from 0 to 6 in half-point intervals. The verbal and quantitative section scores range from 0 to 60. In the 2008–2009 testing year, the test was administered to 265,613 potential business school applicants in more than 90 countries compared with 203,181 in the 2003–2004 testing year (GMAC, 2009).

Benefits of Admissions Tests

All of the admission tests discussed thus far have been, and continue to be, the focus of attention and questions about their use in admissions decisions. Because of these persistent questions, a committee of the National Research Council was charged with examining the evidence centered on the use of these tests. The committee identified a number of key benefits accrued with the use of the tests, such as the following:

- Standardization—curricular, grading standards, and course content vary enormously across schools and admissions; tests offer an efficient source of comparative information for which there is no substitute.
- Efficiency—admissions tests are provided at relatively low cost to students and are efficient for institutions comparing hundreds or thousands of applicants in a very short period of time.
- Opportunity—standardized tests provide an opportunity to demonstrate talent for students whose academic records are not particularly strong, who have not attended the most prestigious prior institutions, or who have not taken the most rigorous courses (Beatty et al., 1999).

RESEARCH TOPICS: VALIDITY

As would be expected with tests whose primary focus is on admissions, the majority of validity evidence for these tests is dedicated to demonstrating their efficacy in the context of admissions decisions (see Volume 1, Chapter 4, this handbook). The earliest conception of validity focused on prediction, and for several decades, validity centered on the basis of predictive accuracy (Brennan, 2006). In fact, the first validity study on admissions tests was actually conducted as students completed the first SAT in 1926 and earned college grades (Mattern, Kobrin, Patterson, Shaw, & Camara, 2009). This section identifies key themes or patterns observed across various validity studies conducted on the major admissions testing programs.

Accuracy of Prediction

Admissions officers, whether working at an undergraduate institution, a law school, medical school, or graduate program, are all faced with the difficult decision of who to accept and who to reject. To aid in this decision-making process, admissions officers look to a wide variety of information, including high school or college grades, the rigor of curricula, test scores, letters of recommendation, and extracurricular activities. Nonetheless, the two measures that carry the most weight are student GPA and test scores. A recent report from the National Association for College Admission Counseling (Clinedinst & Hawkins, 2009) estimated that approximately 93% of postsecondary schools place moderate or considerable importance on high school GPA, whereas approximately 88% did the same with test scores. Given the importance and value of these two measures, it is not surprising that these two measures have received the most focus in research focused on school admissions.

Virtually all of the admissions testing programs provide extensive evidence demonstrating their predictive accuracy in identifying students likely to succeed in college (Burton & Ramist, 2001; Julian, 2005; Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008; Kuncel, Crede, & Thomas, 2007; Kuncel & Hezlett, 2007; Kuncel, Hezlett, & Ones, 2001; Noble & Sawyer, 2002; Stilwell, Dalessandro, & Reese, 2007). All of these studies looked at not only the admissions tests' ability to predict successful performance but also their relationship to grades in college, graduate programs, or professional

programs. Rather than reviewing all of the studies that have been produced demonstrating the predictive validity of the admissions tests, we will identify and review key research studies or reviews that provide a more comprehensive approach to validity evidence of these examinations.

A meta-analysis of SAT validity was conducted in relation to college grades after one semester, and each year of college, including cumulative grades (Hezlett et al., 2001). Results for 1st-year college grades were based on more than 1,734 studies, with aggregate sample sizes ranging from 146,000 to more than 1 million. The average, sample-weighted, observed validity coefficients for 1st-year college GPA (FYGPA) ranged from .30 to .36. The operational validities of the SAT verbal (SAT-V), SAT math (SAT-M), and SAT-T (SAT-V + SAT-M) in predicting GPA for first semester and first year of college ranged from .44 to .62. None of the 90% credibility intervals around the operational validities included zero, and the standard deviations of the operational validities (SD_ρ) ranged from .06 to .20. These small to moderate values suggest that either predictive validity values were not affected by moderator variables or that the effects of any moderators were relatively small. Collectively, these results demonstrated that SAT scores are valid predictors of performance early in college. The operational validities of the SAT-V and SAT-M for predicting noncumulative GPA in the second, third, and fourth years of college also ranged from the mid-30s to the mid-40s. Results for 2-year and 4-year cumulative college grades were similarly robust, with aggregated sample sizes of at least 10,000, with observed validities ranging from .29 to .37 and operational validities from .40 to .50. By reporting results from published and unpublished studies, the authors completed the largest meta-analysis published in the social sciences and found that the SAT predicted a wide range of academic performance, study habits, and withdrawal in academic settings.

Burton and Ramist (2001) conducted an extensive review of studies evaluating the ability of SAT scores and high GPA to predict successful performance in college. They examined the relationship between these predictors and a variety of measures of successful college performance, including FYGPA, cumulative GPA, college graduation, academic honors, and other nonacademic indicators of successful performance. This review found that both SAT scores and high school GPA made significant contributions to the prediction of FYGPA, cumulative GPA, and eventual college graduation. In all cases, the combination of the two variables provided notably more accurate predictions than using either one alone. Both predictors also seemed to show strong evidence for their ability to predict other academic behaviors, such as awards of academic distinction and departmental honors. The predictors also demonstrated notably lower but still significant relationships between most of the nonacademic variables that were identified and occurred within the college setting. Examples of these types of behavior included taking leadership positions within the school, active involvement in school or community activities, and artistic endeavors.

A common concern cited with validity studies is that different grades have different meaning at different colleges, making it difficult to compare FYGPA across institutions and professors. Berry and Sackett (2008) proposed a solution by examining the validity of admissions tests at the individual course level within an institution. Overall, they found a correlation of .58 between SAT scores and course grade composites compared with a correlation of .47 using FYGPA as the criterion. The correlation of HSGPA and course grade composite was .58 compared with .51 for FYGPA. They concluded that the predictive validity of the SAT was reduced by 19% because of the noise that is added as a result of taking different courses across different institutions. In fact, several studies have demonstrated that the validity of admissions test scores often increases when the criterion is course grades rather than FYGPA—this is particularly true in science and math courses (Camara, 2009).

Kuncel and Hezlett (2007) recently released a synthesis of meta-analyses that investigated the ability of admissions test scores to predict performance in graduate and professional programs. They identified four key results or findings:

1. Standardized tests are effective predictors of performance in graduate school.
2. Both tests and undergraduate grades predict important academic outcomes beyond grades earned in graduate school.

3. Standardized admissions test predict most measures of successful performance better than college academic records do.
4. The combination of tests and grades yields the most accurate predictions of success. (p. 1080)

The results identified by Kuncel and Hezlett (2007), as well as by Burton and Ramist (2001), have been consistently confirmed across a wide variety of studies covering a wide range of admissions tests. Research consistently demonstrates similar findings and their ability to predict successful performance in subsequent educational environments (Julian, 2005; Kobrin et al., 2008; Kuncel, Crede, & Thomas, 2007; Kuncel & Hezlett, 2007; Kuncel et al., 2001; Noble & Sawyer, 2002; Stilwell et al., 2007). Researchers have found that GPA, at the college or high school level, predicts performance moderately well, as does the admission test score for students. Normally, the GPAs for students have a slightly greater value in predicting performance, although in the graduate and professional programs, numerous studies have demonstrated superior predictive validity for the admissions tests. The studies have shown some variability depending on the criterion used to define successful performance.

More important, research has consistently demonstrated that although both successfully predict performance, the combination of the two always consistently outperforms either one variable alone. This relationship is observed with the SAT, where a recent report (Kobrin et al., 2008) demonstrated an incremental increase in predictive validity of .08 when SAT scores were added to the prediction equation using high school GPA alone as well as the ACT (Noble, 2003). This relationship also holds with the MCAT examination (Julian, 2005), the LSAT (Stilwell et al., 2007), the GRE (Burton and Wang, 2005; Kuncel et al., 2001), and the GMAT (Kuncel, Crede, & Thomas, 2007). Interestingly, the LSAT has shown consistently slightly higher utility in predicting law school performance than college GPA.

TYPES OF CRITERION MEASURES

As was just described, the evidence to support the predictive validity of the various admissions tests covers a wide variety of topic areas and students. An interesting component of the validity work being conducted on these examinations is the use of criterion variables and how these variables can affect the results of these studies. Historically, almost all validity studies have focused on predicting students GPA in their respective schools, and more often than not, the FYGPA for students.

The traditional FYGPA variable used in most of these studies does have a variety of important benefits. Probably the greatest benefit is that it allows for the most uniform comparison of students that is available during college or graduate school careers. Students generally complete a similar set of courses during their first year. After a student's first year, as they begin to pursue the courses for their major or discipline, the differences in course-taking patterns become more pronounced and would have a more notable affect on the consistency of GPA as a measure of student performance.

Cumulative grades, which represent the entire academic performance of a student at college, seem to instinctively be the best criterion measure for admissions tests. Camara and Echternacht (2000) noted that there are a number of problems in relying on cumulative grades to evaluate the utility of admissions measures. First, there are significant differences in courses taken and course difficulty across majors. Second, Willingham (1985) noted that there is far less variance in grades across upper level courses (fewer students are getting Cs). Third, there are little to no differences between large validity studies that use first-year grades as the criterion, and those studies using 2nd-, 3rd-, and 4th-year and cumulative grades (Hezlett et al., 2001). Within undergraduate institutions, FYGPA has been shown to be a very strong predictor of eventual success in college (Allen, 1999; Murtaugh, Burns, & Schuster, 1999). The strong link between FYGPA and eventual college success provides strong support for its use as a criterion variable.

Whereas FYGPA has shown itself to be a useful criterion, the predictive validity of these tests has also been evaluated across a wide variety of other factors. For all of these tests, the results have demonstrated an impressive consistency in their ability to predict performance, and also have revealed some

variability in this relationship, depending on the criterion variable investigated. At the undergraduate level, studies are increasingly evaluating other variables, such as retention to a second year at the school and even subsequent degree attainment. At the graduate level, studies are increasingly looking at such variables as eventual degree attainment, success at passing licensure or certification exams, and graduation with distinction or honors.

The studies have shown fairly strong relationships between their respective admissions test scores and the various measures of successful student performance. Mattern and Patterson (2011) collected data from 66 higher education institutions and analyzed the predictive validity of the SAT for predicting second-year GPA. Their study demonstrated that the predictive validity of the three combined SAT section had a predictive validity (.55) that was almost equivalent to that of high school GPA (.56). As with FYGPA, the most predictive variables were a combination of high school GPA and the SAT. Similar results were reported when investigating 3rd-year GPA as well (Mattern & Patterson, 2011).

Robbins, Allen, Casillas, Peterson, and Le (2006) demonstrated that the ACT is a strong predictor of retention to a second year of college. Allen, Robbins, Casillas, and Oh (2008) found similar results when looking at third-year retention rates for undergraduate students. Mattern and Patterson (2009) also reported that students with the highest SAT scores (composite scores of between 2,100 and 2,400) have approximately a 95% likelihood of returning for their second year, as compared with a rate of approximately 64% for students with the lowest SAT scores (composite scores of between 600 and 890).

Interestingly, studies have found that admissions tests such as the ACT and SAT have less utility in predicting college graduation when compared with the prediction of college grades (Bowen, Chingos, & McPherson, 2009). Nonetheless, whereas the relationship does decrease, it has been shown that these tests are still fairly effective predictors of eventual college graduation. But unlike in the prediction of FYGPA, the addition of these admission tests to high school GPA does not add that much to the overall accuracy of the prediction equation. If other measures of student success are used, however, such as cumulative GPA or the likelihood of proceeding forward to obtain a graduate or advanced degree, the admissions tests appear to be as likely, or more likely, to predict success on these measures (Burton & Ramist, 2001; Noble & Sawyer, 2002).

The recent work of Kuncel et al. (2001) found that the GRE was an effective predictor of more than just FYGPA for graduate students as well. Kuncel et al. conducted a meta-analysis of more than 1,700 independent samples using a wide variety of criteria to define successful performance. As expected, the study demonstrated a strong relationship between the GRE and FYGPA for graduate students. Perhaps more notably, the study also demonstrated that performance on the GRE was strongly associated with other measures of success, such as scores on certification and licensure tests, publication or citation counts, and faculty ratings.

Admissions tests for professional programs also have evidence for the association between admissions test scores and passing the requirements to practice as a professional in their field (Julian, 2005). The Association of American Medical Colleges has evaluated the relationship between MCAT scores and performance on the U.S. Medical Licensing Examination (USMLE) Step examinations. They also have evaluated how well the MCAT can predict students who will graduate from medical school with distinction or experience difficulty during their time in medical school. They have observed a strong consistent relationship between MCAT test scores and all of these criterion measures, although the relationship between MCAT test scores and performance on the USMLE Step examinations was notably stronger than the others.

Not too surprisingly, as the criterion variables move further away temporally from the test administration, the utility of scores in predicting future success is reduced and practical challenges in conducting such studies increase (Camara & Echternacht, 2000; Mattern et al., 2009). Even the use of a criterion variable such as retention to second year of college presents numerous logistic difficulties that are notably greater than those observed when using FYGPA. Colleges can typically report whether or not students have returned for their second year at their institution. When students have not returned,

however, it is very rare for there to be a systematic and reliable classification scheme for why the student has chosen not to return. This becomes an important distinction because, from a validity perspective, the student who elects to leave a school because of a family emergency is notably different from the student who elects to not return to a school because of academic difficulty, and both are distinct from the student who elects to transfer to a different school.

The increasing number of criteria used in studies has added a layer of complexity into the interpretation of these studies. For example, as the length of time increases from a student taking an admissions test, the question needs to be raised regarding how well we should expect a one-time examination to be able to predict performance. As the length of time increases, the degree of other extraneous or confounding factors that can influence the relationship will continue to grow. Some of these variables can be academic in nature, such as the choice of a challenging major (e.g., chemical engineering), which can affect the cumulative GPA obtained by a student. Other variables can be more personal, as students have increased exposure to new opportunities or ideas that can change their academic plans.

All of the criterion variables available for these studies provide a unique and important contribution to understanding how well each of these tests is able to predict eventual performance in the schools that students are applying to. It is imperative that the test sponsors continue to explore as many of these criterion variables as possible.

Fairness—Performance of Underrepresented Minority Students and Female Students

The large and persistent score differences in mean test scores between underrepresented minorities and other students has been a major source of criticism with admissions tests as well as with all cognitive ability tests (see Chapter 27, this volume). Differences seem to hover close to one standard deviation between the mean score for African American students and White students, whereas the difference is closer to .67 between White and Hispanic students. Roth, Bevier, Bobko, Switzer, and Tyler (2001) conducted a comprehensive meta-analysis of group differences by ethnicity on many of the admissions tests. They found effect size differences of approximately 1.0 for both the SAT and the ACT, with slightly larger estimates (1.34) for the GRE. Schmidt and Camara (2004) reported slightly smaller standardized differences on the ACT and SAT reading and English tests than African Americans when compared with math and science tests. These differences were slightly smaller than gaps found on the GRE, GMAT, LSAT, and MCAT.

The observed differences by gender have been more difficult to characterize. On some tests, such as the SAT Critical Reading and Writing tests and ACT Reading and English tests, performance differences between male and female students has either disappeared or now favor female students. On others, such as the ACT and SAT Math tests, the score gap has been more consistent and pronounced (Hedges & Nowell, 1995; Willingham & Cole, 1997). Across all of the different admissions tests described in this chapter, the gap in performance between male and female students is a notably lower magnitude than the gaps observed for underrepresented minority students. In most cases, the performance gap by gender is closer to 0.10 or 0.25 standard deviation points (Sackett, Borneman, & Connelly, 2008).

These differences in performance have frequently been cited as evidence of test bias against underrepresented minority or female students. However, the idea that fairness or lack of bias is defined by equal performance on these tests for all groups has been consistently refuted within the professional measurement community. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) have stated that "the idea that fairness requires equality in overall passing rates for different groups has been almost entirely repudiated in the professional testing literature" (p. 74). Instead, most researchers have acknowledged that the mean score differences observed reflect the unfortunate disparity in educational opportunities afforded to different groups within society.

Instead, admissions tests routinely apply the regression model first advanced by Cleary (1968) to investigate their examination scores for bias against any particular group. This model uses the regression lines obtained when predicting performance as the criterion to evaluate the fairness of the test in question. The key question asked by the Cleary regression model is whether the expected performance of students with identical or similar predictor scores turns out to be same. In other words, using the SAT as an example, what would be the predicted performance of White students who had an overall SAT composite score of 1600, and how would that compare to the predicted performance of African American students? A test would be considered biased in this model if the predicted performance of African American students was consistently underpredicted by the regression equation. If the regression model predicted a FYGPA for all students of 2.70, but the mean FYGPA for African American students was actually 2.95, the results would support the theory that there was some bias in the use of the test score in the admissions decision. A comprehensive evaluation of test bias would look across the entire score scale, and also would look at the distribution of scores in relation to the predicted scores, to ensure that the test was equally accurate at each score point.

Empirical evidence has consistently shown that test scores do not systematically demonstrate evidence of bias against underrepresented minority students. Both Linn (1973) and Young and Kobrin (2001) conducted extensive reviews of the available studies and found similar results. Both reviews found that admissions test scores slightly overpredict the performance of underrepresented minority students. At the undergraduate level, whereas the amount or degree of overprediction did vary across studies, the amount of overprediction was, on average, approximately 0.20 on a 4.0 GPA scale.

Very similar results have been observed by individual studies conducted by independent researchers as well as researchers associated with each testing program. At the undergraduate level, a study conducted by Noble (2003) investigated the predictive validity of the ACT for underrepresented minority students, whereas a recent study by Mattern, Patterson, Shaw, Kobrin, and Barbuti (2008) did the same for the SAT. Both studies showed consistent results, with their respective admissions test scores slightly overpredicting the performance of underrepresented minority students. Interestingly, they both also observed that the degree of overprediction was even greater for high school GPA than it was for their admissions test scores. In both cases, underrepresented minorities obtained slightly lower college grades than White students who attained the same scores on admissions tests and by using both high school GPA and the admissions test scores, the magnitude of overprediction was notably reduced.

The same pattern of results is observed when investigating the tests used at the graduate level. Recent research focused on the LSAT (Norton, Suto, & Reese, 2006) demonstrated results that were consistent with results observed at the undergraduate level. As with these other studies, GPA and the LSAT slightly overpredicted the performance of underrepresented minority students, with the combination of the two providing the least amount of overprediction.

Interestingly, whereas the score gap between male and female students is notably smaller in magnitude than the gap for ethnic groups, there does seem to be some evidence for a very slight underprediction of females using admissions test scores. Using the same Cleary (1968) regression model, most admissions-testing programs have investigated how well each of their tests predict performance across male and female students. Although the effect is quite small (approximately 0.1 on a 4.0 GPA scale), the underprediction of female student's college GPA has been found in a number of different studies (Leonard & Jiang, 1999; Mattern & Patterson, 2009).

Some studies have investigated different hypotheses for why the underprediction of female students could exist. Some studies have focused on the course selection of students (Ramist, Lewis, & McCamley-Jenkins, 1994) and observed that female students were more likely to enroll in majors in the humanities or social sciences, which had higher overall, mean GPAs than their counterparts in the math and science disciplines. Other studies have examined study habits of students and observed

that female students, in general, exhibited better study habits than male students (Stricker, Rock, & Burton, 1991).

What is known about the large and persistent score gaps between ethnic and racial groups on standardized tests is that similar gaps are found across all types of tests, including performance assessments, standardized tests, and national survey tests (e.g., National Assessment of Educational Progress) as well as college grades, college remediation rates, and college graduation (Camara, 2009). The mere presence of large score gaps is not evidence of test bias or a lack of fairness in test use (Jencks, 1998). The following quote of this issue in a report by the National Research Council, as well as the substantial research on this issue (Cooper, Kuncel, Sackett, Waters, & Arneson, 2006; Sackett, Borneman & Connelly, 2008), should finally silence this allegation:

> Whatever the problems in the construction of the earlier instruments, a considerable body of research has explored the possibility of bias in the current admissions tests, and it has not substantiated the claim that the test bias accounts for score disparities among groups. (Beatty, Greenwood, & Linn, 1999, p. 21)

Role of Socioeconomic Status

One of the more frequent criticisms aimed at standardized testing overall, and admissions testing in particular, is that the tests add no value beyond a reflection of the socioeconomic status (SES) of the test takers. Some of the critics of the SAT have asserted that the SAT "merely measures the size of students' houses" (Kohn, 2001) or that "the only thing that the SAT predicts well is socioeconomic status" (Colvin, 1997). The critics contend that tests like the SAT lose any ability to predict performance in college once variables such as SES are accounted for (Crosby, Iyer, Clayton, & Downing, 2003; Geiser & Studley, 2001).

Sackett, Kuncel, Arneson, Cooper, and Waters (2009) have investigated these criticisms by looking closely at the correlation between SAT and SES, along with the relationship between SAT and college FYGPA and the relationship between SES and FYGPA with data from 41 institutions. In this analysis, the authors were able to analyze the change in the correlation between SAT and FYGPA, after controlling for the relationship between FYGPA and SES. If, as the critics of the SAT maintained, the SAT was solely a proxy for student SES, the correlation between SAT and FYGPA would be significantly reduced to at or near a value of zero. When they conducted the first step of the analysis, they estimated the correlation between the SAT and FYGPA to be equivalent to .47 across the 41 schools in the study. Once they controlled for student SES, the correlation of SAT to FYGPA was reduced by only .03 to .44. The fact that the SAT retained almost all of its predictive validity counters the notion that it offers no utility beyond reflecting a student's SES.

Sackett et al. (2009) also conducted a meta-analysis using information collected from 17 studies that examined the predictive validity of the SAT, ACT, and other examinations. From this data, they estimated the correlation between the admissions test scores and FYGPA to be approximately .37 across these studies. They also estimated the correlation between SES and FYGPA to be .09. Once they controlled for the relationship between FYGPA and SES, the correlation between FYGPA and the admissions test scores was reduced from .37 to .36. This reduction of only .01 in the correlation again provided strong evidence that the admission test scores being used measured something significantly beyond just student SES.

This research does not eliminate the importance that family background, educational quality, or accumulated experiences provide students of privilege. Moderate correlations between test scores and SES are consistently found, but similar correlations exist with other educational predictors and outcomes. For example, raw correlations of .10 and .20 were reported between SAT scores and family income and parental education, respectively. Larger correlations were found, however, between the academic rigor of high school courses taken and family income and parental education, .16 and .25, respectively (Camara, Kobrin, & Sathy, 2005).

Impact of Coaching

Commercial coaching firms have long claimed their coaching courses would lead to score gains of 100 points on the SAT or GRE, and similar gains on other admissions tests. Whereas such claims are obviously attention getting and appealing to students, there is little to no documented evidence to support these claims. In fact, the impact of coaching classes is an area for which very little published research can be found. Research in this area is particularly challenging because the methodology required to conduct a true experimental study would require randomly assigning students to coaching classes or to no preparation at all, which can clearly not be done in the real world. Instead, researchers are left trying to create quasi-experimental studies by comparing students who self-select into coaching classes with those who do not. To make matters even more complicated, almost all studies find that students who enroll and participate in commercial coaching classes are substantially different than students who do not. Because students are not randomly selected into these groups (those who receive coaching and those who do not), such studies continue to be complex and difficult to conduct.

Some studies have managed to create comparable groups using a student's first score on a test like the SAT. When students take the test for a second time, the score gains for the coached students can be compared with the score gain for students who simply retest without coaching or test preparation. Unfortunately, this type of study still does not account for other potential differences in students who participate in coaching, such as motivation, as students who take the time to actually enroll and pay for these classes may be more motivated than those who do not. Even with these limitations, when studies such as Powers and Rock (1998) evaluated the impact of coaching classes on comparable groups, they estimated that the impact of these classes was close to between 9 and 15 points on the SAT-V scale, and between 15 and 18 points on the SAT-M scale.

Scholes and Lain (1997) conducted a similar study using students who took the ACT more than once. Scores from the first time students took the ACT were compared with their scores upon retaking the test. Students were classified into three different groups: (a) those who did no preparation for their second testing, (b) those who used a professional coaching program, and (c) those who prepared on their own using workbooks and other similar preparatory methods. The change in scores for these students was then compared across the three groups. Interestingly, students who prepared on their own actually gained more than the other two groups, although the actual difference in gains across the three groups was so small that the authors did not consider them to be practically significant.

Becker (1990) conducted a meta-analysis of the impact of coaching on SAT test scores. Using all available published materials, Becker estimated that the impact of coaching classes was approximately a 9-point increase in the verbal section, and approximately a 16-point increase in the math section. Although this study is somewhat dated, the results are consistent with the research described in this chapter as well as other coaching research (Powers & Camara, 1999; Powers & Rock, 1998; Scholes & Lain, 1997; Scholes & McCoy, 1998; Zwick, 2002). According to Briggs (2004), there is an emerging consensus that particular forms of coaching can improve scores on admissions tests, but the magnitude of the effect and whether it is worth the associated costs remains in dispute. On the basis of analysis of the National Education Longitudinal Study of 1988, Briggs (2004) reported a coaching effect of about 11 points on SAT-V and 20 points for SAT-M, which is generally consistent with past results (Powers & Camara, 1999).

The National Association of College Admission Counseling commissioned an independent review of the impact of coaching on the scores for students (Briggs, 2009). This study reviewed the available literature on the impact of coaching courses on ACT and SAT scores. Briggs identified more than 30 unique studies that had been conducted on the impact of coaching on SAT scores, but noted that many of these studies had rather small sample sizes or other methodological issues. Instead, he based his conclusions primarily on three large-scale studies that had rather large sample sizes and more rigorous methodology than the

others. Contrary to popular perception, the results did not show large score gains. Instead, scores on the SAT were seen to increase by approximately 30 points for students who took professional coaching courses compared with those who did not. The 30-point increase reflects the increase in SAT scores before 2005, which means the increase is for the 400–1600 scale, not the 600–2400 scale used today.

The review, however, did point to some key limitations that researchers should be aware of when evaluating these studies. First, Briggs (2009) noted that most of the available research had been conducted on the SAT. Few, if any, studies were available on the other major testing programs. The author also noted that most of the research that had been conducted on the SAT was completed before the introduction of the revised SAT in March of 2005. The impact of coaching on this new test has yet to be fully explored. Briggs also noted the inherent limitation of these studies as described thus far. Because true experimental or controlled studies cannot be conducted in this area, Briggs urged all consumers of this research to proceed with some degree of caution in interpreting these results.

Less published research is available on coaching for graduate admissions tests and much of what has been published is now dated. The LSAT program has produced a series of reports evaluating the preparation methods of LSAT test takers, the characteristics of those who prepare using different methods, and their resulting performance (Evans, Thornton, & Reese, 2008; Thornton, Suto, & Reese, 2005). These researchers have consistently shown that students who prepared for the LSAT using workbooks and materials produced by the LSAC consistently outperform those who did not. The research also reports that users of test preparation materials produced by the LSAC tended to be female students and slightly older. These studies did not directly look at the impact of these different preparatory methods. So whereas they are informative and provide a useful snapshot of student preparation for the LSAT, they do not really provide information that allows for an estimate of the preparation effect these methods have on test performance.

THE EVOLUTION OF ADMISSIONS TESTING

No discussion of undergraduate and graduate admissions testing would be complete without mentioning the ever-evolving nature of admissions systems and the role that standardized admissions tests play in that process. Schools and universities are under intense pressure to admit incoming classes that are the most prestigious, the most diverse, and the most dynamic. Because of this enormous pressure, institutions are constantly evaluating their admissions systems and searching to find ways to improve them. Some of the ways this can be seen is through the evolving discussion centered around the appropriate use of admissions tests and the focus on noncognitive measures and high school achievement tests for use in admissions.

Fair Test lists 830 schools that have adopted some form of test-optional policy for admission (Fair Test, 2010). The exact definition of test optional varies by school and can include schools that have decided to not use admission test scores at all when making admission decisions as well as schools that do not require admission test scores for students with high GPAs or students applying to certain programs. Milewski and Camara (2002) inspected the then-current list of 391 schools available from Fair Test and found that a large majority of them were either less competitive or noncompetitive schools. They also found that a significant majority of the schools listed still required an admissions test for most of their students but exempted a percentage of them, such as students who ranked in the top 10% of their graduating class. Nonetheless, there are some well-known and competitive schools on the list, many of which are small liberal arts colleges or technical schools (Fair Test, 2010).

Postsecondary institutions and researchers cite many reasons for choosing and recommending a test-optional policy. The reason given most often is to increase the diversity of the applicant pool to admit greater numbers of racial and ethnic minorities, females, and rural, low-SES, and first-generation college students (Rooney & Schaeffer, 1998). In addition, not requiring test scores encourages

students with a discrepancy between their test scores and high school GPA to apply.

Another challenge within this area is the production of research regarding the utility of admissions programs using a test-optional program. Research on the success of test-optional policies is often cited but suffers from common methodological limitations of study design and sample size. For example, a summary of the research conducted at Bates College concluded that not requiring standardized test scores was linked to an increase in the academic survival of students. Yet, this analysis was based on a sample of only 14 students (Rooney & Schaeffer, 1998).

Test-optional policies are not unique to undergraduate programs, as graduate schools have also examined the possibility of decreasing the weight of the LSAT, MCAT, GMAT (Shultz & Zedeck, 2005), and GRE (Kuncel et al., 2001) in favor of other cognitive and noncognitive measures for admissions.

A range of noncognitive measures traditionally has been used in the admissions process at both the undergraduate and graduate level. The value added of essays, interviews, student interest inventories, and other noncognitive assessments (e.g., measures of personality or study skills) to the admissions process is not consistent or well documented. The main critique of these measures is their lack of standardization and susceptibility to coaching and faking (e.g., Robbins, Lauver, Le, Davis, Langley, & Colstrom, 2004). In addition, there is little agreement on the characteristics that should be measured or the most appropriate methodology for doing so. Finally, validity research conducted on these measures is often confounded by a lack of examinee motivation or social desirability response tendencies (Robbins et al., 2006; Schmitt et al., 2009).

Despite these challenges, evidence for the incremental validity of noncognitive measures to predict program success above and beyond academic record and standardized test scores is growing. Camara and Kimmel (2005) have pointed out that noncognitive measures have their greatest utility in predicting less traditional outcomes, such as leadership, retention, and engagement, across education, employment, and military settings. Other studies have shown that scores on a noncognitive inventory can aid in the prediction of positive college outcomes (e.g., Robbins et al., 2006; Schmitt et al., 2009). For example, the Multiple Mini-Interview (MMI) was created to standardize the administration and scoring of the international medical school admissions interview. The MMI is composed of nine standardized interview prompts that assess advocacy, ambiguity, collegiality and collaboration, cultural sensitivity, empathy, ethics, honesty and integrity, responsibility and reliability, and self-assessment. Research showed that the MMI was able to distinguish between candidates who were accepted or placed on the waiting list for admission (Lemay, Lockyer, Collin, & Brownell, 2007). In addition, the Law School Admission Project has spent the past 9 years focused on creating assessments that would be more appropriate than the LSAT for predicting success as a lawyer, beyond predicting success as a law student (Shultz & Zedeck, 2005).

Researchers have begun to focus on the value of high school accountability tests for postsecondary decisions (e.g., Cimetta, D'Agostino, & Levin, 2010). According to NCLB, all high school students are required to take at least one achievement test in English language arts, mathematics, and science before graduation and in some states passing these tests is required for graduation. Cimetta et al. investigated how well the Arizona Instrument to Measure Standards, the state NCLB test, could predict student college performance, as measured with FYGPA. They found that it accounted for as much variance as the SAT in a model that was combined with high school GPA when looking at students enrolled in the University of Arizona.

The content and skills assessed on state tests are based on each state's unique standards. According to Achieve, Inc. (2004), these standards are often inadequate because they focus on entry-level high school concepts and modest expectations that are typically below the benchmarks of college and career readiness. As such, educational policy makers have called for new tests that are specifically developed to predict readiness for college and career (e.g., Conley, 2007). The Common Core State Standards, which were released in June 2010 and provide rigorous K–12 standards in English language arts and mathematics, may have a long-term impact linking student

performance and expectations from high school to college (Achieve Inc., 2010). At this time, two consortiums of states have formed and received government funding to develop assessments that could potentially measure students' preparedness for college-level work and could also serve as placement tests for students as they enter higher education (see Chapter 15, this volume).

An additional evolving question is about the most efficient and appropriate delivery model for admissions tests. At the professional level, the GRE, GMAT, and MCAT have been successfully transitioned to computer delivery. But the LSAT and the GRE Subject Tests are delivered via paper and pencil. At the undergraduate level, both the ACT and SAT are delivered via paper and pencil. For examinations like the ACT and the SAT, which are delivered to millions of students each year at fairly low costs, one of the biggest challenges continues to be finding a viable alternate delivery model. Schools with computer labs are rarely set up to handle large-scale secure testing requirements, and the costs associated with the delivery in professional testing locations would add greatly to the costs for students. Nonetheless, as more and more course delivery is completed online, students and educators will increasingly demand that these tests move online.

Admissions into postsecondary institutions and graduate programs will always be desired and will always include difficult decisions for all parties involved. The tests reviewed within this chapter can inform and assist with this process and continue to provide critical information to admissions officers across a variety of programs. Because of the potentially competitive nature of admissions, the role that standardized tests play in admissions decisions will always be the subject of intense scrutiny. As the higher education community begins to face the increasing pressure for accountability systems, it is likely to lead to even more pressure to enhance the effectiveness, efficiency, and equity of admissions decisions in schools and the admissions tests used as part of the process.

References

Achieve, Inc. (2010). *Achieving the common core.* Retrieved from http://www.achieve.org/achievingcommoncore

Achieve, Inc., The Education Trust, & Thomas B. Fordham Foundation. (2004). *Ready or not: Creating a high school diploma that counts.* Washington, DC: Achieve.

ACT. (2007). *The ACT technical manual.* Retrieved from http://www.act.org/research/researchers/techmanuals.html

ACT. (2009). *ACT profile report–National.* Retrieved from http://www.act.org/news/data/09/pdf/National2009.pdf

ACT. (2010). *Facts about the ACT.* Retrieved from http://www.act.org/news/aapfacts.html

Allen, D. (1999). Desire to finish college: An empirical link between motivation and persistence. *Research in Higher Education, 40*, 461–485. doi:10.1023/A:1018740226006

Allen, J., Robbins, S., Casillas, A., & Oh, I. (2008). Effects of academic performance, motivation, and social connectedness on third-year college retention and transfer. *Research in Higher Education, 49*, 647–664. doi:10.1007/s11162-008-9098-3

Allen, J., & Sconing, J. (2005). *Using ACT assessment scores to set benchmarks for college readiness* (ACT Research Report RR2003-1). Iowa City, IA: ACT.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Association of American Medical Colleges. (2010). *U.S. medical school applicants and students 1982–83 to 2009–2010.* Retrieved from http://www.aamc.org/data/facts

Baum, S., & Ma, J. (2007). *Education pays: The benefits of higher education for individuals and society*. New York, NY: The College Board.

Beatty, A., Greenwood, M. R. C., & Linn, R. L. (Eds.). (1999). *Myths and tradeoffs: The role of tests in undergraduate admissions*. Washington, DC: National Research Council.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60*, 373–417.

Berry, C. M., & Sackett, P. R. (2008, March). *The validity of the SAT at the individual course level.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Bowen, W. G., & Bok, D. (1998). *The shape of the river: Long-term consequences of considering race in college and university admissions.* Princeton, NJ: Princeton University Press.

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (2009). *Crossing the finish line: Completing college at America's public universities.* Princeton, NJ: Princeton University Press.

Brennan, R. L. (2006). *Educational measurement* (4th ed.). Westport, CT: American Council on Education/ Praeger.

Briggs, D. C. (2004). Evaluating SAT coaching: Gains, effects and self-selection. In R. Zwick (Ed.), *Rethinking the SAT* (pp. 217–233). New York, NY: Routledge-Falmer.

Briggs, D. C. (2009). Preparation for college admission exams (*Discussion paper for the National Association for College Admission Counseling*). Arlington, VA: National Association for College Admission Counseling.

Burnsed, B. (2010, May 14). *GRE is fast becoming a GMAT alternative for B-school applicants.* Retrieved from http://www.usnews.com/articles/education/best-business-schools/2010/05/14/gre-is-fast-becoming-a-gmat-alternative-for-b-school-applicants.html

Burton, N. W., & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980* (Research Report No. 2001-2). New York, NY: The College Board.

Burton, N. W., & Wang, M. L. (2005). *Predicting long-term success in graduate school: A collaborative validity study* (Research Report No. 2005-3). Princeton, NJ: ETS.

Camara, W. J. (2009). College admissions testing: Myths and realities in an age of admissions hype. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 147–180). Washington, DC: American Psychological Association. doi:10.1037/11861-004

Camara, W. J., & Echternacht (2000). *The SAT and high school grades: Utility in predicting success in college* (College Board Research Note RN-10). Retrieved from http://professionals.collegeboard.com/profdownload/pdf/rn10_10755.pdf

Camara, W. J., & Kimmel, E. W. (2005). *Choosing students: Higher education admissions tools for the 21st century*. Mahwah, NJ: Erlbaum.

Camara, W. J., Kobrin, J. L., & Sathy, V. (2005, April). *Is there an SES advantage for the SAT and college success?* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Cimetta, A. D., D'Agostino, J. V., & Levin, J. R. (2010). Can high school achievement tests serve to select college students? *Educational Measurement: Issues and Practice*, *29*(2), 3–12. doi:10.1111/j.1745-3992.2010.00171.x

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. doi:10.1111/j.1745-3984.1968.tb00613.x

Clinedinst, M., & Hawkins, D. (2009). *State of college admission*. Alexandria, VA: National Association for College Admission Counseling.

College Board. (2009a). *College bound juniors and sophomores 2009*. Retrieved from http://professionals.collegeboard.com/data-reports-research/psat/cb-jr-soph

College Board. (2009b). *2009 college-bound seniors: Total group profile report*. Retrieved from http://professionals.collegeboard.com/profdownload/cbs-2009-national-TOTAL-GROUP.pdf

College Board. (2010). *About the SAT*. Retrieved from http://professionals.collegeboard.com/testing/sat-reasoning/about

Colvin, R. L. (1997, October 1). Q & A: Should UC do away with the SAT? *Los Angeles Times,* p. B2.

Conley, D. T. (2007). *College readiness*. Eugene, OR: Educational Policy Improvement Center.

Cooper, S. R., Kuncel, N. R., Sackett, P. R., Waters, J., & Arneson, S. D. (2006, April). *The role of SES in the ability–performance relationship: Results from national longitudinal studies*. Poster session presented at the annual meeting of The Society for Industrial and Organizational Psychology, Dallas, TX.

Crosby, F. J., Iyer, A., Clayton, S., & Downing, R. A. (2003). Affirmative action: Psychological data and the policy debates. *American Psychologist, 58*, 93–115. doi:10.1037/0003-066X.58.2.93

Educational Testing Service. (2010a). *GRE*. Retrieved from http://www.ets.org/gre

Educational Testing Service. (2010b). *The GRE Revised General Test: Better by design*. Retrieved from http://www.ets.org/Media/Tests/GRE/pdf/12903_GREr_overview_brochure.pdf

Educational Testing Service. (2011). *Taking the GRE general test for business school*. Retrieved from http://www.ets.org/gre/general/about/mba

Evans, J., Thornton, A. E., & Reese, L. M. (2008). *Summary of self-reported methods of test preparation by LSAT takers for testing years 2005–2006 through 2007–2008* (LSAT Technical Report No. 08-04). Newtown, PA: Law School Admission Council.

FairTest. (2010). *Test score optional list*. Retrieved from http://www.fairtest.org/university/optional

Fisher, J. B., & Resnick, D. A. (1990). Standardized testing and graduate business school admission: A review of issues and an analysis of a Baruch College MBA cohort. *College and University, 65*, 137–148.

Geiser, S., & Studley, R. (2001). *UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California*. Retrieved from http://www.ucop.edu/sas/research/researchandplanning/pdf/sat_study.pdf

Goldberg, J., & Smith, J. (2008). The effects of education on labor market outcomes. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 688–708). New York, NY: Routledge.

Graduate Management Admission Council. (2009). *Profile of Graduate Management Admission Test Candidates.* Retrieved from http://www.gmac.com/NR/rdonlyres/EEFE1A18-4FCE-421D-ACE1-67C9D86A444B/0/ProfileofGMATCandidates0509.pdf

Graduate Management Admission Council. (2010). *The GMAT exam.* Retrieved from http://www.gmac.com/gmac/thegmat

Handwerk, P. (2009). *National applicant trends—2008.* Retrieved from http://www.lsac.org/LSACResources/Data/national-applicant-trends.sps

Hedges, L. V., & Nowell, A. (1995). Sex difference in mental test scores, variability, and number of high scoring individuals. *Science, 269*, 41–45. doi:10.1126/science.7604277

Hezlett, S. A., Kuncel, N. R., & Vey, M. A. Ahart, A. M., Ones, D. S., Campbell, J. P., & Camara, W. J. (2001, April). *The effectiveness of the SAT in predicting success early and late in college: A comprehensive meta-analysis.* Paper presented at the annual meeting of The Society of Industrial-Organizational Psychology, San Diego, CA.

Jencks, C. (1998). Racial bias in testing. In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 55–85). Washington, DC: Brookings Institution Press.

Julian, E. R. (2005). Validity of the Medical College Admission Test for predicting medical school performance. *Academic Medicine, 80*, 910–917. doi:10.1097/00001888-200510000-00010

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average* (Research Report No. 2008-5). New York, NY: The College Board.

Kohn, A. (2001, March 9). Two cheers for an end to the SAT. *Chronicle of Higher Education*, p. B12.

Kuncel, N. R., Crede, M., & Thomas, L. L. (2007). A meta-analysis of the predictive validity of the graduate management admission test (GMAT) and undergraduate grade point average for graduate student academic performance. *Academy of Management Learning and Education, 6*, 51–68. doi:10.5465/AMLE.2007.24401702

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science, 315*, 1080–1081. doi:10.1126/science.1136618

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate students election and performance. *Psychological Bulletin*, *127*, 162–181. doi:10.1037/0033-2909.127.1.162

Lacey, J. N., & Crosby, O. (2004). Job outlook for college graduates. *Occupational Outlook Quarterly*, *48*(4), 15–27.

Law School Admissions Council. (2010). *The LSAT: About the LSAT.* Retrieved from http://www.lsac.org/LSAT/about-the-lsat.asp

Lemay, J.-F., Lockyer, J. M., Collin, V. T., & Brownell, K. W. (2007). Assessment of non-cognitive traits through the admissions multiple mini-interview. *Medical Education, 41*, 573–579. doi:10.1111/j.1365-2923.2007.02767.x

Leonard, D. K., & Jiang, J. M. (1999). Gender bias and the college predictions of the SATs: A cry of despair. *Research in Higher Education, 40*, 375–407. doi:10.1023/A:1018759308259

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43*, 139–161.

Mattern, K. D., Kobrin, J. L., Patterson, B. F., Shaw, E. J., & Camara, W. J. (2009). Validity is in the eye of the beholder: Conveying the SAT research findings to the general public. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 213–240). Charlotte, NC: Information Age.

Mattern, K. D., & Patterson, B. F. (2009). *Is performance on the SAT related to college retention?* (Research Report No. 2009-7). New York, NY: The College Board.

Mattern, K. D., & Patterson, B. F. (2011). *Validity of the SAT for predicting second year grades: 2006 validity sample* (Research Report No. 2011-1). New York, NY: The College Board.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of the SAT* (Research Report No. 2008-4). New York, NY: The College Board.

Milewski, G. B., & Camara, W. J. (2002). *Colleges and universities that do not require SAT or ACT scores* (Research Note RN-18). New York, NY: The College Board.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education, 40*, 355–371. doi:10.1023/A:1018755201899

Noble, J. (2003). *The effect of using ACT composite score and high school average on college admission decisions for racial/ethnic groups* (ACT Research Report RR2003-1). Iowa City, IA: ACT.

Noble, J., & Sawyer, R. (2002). *Predicting different levels of academic success in college using high school GPA and ACT composite score* (ACT Research Report No. 2002-4). Iowa City, IA: ACT.

Norton, L. L., Suto, D. A., & Reese, L. M. (2006). *Analysis of differential prediction of law school performance by racial/ethnic subgroups based on 2002–2004 entering law school classes* (LSAT Technical Report No. 06-01). Newtown, PA: Law School Admission Council.

Powers, D. E., & Camara, W. C. (1999). *Coaching and the SAT I* (Research Note RN-06). New York, NY: The College Board.

Powers, D. E., & Rock, D. A. (1998). *Effects of coaching on SAT I: Reasoning scores* (Research Report No. 98-6). New York, NY: The College Board.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (Research Report No. 93-1). New York, NY: The College Board.

Robbins, S. B., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology*, *98*, 598–616. doi:10.1037/0022-0663.98.3.598

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Colstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, *130*, 261–288. doi:10.1037/0033-2909.130.2.261

Rooney, C., & Schaeffer, B. (1998). *Test scores do not equal merit: Enhancing equity and excellence in college admissions by deemphasizing SAT and ACT results.* Cambridge, MA: FairTest.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297–330. doi:10.1111/j.1744-6570.2001.tb00094.x

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*, 215–227. doi:10.1037/0003-066X.63.4.215

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). *Socio-economic status and the relationship between the SAT and freshman GPA: An analysis of data from 41 colleges and universities* (Technical Report No. 2009-1). New York, NY: The College Board.

Schmidt, A. E., & Camara, W. J. (2004). Group differences in standardized test scores and other educational indicators. In R. Zwick (Ed.), *Rethinking the SAT* (pp. 189–216). New York, NY: Routledge-Falmer.

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact of demographic status of admitted students. *Journal of Applied Psychology*, *94*, 1479–1497. doi:10.1037/a0016810

Scholes, R. J., & Lain, M. M. (1997, March). *The effects of test preparation activities on ACT assessment scores.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Scholes, R. J., & McCoy, T. R. (1998, April). *The effects of type, length, and content of test preparation activities on ACT Assessment scores.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Shelton, P. D. (1997). The LSAT: Good—but not that good. *Law School Services Report, 97*, 2–3.

Shultz, M., & Zedeck, S. (2005). *What makes for good lawyering? A multi-year study looks beyond the LSAT.* Retrieved from http://www.law.berkeley.edu/beyondlsat/transcript.pdf

Stilwell, L. A., Dalessandro, S. P., & Reese, L. M. (2007). *Predictive validity of the LSAT: A national summary of the 2005–2006 correlation.* Newtown, PA: Law School Admission Council.

Stricker, L., Rock, D., & Burton, N. (1991). *Sex differences in SAT prediction of college grades* (Research Report No. 91-2). New York, NY: The College Board.

Thornton, A. E., Suto, D. A., & Reese, L. M. (2005). *Summary of self-reported methods of test preparation by LSAT takers for testing years 2003–2004 and 2004–2005* (LSAT Technical Report No. 05-01). Newtown, PA: Law School Admission Council.

U.S. Census Bureau. (2002). *The big payoff: Educational attainment and synthetic estimates of work-life earnings.* Retrieved from http://www.census.gov/prod/2002pubs/p23-210.pdf

Wightman, L. F., & Jaeger, R. M. (1998). *High stakes and ubiquitous presence: An overview and comparison of standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Wiley, A., Wyatt, J. N., & Camara, W. J. (2010). *The development of a multidimensional index of college readiness for SAT students* (Research Report No. 2010-3). New York, NY: The College Board.

Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability*. New York, NY: The College Board.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Erlbaum.

Young, J. W., & Kobrin, J. L. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (Research Report No. 2001-6). New York, NY: The College Board.

Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education.* New York, NY: Routledge-Falmer.

CHAPTER 15

ASSESSMENT IN HIGHER EDUCATION: ADMISSIONS AND OUTCOMES

Diane F. Halpern and Heather A. Butler

For most people, the word *assessment* conjures images of stacks of paper exams replete with multiple choice questions, but assessment is a broad term that includes much more than testing. We define *assessment* in the context of higher education as the process of gathering information that can be used to inform decisions related to teaching and learning. We focus on two crucial points in time in the life cycle of higher education—assessment related to college and university admissions and assessment of student learning outcomes. Many people have a stake in assessments in higher education. For students and their parents, assessments are important because they can influence what college the student gets into, whether the student receives funding while in college, whether the student graduates, and whether the student gets accepted into graduate school or heads into a choice career. For colleges and universities, assessment outcomes can affect funding, the prestige of the university, which students choose to attend, and faculty salaries and tenure. In this chapter, we examine the wide variety of assessments associated with higher education, with a special emphasis on their validity, the potential and real biases associated with each type of assessment, and the added value of assessment activities beyond grade point averages (GPAs). We also examine the influence that assessment has on college rankings, group differences in performance, stereotype threat, and the recent inclusion of noncognitive measures in assessments of college admissions and student learning outcomes.

USING ASSESSMENTS TO DETERMINE COLLEGE ADMISSIONS

A college degree has become the passport to the middle class. It is the difference between having an interesting career with possibilities for advancement and a dead-end job at low wages. An educated citizenry is also essential for the economy of every country. A recently released report in the United States found that

> by 2018, we will need 22 million new workers with college degrees—but will fall short of that number by at least 3 million postsecondary degrees . . . At a time when every job is precious, this shortfall will mean lost economic opportunity for millions of American workers. (Carnevale, Smith, & Strohl, 2010, para. 1)

With stakes this high it is not surprising that many individuals are critical of standardized testing, which serves as the entrance gate for most institutions of higher education. Some institutions have open admissions policies or policies that allow students with high enough GPAs entrance without standardized admissions exam scores, but the vast majority of colleges and universities in the United States require these standardized assessments. Although many community colleges and other open-admissions institutions have produced exceptionally talented graduates, most open-admissions institutions have low prestige and high dropout rates (Horn, Nevill, & Griffith, 2006).

DOI: 10.1037/14049-015
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

Scores on standardized tests are a valuable tool for college administrators because they provide a more objective assessment of ability than high school GPAs. First, although GPA certainly will vary based on ability, GPA also will vary based on the type of classes taken, the intensity of the classes, and the quality of the school. For example, two students of equal ability may have different GPAs because one student took remedial or "easier" courses, whereas the other student took challenging courses. Second, over the past few decades, grade inflation has become a problem in the United States, Canada, England, and other countries. Not only are grades increasing at an accelerating rate, but also the average grade given by an instructor is highly correlated with positive course evaluations (Blackhart, Peruche, DeWall, & Joiner, 2006). A report from the Higher Education Research Institute (1999) found that 34.1% of college freshmen claim that they finished high school with an A average, a figure that has increased steadily since 1969. Alexander Astin, founding director of the Higher Education Research Institute, suggested that grade inflation in high school is increasing because students and their parents pressure teachers to help them become more competitive for college (de Vries, 2003). The numerous causes of grade inflation are beyond the scope of this chapter; the germane point is that grade inflation has greatly reduced the predictive power of high school GPAs, rendering them less useful for college admissions committees or as legitimate measures of student learning. Although high school GPA may be limited in its usefulness as a college admissions variable, standardized college entrance exams also have limitations.

In general, there are two types of tests, achievement tests and aptitude tests. Achievement tests are designed to measure past learning; aptitude tests are designed to measure ability and predict future performance. Higher scores on aptitude tests correspond to higher developed ability, which is often used as a proxy for intelligence. In the United States, the two most common tests for college admissions are the SAT, which is closer in its conceptualization to an aptitude test, and the American College Test (ACT), which is closer in its conceptualization to an achievement test. Taken together, these two admissions tests have powerful effects on the future lives of many people, especially for those who engage in the "ferocious competition" for admission to highly selective institutions. In 2010, more than 1.5 million students took the SAT (The College Board, 2010), more than 1.5 million took the ACT (Sawyer, 2010), and many students took both exams. The current SAT measures critical reading ability (formerly verbal ability), math ability, and writing ability. Scores on the exam range from 600 to 2400. The SAT II refers to subject area tests that are frequently taken along with tests of critical reading, mathematics, and writing, which collectively are referred to as the SAT I. The ACT was designed as an achievement test, and thus it should more closely match what is taught in high school. As Atkinson and Geiser (2009) pointed out, the major problem with this approach is that the United States does not have a national curriculum, and even though the ACT attempts to match what is taught in an average curriculum, there can be large differences between the materials that are assessed on the ACT and any student's actual high school curriculum. Over the many years that these two tests have coexisted, the ACT and SAT have become more similar, and many college and universities accept either test score. Like all psychological constructs, the SAT and ACT are imperfect measures that depend in part on past performance, socioeconomic status, conscientiousness, and a variety of other variables. The next section of this chapter mostly focuses on research conducted on the SAT; readers interested in the relevant research on the ACT should see the Chapter 14 in this volume.

The Validity of College Admission Assessments

The validity of a college entrance exam concerns values, test usage, and statistics. A valid measure must demonstrate that it is measuring something "real" and accurate as determined by society. For example, most college entrance exams were developed to identify students who have the ability to succeed in college, but what variables can accurately predict an individual's ability to succeed? Is ability a product of innate intelligence, past learning, motivation, or some combination of these and other traits?

Is success in college best measured by a student's GPA, breadth and depth of learning, or whether that student finds a good job after graduation? Or even more broadly, is success measured by how much money a graduate earns or how happy the graduate is? Answers to these murky questions greatly influence how an assessment is developed, used, and validated.

The question of whether a measure is valid will depend on its use. It is a question of "valid for what purpose?" For example, consider scores on the Graduate Record Examination (GRE), which is a commonly used test to inform graduate school admission decisions. Although it is designed to be taken at the end (or near the end) of an undergraduate education and the subject area tests include some content that is learned and developed in baccalaureate programs, it is not a valid assessment of what students have learned in their undergraduate program because it is designed to provide information about students who are applying to graduate school, and the vast majority of undergraduates are not planning graduate study. For this reason, using scores on the GRE to determine whether students should receive their bachelor's degree is an inappropriate use of a measure that is valid when used for other purposes.

Valid measures must demonstrate certain statistical properties. That is, a strong relationship between scores on an assessment and the criterion should be found. For example, statistical validity would be apparent if scores on a college entrance exam predicted college GPA (or another operationalization of college success). Two types of statistical validity that are of greatest concern with regard to college entrance exams are construct validity and predictive validity. Construct validity refers to whether an assessment (or an operationalization of the construct) measures what it was intended to measure. For example, does an intelligence quotient test accurately measure intelligence? Does a score on scholastic aptitude assessment accurately measure an individual's ability to succeed in college? Predictive validity refers to whether scores on an assessment can accurately predict future performance. For example, if individuals receive high scores on a critical thinking assessment, will they critically analyze a political speech, avoid clever Internet scams, or make more informed decisions in their personal and professional life? Will an individual who scores high on a scholastic aptitude assessment do well in college? How well do these examinations predict first-year grades in college?

The current SAT consists of three sections, a critical reading section (SAT-CR), a mathematics section (SAT-M), and a writing section (SAT-W). To determine the predictive validity of the newly revised SAT, The College Board, conducted a study of more than 196,000 students from 110 colleges and universities in the United States (Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008). The students took the SAT in 2006 and completed their first year in college in 2007. The students' SAT scores were correlated with their high school GPA and their first-year college GPA. The range of students was necessarily restricted because the sample included only those students who were admitted to college, not the entire population of students who took the SAT. Accordingly, The College Board made a statistical correction for the restricted range by using the Pearson–Lawley multivariate correction (Kobrin et al., 2008), and adjusted values are reported here.

Overall, the SAT predicted 1st-year GPA in college quite well, but some portions of the assessment were better predictors than others (Kobrin et al., 2008). The combined SAT score is correlated with first-year college GPA ($r_{adjusted} = .53$). Of the three SAT portions, the new writing portion is the best predictor of first-year college GPA ($r_{adjusted} = .51$), followed by the critical reading portion ($r_{adjusted} = .48$), and then the math proportion ($r_{adjusted} = .47$). Controlling for high school GPA, the SAT provides an additional increment of .08 to the predictive validity of the assessment. It seems that this is a very small value for incremental validity, but some have argued that this seemingly small increase in explained variance has a meaningful effect when predicting the percentage of students who succeed in college, especially for highly selective colleges (Bridgeman, Pollack, & Burton, 2004). Furthermore, if a school used only high school GPA for college admissions, this measure would underpredict the college performance of students who performed significantly

higher on the SAT than on their high school GPA. Thus, The College Board recommends that both measures be used to make college admissions decisions (Kobrin et al., 2008). Other researchers also "correct" for other statistical problems that occur, including the unreliability of the criterion variable, which in this case is college GPAs (e.g., Ramist, Lewis, & McCamley-Jenkins, 1994). These corrections increase the predictive validity of the SATs.

There are many strident critics of the use of standardized tests for the purpose of college admissions. Fair Test (see http://www.fairtest.org/university/ACT-SAT), a national group that is generally opposed to the use of standardized tests, calls the SAT inaccurate, biased, and susceptible to coaching. The use of standardized exams is a hot-button topic for many people. It may seem as though an open review of the research literature on standardized exams would settle the question about their validity and usefulness, but despite a growing body of quality research on this question, the controversy is far from resolved.

Simple questions, complex answers. On the basis of data supplied from The College Board, it may seem that the SATs add little incremental validity beyond high school grades in predicting college success. But recent research by Berry and Sackett (2009) has shown that even The College Board's own data underestimate the validity of the SATs. These researchers argued that usual estimates of validity are contaminated by fact that the outcome we want these tests to predict is academic performance, not college GPAs. Of course, college GPAs reflect academic performance, but many other variables that are unrelated to academic performance go into GPAs, including differences in how professors assign grades and differences in the types of courses that students take. To reduce some of the "error variance" in GPAs, these researchers used individual course grades as their criterion variable. Individual course grades are the components of GPAs, but when analyzed on the course level, the researchers argued, they are measuring academic performance without many of the confounds that contaminate GPAs. (Interested readers are referred to Berry & Sackett, 2009, for statistical details.) On the basis of sophisticated analyses using 167,816 students at 41 colleges, they concluded that the validity of the SATs is underestimated when freshman or graduation GPAs are used as the predicted outcome. Berry and Sackett made a few more statistical assumptions (regarding the way students use SAT data to determine which colleges to apply to) and then concluded that .61 and .71 are their best estimates for the predictive validity of SATs and high school GPAs, respectively. These values exceed even the highest estimates published by The College Board. Berry and Sackett also concluded that the SATs and high school GPAs have incremental validity, and both measures should be used simultaneously in making college admissions decisions. If used singly, high school GPAs have slightly higher predictive validity than SATs.

Although Berry and Sackett's (2009) research yielded easy-to-interpret policy recommendations, a major problem is that the reasoning behind those recommendations is uninterpretable by most people because their methods require an advanced understanding of statistical concepts. Concepts like predictive validity are deceptively simple, but even basic testing principles, such as restriction of range and unreliability of the criterion variable, which were discussed earlier, can seem like a smoke screen that is designed to make the predictive validity larger than it is without these adjustments, especially when low scores on high-stakes exams are limiting opportunities for individuals and groups of people.

Alternatives and Additions to the SATs: The Role of Noncognitive Factors

Even the strongest advocates of the SATs (and other standardized high stakes tests) recognize that there is considerable variability in college success for which these tests cannot account. As a way of accounting for variance that is missed with standardized tests, there is an increasing trend in higher education admissions to use noncognitive measures. For example, a recent study found that a personality trait, conscientiousness, predicted success in college better than SAT scores. A conscientious person is careful, thoughtful, self-disciplined, organized, and dependable. Typically, conscientious individuals would be described as having character, being goal oriented, and being hard working. Wagerman and

Funder (2007) collected traditional predictors of college success (i.e., high school GPA, SAT scores), GPA during freshman and senior years, cumulative GPA, and level of conscientiousness (measured with the Big Five inventory) that was rated by both the participants and an informant. Conscientiousness was positively correlated with college success at all academic levels. Combined with the traditional predictors of academic success, 18% of the variance in 1st-year GPA and 37% of the variance in senior-year GPA was explained. Additionally, conscientiousness explained a unique portion of the variance beyond the traditional predictors and the predictive validity did not vary based on race. This finding may lead to some promising discoveries in the future, such as the possible inclusion of measures of conscientiousness for college admissions decisions.

Schmitt et al. (2009) looked to other areas of selection, most notably hiring, for additional means of assessing college success. They used a variety of outcome variables in addition to college GPA, including absences from class, likelihood of graduating, and engagement in college life. These variables form two broad classes of outcomes: (a) biodata, which includes gaining content knowledge, responsible citizenship, physical and psychological health, ethics and integrity, among others; and (b) situation judgment, which is a measure of how well individuals respond to scenarios that are likely to be common on college campuses. The researchers in this study of noncognitive variables concluded that although SATs and high school GPAs have good validity, the use of biodata and situation judgments add incremental validity to decisions about college admissions. Schmitt et al.'s experimental study was not conducted in a context in which the data they collected actually affected college admissions decisions. They noted that their experiment with noncognitive variables may not hold up if the students were taking these tests as part of a "real" college admissions process because it is not known how students would respond to questions when they know the results are being used for a high-stakes purpose. This criticism is common when the inclusion of noncognitive measures in college entrance admission is debated. Many noncognitive factors are self-reported, rendering them susceptible to faking and social desirability biases. Additionally, the authors noted that if their experiment had been used for admissions, a greater proportion of Hispanic and African American students would have been admitted, and a smaller proportion of Asian American and Caucasian students would have been admitted. Given that group differences in college-going rates have fueled much of the debate over the use of standardized exams in college admissions, the finding that the use of additional measures would change the composition of college classes means that the addition of these noncognitive measures are likely to be subject to scrutiny.

The Rainbow and Kaleidoscope Projects were designed to increase predictive validity in the college admissions process while also reducing disparities in admission rates for ethnic minority groups (Sternberg, 2006, 2009; for elaboration on aptitude assessment, see Chapter 13, this volume). These projects are based on a three-part classification of the skills that are needed to succeed in almost any career choice: creativity, analytical ability, and practical intelligence. The researchers report that by adding measures of these three abilities (plus a fourth ability—wisdom, which was added to later studies), they were able to achieve the dual goals of better academic prediction and reduced (or eliminated) differences across ethnic minority groups. The researchers on this project are clear that the goal is not to replace the SATs or other traditional indicators of academic success such as high school GPA and class rank. The measures they are championing are designed to be used to increase predictive validity beyond what is already accounted for by the traditional indicators. Sternberg (2009) concluded that "it is possible to increase academic quality and diversity" (p. 284), while also signaling to all stakeholders in higher education that college applicants are more than the narrow range of skills assessed in standardized admissions tests.

Taken together, these two studies (and other research that has used additional measures for constructs that are not assessed in the SATs) raise important questions about the future of assessment in college admissions decisions. Does the use of noncognitive variables mean that the college admissions process will be fairer for some groups? Can an exam

be fair to some groups and unfair to others at the same time? These are questions about values, statistics, discrimination, and, ultimately, what is right.

Fairness of College Admission Assessments

Like validity, fairness is both a statistical concept and an issue of values. To be fair is to be free from bias. In terms of testing fairness, several concepts of fairness merit consideration (Society for Industrial and Organizational Psychology, Inc., 2003): (a) fairness as equitable treatment in testing conditions (e.g., all test takers take the test under the same conditions), (b) fairness as comparable opportunity to learn the material (e.g., all test takers have the same access to the materials), (c) fairness as equal group outcomes (e.g., equal numbers of male and females pass an assessment), and (d) fairness in terms of being able to predict a test score equally well for all groups (e.g., the score on the measure predicts scores for males equally as well as it predicts scores for females). This last definition of fairness does not mean that two groups will have the same group mean on a particular measure, but that the measure predicts scores on the criterion equally well for both groups. If there are group differences in the percentages that achieve high scores, the test has "adverse impact," which means that it adversely affects some group or groups relative to some other group or groups. According to standard psychometric principles, an assessment can have adverse impact and still be unbiased, as long as it predicts equally well for all groups.

At times, society's idea of fairness will conflict with standard definitions of statistical fairness. Consider, for example, that there is a sensitive period in language development in which children learn the grammatical rules of a given language. Children who learn a second language within this sensitive period (birth to 7 years old) tend to develop a better grammatical understanding of that language than children who learn a second language after this sensitive period (Johnson & Newport, 1989). In this case, a statistically fair measure of grammatical knowledge is not necessarily an assessment in which both of these groups (early language learners vs. late language learners) score equally well. On the basis of what we know of language development, we would expect that early language learners would have higher scores than late language learners. Thus, a statistically fair assessment is an assessment that predicts how well an individual uses grammar in real-life situations, which is influenced by the age at which the child acquired the second language. This idea of statistical fairness directly conflicts with societal notions of fairness as equality. For many people, it just "feels" wrong and unfair when an assessment differentially predicts an individual's score based on group membership, regardless of whether the group is defined as ethnicity/race, gender/sex, socioeconomic status, age at which a second language is learned, or hair color. The debates over fair assessment have never been more heated than the debates over the fairness of college entrance exams.

Consider the definition of fairness in testing proposed by Helms (2006):

> Any time a test yields mean test scores that differ between racial groups, then use of the test to assess individuals is potentially unfair even if considerable evidence exists that the test yields valid and nonbiased scores between and within racial groups. (p. 845)

Thus, she takes the perspective that validity evidence is necessary for test fairness, but it is not sufficient. Helms has argued that racial group is used as a proxy for socialization experiences that affect the way individuals react to the testing situation, and any test with group differences in average scores is unfair.

In the United States, as well as other countries, certain groups may be at a disadvantage with regard to standardized testing. Typically, Caucasian males score higher on many standardized tests than women and most ethnic minorities. A few exceptions to this are in the field of mathematics in which Asian males tend to dominate, and on essay tests in which women tend to dominate (The College Board, 2009). The answer to why these group differences exist is multifaceted and complicated by numerous factors, including socioeconomic status, learning experiences, and parental education level, to name a few. What is clear from the massive

research literature on group differences in tests of cognitive ability is that there is not just one answer to the question of what is causing these differences. Group differences such as these call into question the validity and fairness of assessments used for college admissions.

Ethnic Differences in College Admission Assessments

There are several notable differences in SAT scores based on ethnicity. Differences in SAT scores between African Americans and Caucasians have received the most attention historically, in part because of the size of the difference. African Americans score an average of 209 points (combined math and reading) lower than Caucasians on the SAT (The College Board, 2009). This effect is robust, and it is not uncommon to see a difference of 1 standard deviation between the two groups. In a 1993 study of racial bias, Roy Freedle (2003) argued that the SAT items were biased in favor of Caucasians. His article was criticized by The College Board, which asserted that the SAT was a fair assessment and that the differences in SAT scores were due to an unfair American society (Jaschik, 2010). A more recent examination of the SAT reveals that certain items on the SAT may be racially biased. Santelices and Wilson (2010) used differential item functioning (DIF) with the latest SAT data to test whether certain SAT items favored one ethnic group over another. They concluded that some items of the assessment are racially biased. For example, the easier verbal items favored Caucasians, whereas the more difficult verbal items favored African Americans. Santelices and Wilson argued that the testing community has an obligation to uncover the reasons for these differences. Although DIF values can identify potential biases, they cannot be used alone to prove racial bias. A more thorough analysis of the item needs to be conducted (for more information on the assessment of DIF, see Volume 1, Chapters 7 and 8, this handbook).

Smaller differences in SAT scores, although no less important, were found for every other ethnicity examined. In a study of the validity of the SATs for college decisions in California, researchers found that when all scores were combined, including SAT-II scores, which are more closely tied to subject area knowledge, Latinos scored .9 of a standard deviation below Caucasian students, and native Americans scored .5 of a standard deviation below (Kobrin, Camara, & Milewski, 2002). By contrast, Asian students scored slightly higher than Caucasian students on math tests and slightly lower on verbal tests. These findings need to be considered in light of findings that SATs also show moderate correlations with family income (between .25 and .55) and parental education (between .28 and .58), leading critics to claim that the SATs are just proxy measures of family wealth.

Sackett, Borneman, and Connelly (2008) countered the claim that "the SAT merely measures the size of students' houses" (quote from Kohn, 2001, p. B12). Their argument is both statistical and logical. On the statistical front, they agree that there is a "substantial relationship" between SATs and socioeconomic status, but the predictive validity is "only affected by a small degree" when controlling for socioeconomic status (Sackett et al., 2008, p. 221). Logically, they argued that socioeconomic status is an important contributor to the development of cognitive abilities that predict college success, so it is a relevant variable in understanding academic success. On the basis of a meta-analysis of the effect of socioeconomic status on the predictive validity of the SAT, Sackett, Kuncel, Arneson, Cooper, and Waters (2009) concluded that the SAT retains virtually all of its predictive value when controlling for socioeconomic status.

Several explanations for why standardized exams underestimate the abilities of African Americans and other non-Asian ethnic minorities have been proposed (cf. Helms, 1992). Some scholars have emphasized the biological or genetic differences between African Americans and Caucasians, other scholars have argued that the differences can be explained by environmental factors, and still others have emphasized cultural differences. Furthermore, students' perception of the validity or fairness of an assessment can influence the outcome of an assessment. Chan, Schmitt, DeShon, Clause, and Delbridge (1997) examined the relationships among race, test-taker motivation, perceptions of face validity, and test performance. Not only did

test-taker motivation predict test performance, but also this factor was especially pertinent for African American test takers. There was a stronger relationship between test-taker motivation and test performance for African American test takers than for White test takers. Additionally, perceptions of the face validity of the assessment indirectly influenced test-taking motivation, which in turn predicted test performance. That is, test takers who thought the face validity of the assessment was low were less motivated and scored lower on the tests of cognitive ability. Although it has long been known that student (and teacher) expectations about achievement often predict performance, these test-taker expectations that exist because of their group membership can be especially troublesome because they are so difficult to control.

The stereotypes associated with academic achievement can influence performance on an assessment. Stereotypes are overgeneralizations made about members of particular social groups. In the United States, and several European countries, there is a stereotype that African Americans and other non-Asian ethnic minorities are not as intelligent as Caucasians. The "threat" of this stereotype is enough to depress exam performance (Steele & Aronson, 1995; for clarification of this concept, see Volume 1, Chapter 36, this handbook). This effect is known as stereotype threat and has been found on several college admissions exams including the SAT, ACT, state-mandated standardized tests, and the GRE (Walton & Spencer, 2009). The typical stereotype threat paradigm involves activating the stereotype that one group is superior to another on the assessment simply by instructing the student that the assessment measures intellectual ability (or another stereotype-relevant domain). In a meta-analysis that included more than 3,000 students in five countries, stereotyped students performed worse than nonstereotyped students (Walton, & Spencer, 2009). Whereas nonstereotyped students typically do as well or better in the threat condition than they do in the control condition, stereotyped students perform worse under conditions of threat than they do in the control condition. This effect occurs for students of all levels of abilities, as measured by their prior performance in classes or prior assessments. For low-performing students (those who scored 1 standard deviation below the mean on prior performance measures), the size of the effect was $d = .14$; for medium-performing students (those who scored between 1 standard deviation above and below the mean), the size of the effect was $d = .18$; and for high-performing students (those who scored 1 standard deviation above the mean), the size of the effect was $d = .22$. On the basis of the effect size obtained in the meta-analysis, stereotyped African American students underperform on the SAT by an average of 39 points. The stereotype threat effect occurs not only for non-Asian ethnic minorities but for gender stereotypes as well.

Sex Differences in College Admission Assessments

There are more similarities among men and women than there are differences, but a few sex differences in cognitive ability can influence course grades, scores on standardized exams, admission into college, choice of career, and much more (Hyde, 2005). Generally speaking, women tend to outperform men on some tests of verbal ability, especially writing, but underperform on tests of visuospatial skills and quantitative ability (Halpern et al., 2007). As with the research exploring ethnic differences in cognitive abilities, researchers have attempted to explain these differences from a variety of perspectives, including biological, evolutionary, neuroscientific, social, and environmental perspectives. The variety of explanations for sex differences in cognitive abilities is beyond the scope of this chapter (for a review, see Halpern et al., 2007).

Three main categories of cognitive abilities are most often studied: verbal ability, visuospatial ability, and quantitative ability (e.g., Halpern, 2011). The extent to which sex differences are found depends largely on which ability is being studied, how the ability is measured, the age and context in which the ability is measured, and whether the groups being compared are at the extreme ends of the bell curve where sex differences are more likely to be found (Halpern et al., 2007).

Females outperform males on most tests of verbal ability, but not all. Verbal ability encompasses a wide variety of skills needed for language usage,

including language comprehension, vocabulary, word fluency, grammar, spelling, reading, and so on. Verbal ability has been assessed using a variety of tasks that ask the test taker to select the appropriate synonym or antonym for a group of words, solve verbal analogies, interpret complex reading passages, answer grammatical questions, and write essays. In general, women do better at these tasks than men do, especially when the assessment involves writing. The female writing advantage is seen as early as elementary school, but it is quite large by the end of secondary school (Hedges & Nowell, 1995). In a report by the U.S. Department of Education that included national data from standardized tests of writing ability from 1988 to 1996, females outperformed males in the fourth, eighth, and 11th grades (Bae, Choy, Geddes, Sable, & Snyder, 2000). Women also outperform men on the writing portion of the SAT by 13 points (male = 486, female = 499; The College Board, 2009). Additionally, females evidence a reading advantage over males, an effect that has consistently been seen internationally (Chiu & McBride-Chang, 2006; Mullis, Martin, Gonzalez, & Kennedy, 2003). The advantage that women typically evidence in verbal ability does not transfer to all portions of standardized tests. According to The College Board (2009), men slightly outperformed women on the critical reading portion of the SAT (male = 503, female = 498). This is a very small difference, but it is surprising given the general superiority of women on other tests of verbal ability. Additionally, men performed better on the SAT-Verbal (SAT-V) that included verbal analogies until 2004. Quantitative ability can be described as the specific skills needed to solve mathematics problems. For example, the skills included those needed to do simple arithmetic (add, subtract, multiple, divide), solve word problems, geometry, and calculus. Some researchers have found sex differences in quantitative giftedness in preschool children (Robinson, Abbott, Berninger, & Busse, 1996). More specifically, males tend to be overrepresented in the upper tails of the distribution of quantitative ability. This trend continues throughout the life span. There is more variability in the quantitative abilities of men than women. That is, there are more men in both the upper and lower tails of the quantitative ability distribution. The cause of sex differences in variability is unknown at this time. The size of the sex difference in tests of mathematics increases with the selectivity of the sample—that is, when the sample gets more and more selective, sex differences favoring males are more likely to be found. Internationally, there are few sex differences in mathematics when considering the mean score on tests that are administered in high school (Else-Quest, Hyde, & Linn, 2010). But, when mathematically gifted youth were tested, the ratio of boys to girls among the highest scoring students is between 3:1 and 4:1, and this has not changed over the past 20 years (Wai, Cacchio, Putallaz, & Makel, 2010).

Females earn higher mathematics grades than males in all grades (Association for University Women, 2010; Duckworth & Seligman, 2006; Kimball, 1989). This finding appears to directly contradict the findings that males have greater quantitative abilities than females, until you consider which sex differences are being measured. Females earn higher grades, which are the culmination of several factors, including ability, effort, behavior, and motivation. Duckworth and Seligman (2006) concluded that the female advantage in grades is caused, at least in part, by the finding that females have better self-discipline than males (in general) and thus perform better in school and get higher grades. They also found that females score higher on algebra assessments, which may be reflective of the language-components in algebra. Few sex differences are found in primary school when computational mathematics is learned. However, late in secondary school more mathematical reasoning and spatial skills are required to compute the higher level mathematics problems (e.g., in calculus and geometry) and sex differences become more pronounced. That is, men outperform women on the mathematics portion of the SAT by 35 points (The College Board, 2009).

Sex differences in quantitative ability decrease when the content of the assessment resembles what the students learn in class (Geary, 1996; Halpern, 2011). Compared with the mathematics portion of the SAT (SAT-M) and some international mathematics assessments, the assessment given by the National Center for Education Statistics, known as the National Assessment of Educational Progress

(NAEP), is more closely related to the curriculum. The results of the NAEP suggest that there are essentially no differences between the mathematics scores of males and females (Coley, 2001). The lack of sex differences found in the NAEP is presumably due to its closeness to the subject matter taught in schools and to the fact that the sample that takes these national assessments is not as select as for those who take the SATs. The recent revisions to the SAT were designed to make it more closely approximate what students learn in school, which should reduce the average differences between women and men on this assessment. The average difference on the SAT-M may be as large as it is because of the gender makeup of test takers. Many more women take the SAT than men (818,760 females and 711,368 males took the SAT in 2009; The College Board, 2009), which should result in a lower mean score because, assuming that most top-ability students of both sexes take college admissions exams, more women of lower ability take the SAT than men of lower ability. The differences in number of women and men who take the SAT means that any conclusions about sex differences based on SAT scores should be made with extreme caution. It seems that a number of conditions should lead us to interpret gender differences with caution, including self-selection, differences in variability, course-taking, and sample size.

Visuospatial ability also may influence scores on the quantitative portion of a standardized exam. Visuospatial ability also involves a set of various skills. Halpern and Collaer (2005) explained that visuospatial skill involves the generation, maintenance, transformation, and scanning of images as well as the interaction between the verbal, spatial, and pictorial aspects of mental representations. Sex differences favoring males in visuospatial skills can be reliably detected by 3 to 4 months of age (Moore & Johnson, 2008: Quinn & Liben, 2008) and are found in 53 countries (Lippa, 2010). Although most direct tests of visuospatial ability are inappropriate for a current discussion of assessment in higher education, visuospatial skills may influence quantitative ability. More specifically, the ability to mentally rotate images may be needed to compute higher level mathematics problems. Casey, Nuttall, Pezaris, and Bennow (1995) found that visuospatial ability mediated sex differences on the SAT-M. That is, when the effects of visuospatial ability were controlled for, the sex differences disappeared. Thus, females may be at a disadvantage on the math portion of the SAT because of their lower visuospatial skill.

A series of studies on the predictive validity of spatial skills show that spatial skills are important predictors of college major and occupation (Shea, Lubinski, & Benbow, 2001) and are associated with successful careers in science, technology, engineering, and mathematics (STEM fields; Wai, Lubinski, & Benbow, 2009). Although there is no college-level assessment of spatial skills, it seems that we may be missing an important cognitive dimension that would predict success, in at least some academic fields. It seems that spatial skills are a content-relevant component of the SAT-M, but a separate assessment would provide additional and important information to assess individuals success in fields with a high level of spatial tasks, such as engineering, some areas of mathematics, geography, surgery, and dentistry, among others.

Another factor that may influence scores on a quantitative ability assessment is stereotype threat, which has been shown to affect women's scores on mathematics exams (Walton & Spencer, 2009). The stereotype that women are not good at math may lead women to underperform on quantitative assessments when sex-based math stereotypes are activated. On the basis of the effect size found in the stereotype threat meta-analysis by Walton and Spencer (2009), stereotype threat depresses the abilities of women by 19 to 21 points. This accounts for a substantial proportion of the discrepancies between male and female SAT scores.

Current Trends in College Admission Assessments

Concerns about the validity and fairness of college admission assessments have lead to the exclusion or deemphasis of standardized test scores at many colleges and universities in the United States. The National Center for Fair and Open Testing (see http://www.fairtest.org/university/ACT-SAT) reported that 830 colleges and universities either do not require standardized test scores or deemphasize

the importance of such scores. This number has been challenged elsewhere (see Chapter 14, this volume), but what is certain is that the number of universities with test-optional policies is on the rise. Advocates of test-optional policies argue that test scores are a poor proxy for scholastic merit. Test-optional schools emphasize the poor predictive validity of standardized tests and the anxiety that they create for high school students. Additionally, test-optional policies alleviate concerns about the fairness of the assessments for women and non-Asian ethnic minorities. If there are so many criticisms to the use of standardized tests, why are they still being used?

Many college-ranking services use SAT or ACT scores as one indicator of the prestige of the college or university, and school administrators fear that by not requiring standardized test scores they will reduce their school's ranking. Lower ranking schools are less attractive to potential applicants, resulting in lost revenue and prestige and to fewer high-achieving student applicants. These marketing concerns may not be well founded. College and universities that have opted to exclude standardized tests from their admissions decisions report that the populations of their schools are becoming more diverse while maintaining their quality, and they are pleased with the outcome (see http://www.fairtest.org/university/ACT-SAT). It also has been suggested that colleges and universities with low average SAT scores are less desirable to top students, and that by not requiring the SAT, these institutions lose the disadvantage associated with low SAT scores.

Graduate School Entrance Exams

Many of the same validity concerns that have been discussed with regard to undergraduate entrance exams apply to graduate school entrance exams such as the GRE, the Law School Admissions Test (LSAT), the Graduate Management Admissions Test (GMAT), the Medical College Admissions Test (MCAT), and the Pharmacology College Admissions Test (PCAT), among others. A meta-analysis including 3 to 1,231 studies and 244 to 259,640 students revealed four promising findings regarding standardized tests of graduate school admissions (Kuncel & Hezlett, 2007). First, the standardized exam scores including scores on the GRE, LSAT, GMAT, Miller Analogies Test (MAT), MCAT, and PCAT predicted performance in graduate school as measured by a variety of measures, including 1st-year GPA (corrected correlational values ranges from .41 to .59), GPA at graduation, faculty ratings, and more. Second, both standardized exam scores and undergraduate grades predicted graduate school performance beyond graduate school GPA alone. Third, standardized exam scores predicted most of the performance measures better than undergraduate GPA. Fourth, the combination of standardized exam scores and grades resulted in the most accurate prediction of performance in graduate school. Interestingly, there was generally no evidence of group differences in the predictive validity of the standardized graduate school entrance exams based on gender or ethnicity, and in situations in which differences were found, they favored ethnic minorities. Another meta-analysis of more than 100 studies conducted by Kuncel, Wee, Serafin, and Hezlet (2010) indicated that there were no differences in the predictive validity of the GRE based on whether the graduate students were masters- or doctoral-level students. Although the adjusted correlational values appear to be small to medium, ranging from .21 to .38, the range of students taking the GRE is restricted, which would decrease the value of the correlation. Thus, research findings on the predictive validity of standardized graduate school admissions exams are consistent and promising.

USING ASSESSMENTS TO MEASURE STUDENT LEARNING OUTCOMES

In the preceding section we discussed how assessments can be used as a gateway to higher education, but assessment does not end with admission into college. Student learning needs to be assessed during the college experience and beyond, with the goal of improving teaching and learning (Halpern, 2004). The assessment of student learning outcomes should be, first are foremost, student centered (Halpern, 2004). Allen (2004) has outlined the process of student learning assessment and has indicated the following six steps to her model: (a) develop learning objectives, (b) check for alignment between the

curriculum and the objectives, (c) develop an assessment plan, (d) collect assessment data, (e) use results to improve the program, and (f) routinely examine the assessment process and correct as needed. Thus, before we can discuss the best practices for assessing student learning, we should identify what student should know and be able to do upon graduation.

What Are Our Learning Goals and Objectives?

We include here an example of the learning goals and objectives for psychology majors. The American Psychological Association's (APA's; 2007) Task Force on Psychology Major Competencies identified 10 learning goals for psychology majors, including (a) a general knowledge base of psychology, (b) research methods in psychology, (c) critical-thinking skills in psychology, (d) the application of psychology, (e) values in psychology, (f) information and technological literacy, (g) communication skills, (h) sociocultural and international awareness, (i) personal development, and (j) career planning and development (see APA, 2007, for more detailed information). Similar goals have been developed for most other majors.

How Do We Know Whether Students Are Meeting These Goals and Objectives?

Most faculty members believe they are educating their students. In fact, evidence suggests that most college professors believe they are better educators than the average faculty member (Cross, 1977). This effect is known as the *better-than-average effect* (a.k.a. *illusory superiority*, the *above-average effect*, or the *Lake Wobegon effect*). It occurs in a wide variety of situations with a wide variety of abilities or characteristics, such as the perceptions of intelligence, scores on the SAT relative to peers (Alicke & Goorun, 2005), academic and job performance, desirable personality characteristics (Hoorens, 1993), and popularity ratings (Zuckerman & Jost, 2001). In Cross's (1977) survey of the faculty at the University of Nebraska, 68% reported that they were above-average professors, but how do they *know* that? How do we *know* whether our students are learning? When asked this question, faculty members from colleges and universities around the world responded with either blank stares or anecdotal stories of exceptional students. The problem with anecdotes is that those exceptional students would probably have succeeded at any college. What about the other 99% of our students? Anecdotes are not evidence and as an evidence-based, data-driven field, we need to be more systematic about the assessment of student learning outcomes. Furthermore, as Maki (2002) pointed out, much of what we do as educators and researchers is based on intellectual curiosity, so why *wouldn't* we be curious about whether our students are learning and what aspects of learning could be improved? Nevertheless, the thought of assessing student learning outcomes is terrifying for some faculty members and considered a nuisance by others. Some faculty may view the assessment of student learning as a criticism of their teaching abilities. Departments should work together to articulate the goals of the assessment as well as the policies regarding how that information should be used and shared.

The assessment of student learning is not easy, and many faculty members question why course grades are not enough evidence that learning has occurred. Transcripts are certainly one indication that learning has occurred, but grade inflation has made grades fairly meaningless as an assessment of learning. Grades do not indicate whether that learning has persisted over time, and letter grades (or GPA) are too broad a variable to determine which learning outcomes were mastered. Let us consider two students who received a C in research methods. Did these students learn enough to earn the C, or did the professor feel obligated to give them a C because they exerted so much effort? Will these students be able to design, conduct, and interpret their own studies in the future? Do they understand the difference between correlational research and an experiment? Can they conduct research ethically? A letter grade cannot answer these more interesting and perhaps more relevant questions. Furthermore, Allen (2004) pointed out that course grades are summative and provide no opportunity for improvement, whereas the assessment of student learning outcomes can be formative and thus provide educators with important feedback with which to improve

student learning. Thus, amid the pessimism surrounding student learning outcome assessment is the excited optimism of involved educators.

In 2009, the APA published the second edition of *The Assessment Cyberguide for Learning Goals and Outcomes*. *The Assessment Cyberguide* is divided into four sections, including (a) understanding assessment (departmental, institutional, educational, and societal perspectives), (b) designing viable assessments plans, (c) sustaining an assessment culture, and (d) applying assessment strategies in psychology. Included in *The Assessment Cyberguide* is one of the APA Board of Educational Affairs Task Force on Psychology Major Competencies' top 10 recommendations for best practices in assessment (for more detailed information on the best practices, see APA, 2009). Briefly, these research-based recommendations for student learning outcome assessment include (a) encouraging department ownership to drive the process; (b) defining objectives in the context of each institutional mission; (c) focusing on collaboration and teamwork; (d) clarifying the purpose of assessment; (e) identifying clear, measurable, and developmental student learning; (f) using multiple measures and sources consistent with resources; (g) implementing continuous assessment with clear, manageable timelines; (h) helping students succeed on assessment tasks; (i) interpreting and using assessment results appropriately; and (j) evaluating your assessment practices.

There are many continua on which students learning assessments can be categorized. For example, an assessment can be direct or indirect, traditional or performance based, quantitative or qualitative, value added (did they improve?) or absolute (did they met a specified criteria?), formative (provides feedback) or summative (descriptive only), authentic (completing a real task), developmental (a series of steps, hurdles, or tasks that must be completed in sequence), or even embedded within a course (Allen, 2004). The bulk of *The Assessment Cyberguide* (APA, 2009) examines the relative advantages and disadvantages of each assessment strategy (or potential source of assessment data) based on the learning outcome of interest (e.g., critical thinking, research methods) to suggest an optimal method for assessing each learning outcome, objective, or goal. A wide variety of assessment strategies were considered including course data (objective tests, essay tests, embedded questions, or assignments), individual projects and performance assessments (written products such as term papers, lab reports or critiques, oral presentations, graphic tests and displays, poster presentations, structural or situational assessments), summative performance assessments (standardized tests, locally developed exams, capstone experiences, internships or professional applications, portfolios, assessments center methods such as in-baskets and guided problem solving), self-assessment and reflection (student journals of self-critique), collaboration (research teams, groups projects, online group activities), interviews and surveys (attitude measurement using satisfaction measures from seniors, alumni, employers, graduate school advisors, or parents; performance reviews from alumni, employers, or graduate school advisors; exit interviews, focus groups, follow-up alumni interviews, and external examiner interviews), and archival measures (transcript analysis, analysis of transfer patterns, syllabus audit, demographic data analysis, alumni database, library use statistics, or website hits).

Other Trends in the Assessment of Student Learning Outcomes

The use of portfolios in student learning outcome assessments is becoming increasingly popular (cf. Allen, 2004, for examples). A portfolio is a collection of work a student has done. It is typically accompanied by a reflective essay that requires students to think about their academic experiences. Portfolios can either be developmental in that they can demonstrate how the student has progressed, or they can showcase the student's best pieces of work. Additionally, portfolios can be used within one course or could be completed as a capstone project that reflects students' work during their entire academic career. According to Allen (2004), portfolios encourage students to take part in the assessment process and claim ownership over their education and learning processes. There is also a growing body of evidence that completing portfolios improves metacognition (Meyer, Abrami, Wade, Aslan, & Deault, 2010; Scott, 2010). Empirical research that

empirically evaluates the validity of portfolios is scant at this time and has been plagued with problems of implementation and grading reliability (Holland, 2007). Regardless of the lack of empirical validation, some colleges require that student portfolios be evaluated and approved before students are allowed to graduate (New Century College, 2002). Furthermore, in some parts of the United States, school principals expect teaching position applicants to bring their teaching portfolios with them when interviewed (Allen, 2004). Many colleges and universities are considering the utility of a digital portfolio or webfolio.

Postgraduation surveys of alumni and their employers are increasingly popular forms of assessment. Colleges and universities can potentially determine how well their students are prepared for the real world based on how successful their graduates are in finding employment, how well they are being paid, and the prestige of their position. Although using terms such as "success" is problematic, colleges and universities can use such means to gauge how well their students believe they were prepared for life after graduation. College graduates who have obtained employment can provide a wealth of information about their satisfaction with their employment and how well their college or university prepared them for their current position. Additionally, employer satisfaction surveys can be especially probative to the issue of whether college graduates are prepared for the workforce. Many question whether 21st-century graduates will be prepared for tomorrow's workforce (Association of American Colleges and Universities [AAC&U], 2010; Hunt, 1995; National Science Board, 2005). In a recent employer survey conducted by the Association of American Colleges and Universities, 63% of employers reported that recent college graduates were not prepared enough to compete in a global market and lacked the essential skills that they need to do so (AAC&U, 2010). Employers want colleges and universities to place greater emphasis on teaching a variety of skills, including written or oral communication (89% of employers) and critical-thinking skills (81% of employers). The assessment of employer perceptions of alumni is a valuable tool for departments to use to meet the demands of an ever-changing workplace.

Employers believe that colleges and universities should place greater emphasis on teaching critical-thinking skills (AAC&U, 2010). Critical thinking is not only a learned skill but also a disposition (Halpern, 1998, 2003). Critical thinking is a desirable skill or disposition to have because critical thinkers are more flexible, more willfully process information, make more informed decisions, and are more persistent than noncritical thinkers. Critical thinking can be learned (for reviews, see Chance, 1986; Halpern, 2003; Moseley et al., 2005; Nisbett, 1992) and is currently being taught directly in the form of critical-thinking and problem-solving courses as well as indirectly in other courses. There are a variety of known clearly identifiable critical-thinking skills, and one of the greatest challenges of a critical thinker is to identify which skill is appropriate for a given situation. Critical thinking can be assessed by colleges and universities or by employers. The Halpern Critical Thinking Assessment (HCTA; Halpern, 2010) measures a variety of skills, including (a) verbal reasoning skills, (b) argument analysis skills, (c) skills in thinking as hypothesis testing, (d) using likelihood and uncertainty, and (e) decision-making and problem-solving skills. Many of these skills were identified by employers as desirable skills that colleges and universities should emphasize. The HCTA has the added benefit of having face validity and being easy to communicate to non-academics.

Recently, at least one graduate school admission exam, the GMAT, has integrated an assessment of reasoning. The GMAT is the standardized exam for those pursuing a graduate degree in business. The inclusion of the "integrated reasoning" section is meant to examine the student's ability to analyze information. Students will be presented with a table or spreadsheet of information and asked questions about it. The new section will replace one of the essay questions in the 2012 administration. This change in the exam reflects a general trend toward the inclusion of critical-thinking skills in assessments, but it remains problematic as a measure of learning outcomes because it is designed to select students for competitive programs in business.

Ultimately, the goal of any student-learning assessment should be for transfer. That is, educators should teach for transfer to new situations and

assess whether students can exhibit evidence of transfer. It will be of no benefit to our students to learn what critical-thinking skills are, if they cannot use them in the real world to make more informed decisions. It is of no use for a student to recite Piaget's stages of development, if it does not make them a better parents or caretaker.

Assessment is difficult and fraught with pitfalls. Educators must answer the challenge of the public for more accountability in higher education with empirically derived answers. As long as the stakeholders remember that the purpose of assessing student learning outcomes is to improve teaching and learning, who can argue against that?

References

Alicke, M. D., & Goorun, O. (2005). The better-than-average effect. In M. D. Alicke, D. A. Dunning, & J. I. Krueger (Eds.), *The self in social judgment* (pp. 85–106). New York, NY: Psychology Press.

Allen, M. J. (2004). *Assessing academic programs in higher education.* San Francisco, CA: Jossey-Bass.

American Psychological Association. (2007). *APA guidelines for the undergraduate psychology major.* Washington, DC: Author. Retrieved from http://www.apa.org/ed/precollege/about/psymajor-guidelines.pdf

American Psychological Association. (2009). *The assessment cyberguide for learning goals and outcomes.* Washington, DC: Author. Retrieved from http://www.apa.org/ed/governance/bea/assessment-cyberguide-v2.pdf

Association for University Women. (2010). *Why so few?* Washington, DC: Author. Retrieved from http://www.aauw.org/learn/research/upload/whysofew.pdf

Association of American Colleges & Universities. (2010). *Raising the bar: Employers' views on college learning in the wake of the economic downturn.* Retrieved from http://www.aacu.org/leap

Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher, 38*, 665–676. doi:10.3102/0013189X09351981

Bae, Y., Choy, S., Geddes, C. Sable, J., & Snyder, T. (2000). *Educational equity for girls and women* (NCES 2000–030). National Center for Education Statistics, U.S. Department of Education. Washington, DC: U.S. Government Printing Office. Retrieved from http://nces.ed.gov/pubs2000/2000030.pdf

Berry, C. M., & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of the validity of college admissions systems. *Psychological Science, 20*, 822–830. doi:10.1111/j.1467-9280.2009.02368.x

Blackhart, G. C., Peruche, B. M., DeWall, C. N., & Joiner, T. E., Jr. (2006). Factors influencing teaching evaluations in higher education. *Teaching of Psychology, 33*, 37–63. doi:10.1207/s15328023top3301_9

Bridgeman, B., Pollack, J., & Burton, N. (2004). *Understanding what SAT reasoning test scores add to high school grades: A straightforward approach* (Research Report No. 2004-4). New York, NY: The College Board.

Carnevale, A. P., Smith, N., & Strohl, J. (2010). *Help wanted: Projections of jobs and education requirements through 2018.* Georgetown Center on Education and the Workforce. Retrieved from http://cew.georgetown.edu/JOBS2018

Casey, M. B., Nuttall, R., Pezaris, E., & Bennow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31*, 697–705. doi:10.1037/0012-1649.31.4.697

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivations. *Journal of Applied Psychology, 82*, 300–310. doi:10.1037/0021-9010.82.2.300

Chance, P. (1986). *Thinking in the classroom: A survey of programs.* New York, NY: Teachers College.

Chiu, M. M., & McBride-Chang, C. (2006). Gender, context, and reading: A comparison of students in 43 countries. *Scientific Studies of Reading, 10*, 331–362. doi:10.1207/s1532799xssr1004_1

Coley, R. (2001). *Differences in the gender gap: Comparisons across racial/ethnic groups in education and work.* Princeton, NJ: Educational Testing Service, Policy Information Center. Retrieved from http://www.ets.org/research/pic

College Board. (2009). *Validity of the SAT for predicting FYGPA: 2007 SAT validity sample* (Statistical Report No. 2009-1). Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/statisticalreport-2009-1-validity-sat-1st-yr-gpa-2007-sample.pdf

College Board. (2010). *2010 College-bound seniors: Total group profile.* Retrieved from http://professionals.collegeboard.com/profdownload/2010-total-group-profile-report-cbs.pdf

Cross, P. (1977). Not can but will college teachers be improved? *New Directions for Higher Education, 17*, 1–15. doi:10.1002/he.36919771703

de Vries, L. (2003, January 27). *High school grades hit by inflation.* Retrieved from http://www.cbsnews.com/stories/2003/01/27/national/main538000.shtml

Duckworth, A., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98*, 198–208. doi:10.1037/0022-0663.98.1.198

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*, 103–127. doi:10.1037/a0018053

Freedle, R. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, *73*(1), 1–43.

Geary, D. C. (1996). International differences in mathematical achievement: Their nature, causes, and consequences. *Current Directions in Psychological Science, 5*, 133–137. doi:10.1111/1467-8721.ep11512344

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*, 449–455. doi:10.1037/0003-066X.53.4.449

Halpern, D. F. (2003). *Thought and knowledge: An introduction to critical thinking* (4th ed.). Mahwah, NJ: Erlbaum.

Halpern, D. F. (2004). Creating cooperative learning environments. In B. Perlman, L. I. McCann, & S. H. McFadden (Eds.), *Lessons learned: Vol. 2. Practical advice for the teaching of psychology* (pp. 165–173). Cambridge, MA: Cambridge University Press.

Halpern, D. F. (2011). *Sex differences in cognitive abilities* (4th ed.). New York, NY: Psychology Press.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8*, 1–51. doi:10.1111/j.1529-1006.2007.00032.x

Halpern, D. F., & Collaer, M. L. (2005). Sex differences in visuospatial abilities: More than meets the eye. In A. Miyake & P. Shah (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 170–212). New York, NY: Cambridge University Press.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, *269*, 41–45. doi:10.1126/science.7604277

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardization cognitive ability testing? *American Psychologist, 47*, 1083–1101. doi:0OO3-O66X/92

Helms, J. E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist, 61*, 845–859. doi:10.1037/0003-066X.61.8.845

Higher Education Research Institute. (1999). *The American freshman: National norms for fall 1999*. Retrieved from http://www.heri.ucla.edu/pr-display.php?prQry=21

Holland, R. (2007). *Portfolios: A backward step in school accountability*. Lexington Institute. Retrieved from http://www.lexingtoninstitute.org/portfolios-a-backward-step-in-school-accountability

Horn, L., Nevill, S., & Griffith, J. (2006, June). *Profile of undergraduates in U.S. postsecondary institutions: 2003–04, with a special analysis of community college students*. U.S. Department of Education, Institute of Educational Science, National Center for Education Statistics. NCES 2006–184. Retrieved from http://nces.ed.gov/pubs2006/2006184_rev.pdf

Hunt, E. (1995). *Will we be smart enough? A cognitive analysis of the coming workforce*. New York, NY: Russell Sage Foundation.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*, 581–592. doi:10.1037/0003-066X.60.6.581

Jaschik, S. (2010, June 21). *New evidence of racial bias on SAT*. Retrieved from http://www.insidehighered.com/news/2010/06/21/sat

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*, 60–99. doi:10.1016/0010-0285(89)90003-0

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin*, *105*, 198–214. doi:10.1037/0033-2909.105.2.198

Kobrin, J. L., Camara, W. J., & Milewski, G. B. (2002). *The utility of the SAT I and SAT II for admissions decisions in California and the Nation* (Report No. 2002-6). New York, NY: The College Board.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point averages* (Research Report No. 2008-5). New York, NY: The College Board.

Kohn, A. (2001, March 9). Two cheers for an end to the SAT. *Chronicle of Higher Education*, p. B12.

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science, 315*, 1080–1081. doi:10.1126/science.1136618

Kuncel, N. R., Wee, S., Serafin, L., & Hezlet, S. A. (2010). The validity of the graduate record examination for Master's and Doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement, 70*, 340–352. doi:10.1177/0013164409344508

Lippa, R. A. (2010). Sex differences in personality traits and gender-related occupational preferences across 53 nations: Testing evolutionary and social-environmental theories. *Archives of Sexual Behavior, 39*, 619–636. doi:10.1007/s10508-008-9380-7

Maki, P. (2002, May). Moving from paperwork to pedagogy *AAHE Bulletin*. Retrieved from http://www.aahebulletin.com/public/archive/paperwork.asp

Meyer, E., Abrami, P. C., Wade, C. A., Aslan, O., & Deault, L. (2010). Improving literacy and metacognition with electronic portfolios: Teaching and learning with ePEARL. *Computers and Education, 55*, 84–91. doi:10.1016/j.compedu.2009.12.005

Moore, D. S., & Johnson, S. P. (2008). Mental rotation in human infants: A sex difference. *Psychological Science, 19*, 1063–1066. doi:10.1111/j.1467-9280.2008.02200.x

Moseley, D., Baumfield, V., Elliott, J., Gregson, M., Higgins, S., Miller, J., & Newton, D. P. (2005). *Frameworks for thinking: A handbook for teaching and learning*. Cambridge, MA: Cambridge University Press. doi:10.1017/CBO9780511489914

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001: International report*. Boston, MA: Boston College.

National Science Board. (2005). *2020 vision for the National Science Foundation*. Retrieved from http://www.nsf.gov/pubs/2006/nsb05142/nsb05142.pdf

New Century College. (2002). *Integrative studies graduation portfolio*. Retrieved from http://ncc.gmu.edu/graduation/grad_portfolio.html

Nisbett, R. E. (1992). *Rules for reasoning*. Hillsdale, NJ: Erlbaum.

Quinn, P. C., & Liben, L. S. (2008). A sex difference in mental rotation in young infants. *Psychological Science, 19*, 1067–1070. doi:10.1111/j.1467-9280.2008.02201.x

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language and ethnic groups* (ETS RR No. 94–27). The College Entrance Examination Board. New York, NY: The College Board.

Robinson, N. M., Abbott, R. D., Berninger, V. W., & Busse, J. (1996). Structure of abilities in math-precocious young children: Gender similarities and differences. *Journal of Educational Psychology, 88*, 341–352. doi:10.1037/0022-0663.88.2.341

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*, 215–227. doi:10.1037/0003-066X.63.4.215

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin*, *135*, 1–22. doi:10.1037/a0013978

Santelices, M. V., & Wilson, M. (2010). Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review, 80*, 106–133.

Sawyer, R. (2010). *Usefulness of high school average and ACT scores in making college admissions decisions* (ACT Research Report Series Report No. 2010–2). Retrieved from http://www.act.org/research/researchers/reports/pdf/ACT_RR2010-2.pdf

Schmitt, N., Keeney, J., Oswald, F., Pleske, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college performance using cognitive and non-cognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology, 94*, 1479–1497. doi:10.1037/a0016810

Scott, S. G. (2010). Enhancing reflection skills through learning portfolios: An empirical test. *Journal of Management Education, 34*, 430–457. doi:10.1177/1052562909351144

Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology, 93*, 604–614. doi:10.1037/0022-0663.93.3.604

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures*. Retrieved from http://www.siop.org/_Principles/principlesdefault.aspx

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811. doi:10.1037/0022-3514.69.5.797

Sternberg, R. J. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence, 34*, 321–350. doi:10.1016/j.intell.2006.01.002

Sternberg, R. J. (2009). The Rainbow and Kaleidoscope Projects: A new psychological approach to undergraduate admissions. *European Psychologist, 14*, 279–287. doi:10.1027/1016-9040.14.4.279

Wagerman, S. A., & Funder, D. C. (2007). Acquaintance reports of personality and academic achievement: A case for conscientiousness. *Journal of Research in Personality, 41*, 221–229. doi:10.1016/j.jrp.2006.03.001

Wai, J., Cacchio, M., Putallaz, M., & Makel, M. C. (2010). Sex differences in the right tail of cognitive abilities: A 30-year examination. *Intelligence, 38*, 412–423. doi:10.1016/j.intell.2010.04.006

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101*, 817–835. doi:10.1037/a0016127

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science, 20*, 1132–1139. doi:10.1111/j.1467-9280.2009.02417.x

Zuckerman, E. W., & Jost, J. T. (2001). What makes you think you're so popular? Self evaluation maintenance and the subjective side of the friendship paradox. *Social Psychology Quarterly, 64*, 207–223. doi:10.2307/3090112

CHAPTER 16

ACHIEVEMENT TESTING IN K–12 EDUCATION

Carina McCormick

Achievement tests are tests used to measure students' academic skills and knowledge. Achievement testing in K–12 education is especially used to measure mastery of the educational jurisdiction's intended curricula or overarching educational objectives. One historical text (Ebel, 1965) defined an achievement test as "one designed to measure a student's grasp of some body of knowledge or his proficiency in certain skills" and distinguished these tests from aptitude tests "given to determine the potential of an individual for development along a special line" (p. 445). Nunnally (1972) placed achievement tests in the school and temporal context: "The purpose of achievement tests is to measure progress in school up to a particular point in time" (p. 234). Current texts follow similar definitions, specifying that the purpose of achievement tests is to measure individual's knowledge and skills in a particular content area.

Because of this broad definition, a wide variety of tests including classroom assessments to norm-referenced tests, summative state tests, and diagnostic formative assessments can all be considered achievement tests. Although each of these test purposes is unique, they share the common goal of accurately estimating students' level of knowledge and skill in a specified domain. Because achievement tests by definition assess students' content mastery, the intended content and scope of the assessment must be clearly laid out in advance and closely followed in test construction. It is necessary to delineate this type of test from tests of intelligence or aptitude and tests designed to predict future performance because each differs in purpose and construction. However, Thorndike (2005) explained that the difference between achievement and aptitude tests can become blurred and "often lies more in the purpose for which the test results are used than in the nature or content of the test itself" (p. 62). Achievement tests provide descriptive information about students' skill levels, and an underlying premise of their use is that the information can be used to some benefit.

The general issues of validity, reliability, test fairness, and psychometrics discussed elsewhere in this handbook also apply to achievement testing in K–12 education but have certain points that are especially applicable to this type of test. Although classroom assessments certainly are classified as achievement tests, this chapter focuses instead on issues more pertinent to standardized achievement testing. In this chapter, a wide variety of topics under the general umbrella of K–12 achievement testing is introduced. This includes an overview of the important points relating to each topic, especially focusing on considerations for decision makers at various levels. Some important topics are covered by dedicated chapters, and many are cross-referenced in this chapter rather than duplicated.

A clear test purpose is essential to appropriate score interpretation, as is the match between the test's purpose, design, and validation evidence. The chapter begins with some cautions about inappropriate use of achievement tests for a wide variety of purposes. This discussion continues in the following sections by highlighting challenges associated with different achievement test uses. A distinguishing feature of much of current achievement testing in K–12 education is its

DOI: 10.1037/14049-016
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

inextricable tie to federal requirements. Decisions made at the federal level largely influence the practice of measurement in education, and measurement professionals work to simultaneously fulfill the requirements and promote best practices. Therefore, to understand the use of such tests, it is necessary to have some background about the current legislation and how it relates to recent developments in assessment. As part of these changes, the idea of formative and interim assessment has earned greater prominence, and these types of tests and test use are described. In addition to large-scale tests used as part of accountability systems, this chapter briefly describes four other types of achievement tests applicable to K–12 education: norm-referenced assessments, alternate assessment, high school graduation exams, and group-level assessments such as the National Assessment of Educational Progress (NAEP). Achievement tests are used in education as indicators of what students know and can do at a given time (status) and of how much their knowledge and skills have increased over time (growth). Two major issues relating to these purposes are explained: instructional sensitivity and measurement of growth. The chapter concludes with brief summary of how some technical issues—scaling, item response theory, and computerized adaptive assessment—are important in large-scale achievement testing.

Measurement in public education always involves tough choices and a delicate balance of priorities. There will never be enough money to build and implement the test of measurement professionals' dreams. Therefore, it is important to understand many of the challenges of achievement testing in K–12 education. This chapter is intended to assist those involved with these testing programs now and in the future—whether as decision makers, researchers, administrators, measurement professionals, or educators—in supporting high-quality assessment systems that inform change and monitor progress toward the goal of excellent education for all students.

VALIDITY EVIDENCE FOR ACHIEVEMENT TEST USE

When developing, implementing, or using any assessment, it is critical to have a clear idea of the assessment's intended purposes. Achievement tests in K–12 education have been used in many ways, probably for a wider range of purposes than many other tests. Overall, however, the general purpose of achievement testing in K–12 education is to identify students' level of skills and knowledge, typically toward the goal of improving education for individual students or whole groups of students. However, the step between simply providing information and serving as an actionable tool for making change is a large and complicated one. It is critical that the assessment foundation supports the decisions and actions to be made. This foundation is not simply building a "good test" but rather ensuring that the test is appropriately designed for the purposes to which it will be put. This step is an essential component for the validity of scores. Kane (2002) gave an example of a bathroom scale being an excellent assessment of weight, but the use of a person's weight in certain decisions, such as employment, would likely be considered inappropriate. (For a discussion of current validation theory, see Volume 1, Chapter 4, this handbook, and for a discussion of standards for appropriate test use, see Volume 1, Chapter 13, this handbook.)

Kane (2002) wrote that when tests are used to make high-stakes decisions, each inference should be clearly examined in evaluating the interpretive argument. Often, for high-stakes achievement tests in education, the interpretive argument goes beyond inferences about student skill mastery to inferences about positive benefits of the assessment. According to Kane, certain policy assumptions underlie decision-based score interpretations, such as issuing sanctions to schools or withholding a student's diploma, much more than simple descriptive score interpretations about students' level of performance on the tested concepts. Kane argued for more thorough and deliberate evaluation of the claims made for the high-stakes use of tests:

> In fact, if the primary purpose of a testing program is to promote certain outcomes, rather than simply to estimate certain variables, the testing program is functioning as an educational intervention, and therefore merits evaluation of

> the kind routinely mandated for any new educational program. (p. 40)

Bandalos, Ferster, Davis, and Samuelsen (2011) listed four common goals of achievement tests and for each, outlined the underlying assumptions, types of evidence available for evaluation, and possible unintended negative consequences.

Often, especially with end-of-year tests used for accountability, a single high-quality test is used for many purposes. There is a tendency to expect too much from such tests, ranging from diagnostic-level student information to making decisions about the quality of teachers and schools. Sometimes, uses of the test are added after the testing program is in place. This practice is especially troublesome because the test had been designed for specific purposes and may not support the new uses appropriately. According to Braun (2009), "measuring a student's academic achievement is more complex and more subtle than using a scale to measure her height or weight. The complexities only increase when test scores are aggregated to make evaluative judgments about schools" (p. 53). This chapter presents further information about specific achievement test uses.

Test publishers may make many claims about the value and multiple uses of the test. For example, the website for the Stanford Achievement Test, 10th edition, claims that

> administrators obtain critical data to document and monitor the progress of all children and to disaggregate results according to federal mandates. Teachers receive specific information to support instructional planning for individual students and the class as well as to improve their teaching. Parents better understand their child's achievement level and get direction for home involvement. (Pearson Education, Inc., 2011)

Each of these test purposes should be evaluated separately (Kane, 2006). In generally, test users should be wary of claimed test purposes until reviewing evidence supporting those claims. The burden of determining whether validity evidence supports all or any of these uses is shared between the test publisher and the test user (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999).

Sometimes, the test is not appropriate or reliable for making certain decisions about all students being assessed. For example, a test to identify students' reading levels across a wide range of abilities should accurately measure the specific skills that distinguish fluent readers from struggling readers. Tests that are designed to measure grade-level standards may be too advanced for the students who most need remediation and thus cannot accurately pinpoint their true reading level. Especially when such a test is used to make remediation decisions, the reliability of the test needs to be high, specifically at the portion of the scale corresponding to the cut scores. Thus, tests designed for other purposes are likely to produce misleading results when used at extremes of the score scale.

Another common mismatch between desired use and intended use relates to the reporting of subscores. For example, many statewide end-of-year tests are designed to produce accurate overall scores for each student and for groups of students, and for this purpose, they are usually largely successful. Teachers, parents, and administrators, however, often desire more specific scores about areas of student strengths and weaknesses in the form of subscores. The term *subscore* refers to a finer grain score than overall score on a test, such as by objectives in the content standards. Test makers are under much pressure to create tests that can be administered in a relatively short time period and that accurately measure a wide range of student skill levels, so there is a limit to how many subtests can produce scores reliable enough to trust, especially at the student level. As the demand for a larger number of subscores increases, the reliability of each score will almost definitely decrease unless the length of the test increases correspondingly. Subscores typically have more error than overall scores and often do not do a better job of representing student skills in the subdomain than the overall score would (Sinharay, Puhan, & Haberman, 2010). Simply adding a cautionary note on score reports that subtest scores

may have low reliability is not enough to guarantee that these scores will not be misused. Overinterpretation of unreliable subscores can have negative consequences for students and teachers.

STATE TESTS FOR ACCOUNTABILITY

The Elementary and Secondary Education Act (ESEA) is wide ranging and prescribes many aspects of education, especially as they relate to equal rights and federal funding. In 2001, ESEA was reauthorized under the leadership of George W. Bush with the aspirational subtitle No Child Left Behind (NCLB; 2002). Koretz and Hamilton (2006) thoroughly explained major issues facing testing programs under NCLB. Since NCLB became law, it has faced many criticisms about some provisions and other changes in education. The law is currently behind schedule for reauthorization, leading to some challenges for states, testing professionals, and politicians. As a result, states have continued to conduct their testing programs consistent with NCLB, although states are free to add extra components. The specific requirements are expected to change with ESEA reauthorization.

Central to NCLB is a state assessment program tied to state curricular standards. Each state was required to implement its own set of rigorous content standards that would be applied throughout the state (NCLB, 2001). Although this provision allowed states to have control over what was taught in their states, it precluded the use of shared assessments unless states shared content standards. One group of states in particular (Rhode Island, Vermont, Virginia, and recently Maine) formed a partnership that allowed the states to administer the same assessment (the New England Common Assessment Program) by jointly implementing the same content standards. Currently, states are required to administer a summative assessment to students (at least) once each year in Grades 3 through 8 for reading and mathematics, twice overall in Grades 3 through 8 for science, and once in high school for reading, mathematics, and science. These tests must have at least three proficiency levels: one indicating proficiency or sufficient mastery of standards, one for performance below expectations, and one for performance that exceeds expectations (NCLB, 2001; for more information on how proficiency levels are determined through a formal standard-setting process, see Chapter 22, this volume).

With many different tests assessing many different state standards, it is clear that the performance expectations for proficiency classification are not comparable from state to state. Either the cut scores could have been set differently, or the curricular standards themselves could represent different expectations for student learning, which the tests and cut scores accurately reflected. Linn (2005) concluded that holding states responsible for student proficiency but "leaving the definition of proficient achievement to the states has resulted in so much state-to-state variability in the level of achievement required to meet the proficient standard that 'proficient' has become a meaningless designation" (p. 14). Despite these differences, states have been evaluated based on what percent of students were proficient on the basis of each state's own defined levels of proficiency. Clearly, this lack of uniform standards has been a concern, especially for individuals within certain states who believe that their high expectations for students unfairly penalize the schools and districts within the state when they are evaluated based on student performance on the state test.

The percentage of students who are proficient is a limited snapshot of a state's educational quality. This particular indicator is also susceptible to inconsistencies in classification. Readers will likely recall from introductory statistics that in a normal distribution, scores are generally most highly concentrated near the mean. If the proficiency cut score is located near the mean of the distribution, that is, close to where the highest number of students' scores are clustered, then the likelihood of misclassifications increases (Ho, 2008). When weighty decisions are made on this single indicator, increases or decreases in percent proficient can occur based on the precision of student scores used to make classification decisions. Administrators may not mind when the proficiency rate has a drastic increase from the year before, but they may have a harder time explaining sudden decreases. Foley (2011) found that a state was more likely to experience a decrease in percent

proficient when the previous year's percent proficient was high, suggesting that instability in the statistic influenced the figures reported. Plake (2011) wrote that for students whose scores are near the cut scores, "changes in across year performance-level classifications will likely be due more to errors in measurement than to the educational program these students receive" (p. 19). Ho (2008) encouraged greater use of the mean to represent more fully student performance, and he also emphasized the importance of using multiple measures rather than a single achievement test.

Many have argued that the lack of a rigorous national curriculum and national assessment makes reform and improvement especially challenging. For example, on the basis of an analysis of what is now called the Trends in International Mathematics and Science Study (TIMSS), Schmidt, Houang, and Cogan (2002) concluded, "American students and teachers are greatly disadvantaged by our country's lack of a common, coherent curriculum and the texts, materials, and training that match it" (p. 10). Recognizing the limitations of discrepant state curriculum standards and assessments, there has been a push for a unified, comprehensive set of standards that are linked to research about effective education and the needs of the 21st century. As a response to that need, the Common Core State Standards were developed and evaluated through a collaborative and iterative effort involving many stakeholders and experts (National Governors Association & Council of Chief State School Officers, 2010).

Porter, McMaken, Hwang, and Yang (2011) used alignment methodology to compare the Common Core State Standards and found that the cognitive complexity of the new common standards were generally higher than the typical state standards. A complaint about many state standards, compared with those of other countries, has been that they included too many skills, that those skills were repeated for too many grades, and that the learning progression was not coherent (e.g., Schmidt et al., 2002). This problem was deliberately avoided in the new standards: "A particular standard was included in the document only when the best available evidence indicated that its mastery was essential for college and career readiness in a twenty-first-century, globally competitive society" (Common Core State Standards Initiative [CCSSI], 2010, p. 3).

The Race to the Top competition has begun to change assessment and the curricular standards that the assessments will measure. The optional competition pitted states against each other to develop comprehensive reforms to their curriculum, assessments, and teacher evaluations in exchange for substantial grants to fund these changes. Three major requirements were the state's adoption of the Common Core State Standards, removal of any barriers to using test scores to evaluate teachers, and removal of any limits on the number of charter schools in a state (U.S. Department of Education, 2009). As of August 2012, all but five states had adopted the Common Core State Standards (CCSSI, 2012). In a time during which many states were facing budget shortfalls, the opportunity to receive additional federal dollars for education proved to be a powerful motivator. Although the prerequisites for application were not required by law, 40 states and Washington, DC, chose to make the changes and apply in Phase 1. Delaware and Tennessee were awarded $100 million and $500 million, respectively, in Phase 1 (U.S. Department of Education, 2010a), and nine additional states and Washington, DC, were awarded a total of $3.3 billion in Phase 2 (U.S Department of Education, 2010b). In August 2011, Secretary of Education Arne Duncan announced that he would grant waivers of some NCLB requirements to states able to show they were making progress in reforming and improving their educational and accountability systems using similar evaluative criteria as in Race to the Top (Dillon, 2011). It is currently unclear which provisions of NCLB will continue to be enforced and how many states will be granted waivers in the interim before ESEA reauthorization.

An additional component of the Race to the Top competition was related to assessment and allowed consortia of states to propose comprehensive assessment systems that represented a major overhaul of traditional assessment models. The competition explicitly recognized "the dual needs for accountability and instructional improvement" and sought to yield assessments that met both needs by measuring proficiency as well as growth and providing

data that could be used to evaluate schools and teachers (Duncan, 2010, p. 18171). With $160 million available to develop these assessments, the consortia were able to "dream big" in incorporating cutting-edge techniques and technology to reshape K–12 achievement testing.

Two consortia—the Partnership for Assessment of Readiness for College and Careers (PARCC) and the SMARTER Balanced Assessment Consortium (SBAC)—were awarded grants to develop these assessment systems. As of 2011, PARCC represented 23 states and SBAC represented 29 states, although there is some overlap between states (Center for K–12 Assessment Management and Performance at ETS, 2011), and some states have changed their affiliation. Both consortia have developed plans that consist of much more than a single end-of-year test of proficiency. More details about the two planned testing systems will be provided throughout the chapter. For a compendium of both plans, see the K–12 Center's website (http://www.k12center.org/publications.html). The use of performance assessments are called for in both plans (see Volume 1, Chapter 20, this handbook).

Summative, Interim, and Formative Assessment

The type of test that is probably most discussed currently in public discourse is summative achievement test use for accountability. The distinction of "summative" refers to the test being at the end of instruction, currently typically at the end of the year. Another summative assessment schedule type is a series of summative assessments throughout the year, such as at the end of each quarter. In either instance, the purpose of the assessment(s) is to measure what students have learned and use the results for evaluation, traditionally for the evaluation of an individuals or groups of students, and, more recently, in the evaluation of their teachers and schools.

A benefit of having a single test at the end of the year is that students have full opportunity to learn the material before they and their teachers are judged on mastery of the material. Furthermore, when the test is at the end of the year, it allows for teachers to have variation in their pacing of the curriculum, provided they cover the expected material by the time of the test. From a psychometric perspective, researchers are much more familiar with scoring results from a single administration rather than combining responses from multiple administrations into a single score (Wise, 2011).

There are drawbacks of a single end-of-year test, however. Most notably, at the end of the year, it is too late to make changes for the current year based on test results, potentially delaying reform and needed remediation. Because of the length of the test, it is usually not possible to administer the test in a single class period, causing disruption to the school schedule and making the test less similar to regular instruction. By the end of the year, students may have forgotten what they learned earlier, prompting the need for review time or leading to lower scores. A high-stakes score from a single day is especially susceptible to inconsistencies in student performance, such as when not feeling well or not having time for breakfast. It is often difficult to include the test scores in students' grades because scores are available only after student grades need to be submitted, thus potentially lowering students' motivation to try their best on the exam. Despite having a major goal of educational evaluation, a single test is often too limited in scope to provide detailed, rich information to evaluate teaching (Perie, Marion, & Gong, 2009).

Because of these reasons and others, recent changes in assessment may reflect a growing enthusiasm for multiple assessments throughout the year, often referred to as "through-course" assessments. The tests administered before the end of the year can be summative, formative, or interim, depending on their design and use. For example, in a model called cognitively based assessment of, as, and for learning (CBAL), a critical change from traditional accountability models is that the accountability component is separated into multiple summative assessments throughout the year (Bennett & Gitomer, 2008). Class grades are a compilation of many evaluative performances, so it is more consistent with classroom practice to distribute assessments throughout the year as well. Previous design protocols for the PARCC assessments called for

through-course summative assessment (Jerald, Doorey, & Forgione, 2011), but these summative assessments have now been replaced by formative or interim assessments, with the possibility of an additional summative assessment (Center for K–12 Assessment Management and Performance at ETS, 2011).

Interim and Formative Assessments

In contrast to summative assessments, interim and formative assessments are completed during the course of learning, rather than after teaching is completed. The purpose of interim and formative assessments are markedly different because rather than serving as an after-the-fact indicator of what has been learned, they are intended to produce "actionable" results during the same school year. Popham (2008) defined formative assessment as "a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or by students to adjust their current learning tactics" (p. 112). Both assessment consortia that received major funding through Race to the Top, PARCC and SBAC, currently incorporate formative or interim assessment as major components of their planned design (Center for K–12 Assessment Management and Performance at ETS, 2011). CBAL also relies on more frequent formative assessment opportunities designed to promote student learning (Bennett & Gitomer, 2008).

Although some researchers and practitioners do not consistently agree on the distinction between formative and interim assessment, Perie et al. (2009) stated that formative assessment is more frequent than interim assessment and more integrated into the classroom to inform immediate instructional adjustments. These assessments are more flexible than interim assessments and would not be aggregated at the school or district level. Interim assessments, then, have been defined as follows:

> Assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level. The specific interim assessment designs are driven by the purposes and intended uses, but the results of any interim assessment must be reported in a manner allowing aggregation across students, occasions, or concepts. (Perie et al., 2009, p. 6)

Therefore, most large-scale programs administered during the school year rather than at the end are more likely to be considered interim rather than formative assessment under this nomenclature.

The use of interim assessments can be divided into three types of purposes: instructional, evaluative, and predictive (Perie et al., 2009). For instructional purposes, interim assessment is distinguished from formative assessment in part by being less frequently administered, but it shares the function of modifying instruction to improve learning for the students who took the test. For evaluative purposes, the assessment should provide teachers and administrators clear results that help make curricular or instructional changes. When assessments are designed to predict future performance, such as on the end-of-year summative test, this is considered interim assessment. Designing interim assessments requires a balance of these goals. For example, tests used for predictive purposes should be very similar to the test it is intended to predict, but for instructional purposes, the tests need to address a smaller number of curricular objectives to give appropriate and timely information about each objective (Perie et al., 2009).

Although classroom teachers do strive to provide measures throughout the year that relate to student scores on the end-of-year test, it is best when interim assessment systems are centrally supported to provide consistently high-quality assessments. This approach further allows the score interpretation to be linked to instruction in a planned way, consistent with the recommendation by Nichols, Meyers, and Burling (2009). Bennett and Gitomer (2008) emphasized the importance of an internally consistent set of assessments (both summative and formative or interim) that also utilize what is known from research about effective testing and teaching.

Developing an assessment that appropriately meets the needs of different stakeholders is difficult, time-consuming, and often costly. When interim assessments are centrally created, students across the jurisdiction can benefit from the time and expertise that were involved in its creation. In addition, a central database of student records would allow mobile students' interim assessment scores to be accessed and meaningful even if they change schools. After the first year of administration or pretesting of the interim assessment system, the state or jurisdiction should have quality data showing the relationship between scores on the interim assessment and scores on the end-of-year test. They can then use this information to predict student performance more accurately for later administrations. Students whose interim assessment results predict nonproficient performance on the end-of-year test can be targeted for extra help to meet learning goals, particularly when these scores are used in conjunction with other information.

The term *formative* has many connotations of benefits to students and teachers, but "labels can be a powerful way to miscommunicate" (Nichols, Meyers, & Burling, 2009, p. 14). Consistent with the discussion in Volume 1, Chapter 13, this handbook; the Standards for Educational and Psychological Testing (AERA et al., 1999); and Kane (2006), tests should have clearly defined uses that are directly supported by the accumulation of validity evidence. Nichols et al. (2009) have pointed out that the term formative assessment implies that the use of the test directly informs and improves instruction with high-quality, relevant information and yields an increase in achievement, a use that is rarely empirically supported. They wrote that "the argument must causally link information from performance on a particular assessment to the selection of instructional actions whose implementation leads to gains in student learning" (p. 15). The authors provided a multiphase framework for evaluating the extent to which assessments merit these claims. This evaluative framework emphasizes that following the assessment, instruction should be tailored based on test information and that a summative assessment should provide evidence of student improvement.

Norm-Referenced and Criterion-Referenced Assessments

A distinction traditionally has made between two classes of assessments: norm-referenced and criterion-referenced assessments. It is perhaps now appropriate to consider whether interpretations of test scores are norm or criterion referenced, rather than only the tests themselves. In all cases, it is necessary to consider critically whether the test supports the intended score interpretation.

Norm-referenced assessments or interpretations are characterized by their emphasis on comparing students to other students, rather than against a set criterion or performance level. For such score interpretations, a representative norm sample of students takes the test, and that sample's scores (typically weighted) form the basis for the norms when the test is used in the future. When later students take the assessment, they receive scores that represent the percent of the norm population that the student performed better than or equal to. In designing such tests, the emphasis is on items that spread students out over the score continuum. It is critical that these norms are both representative and fairly current in order to provide accurate norm-referenced scores (AERA et al., 1999). For example, norms that are 5 years old will compare current students to students 5 years ago rather than to current students. To the extent that student achievement of the U.S. population is not stable over time, older norms will not represent the current population of students.

In contrast, criterion-referenced assessments report scores in relation to a domain of knowledge and skills rather than in comparison with other test takers. These types of assessments focus on the skills or knowledge called for in the curriculum or instruction. Sometimes, simply percent correct scores can be considered criterion referenced to the extent the items on the test accurately represent the domain. In some cases, there are clearly defined criteria for mastery or proficiency to which student performance is compared. This type of interpretation of scores is currently required by ESEA, as described previously in this chapter.

These criteria or performance standards are typically established through a standard-setting process that identifies the level of performance on

the assessment that corresponds to expectations for student skills. Because scores are interpreted in relation to this criterion, a quality standard-setting procedure is an essential component of appropriate test use (see Chapter 22 in this volume, which more fully describes standard-setting methods). Some criterion-referenced assessments have several categories of student performance expectations, typically with one level designating proficiency and at least one level designating for performance below expectations and one for above expectations. Although the current ESEA legislation mandates such performance levels, these levels can serve as useful indicators of student knowledge and skills, provided that they are meaningfully derived.

Some of the more widely used norm-referenced achievement tests in K–12 education include the Iowa Tests of Basic Skills (ITBS; Hoover et al., 2003) and the Iowa Tests of Educational Development (ITED; Forsyth, Ansley, Feldt, & Alnot, 2006); the Stanford Achievement Test Series (10th ed. [SAT-10]; Harcourt Assessment Inc., 2003); and the TerraNova (3rd ed.; CTB-McGraw Hill, 2011). These tests also provide scale scores in addition to percentiles. These tests are available in multiple forms, which allow them to span a wide range of grades from kindergarten to 12th grade.

Importantly, some norm-referenced tests now provide or support criterion-referenced interpretations, somewhat blurring the traditional distinction. Comparing the websites of these tests, one finds multiple score types provided. For example, the website of the Iowa Testing Program's ITBS explains to potential users that criterion-referenced interpretations can be made by setting clear performance standards (University of Iowa, 2011). The scores provided for the SAT-10 include comparisons to norm groups, with the number correct by topic provided for those seeking a criterion-referenced interpretation. The TerraNova includes a modification of percentage correct by objective and classification into low mastery, moderate mastery, and high mastery (CTB-McGraw Hill, 2011), but the process for classifying mastery level is not clear, especially because objectives may be measured by as few as four items.

It is always necessary that the test development process is appropriate for the intended use of scores. Because traditional norm-referenced test items would be selected for their ability to differentiate students across the distribution, this approach may not accurately select items that best represent the skills students should have mastered. Conversely, when student performance is compared with set performance levels in criterion-referenced assessment, it is most important to maximize information near the cut points for proficiency classification but may not rank students well further from the mean. Thus, the strategies used to create tests for these two different purposes differ (see Volume 1, Chapter 7, this handbook, for procedures used to select test items for both types of assessment purposes).

As discussed in this chapter, results on state-based criterion-referenced assessments are not comparable between states. Because norm-referenced assessments include national norms, such tests provide one way to compare achievement between students in different states. Because a nationally used test has not been able to match the content standards of each state, however, interpretation of scores from national test is limited by any mismatch between test content and state curriculum. For achievement tests, it is important to understand the match between local curricular expectations and the test content when interpreting scores, keeping in mind that student performance can be interpreted only in relation to the test content.

Graduation Examinations

This chapter has already addressed a major type of large-scale summative assessment in K–12 achievement testing—state tests for accountability—but a type of test with perhaps an even larger direct consequence for students is graduation examinations. There is an enormous disparity among the skills demonstrated by 12th graders within a state. Especially in some very low-performing schools, large numbers of students do not possess the level of skill required for proficiency in reading, writing, and mathematics. As a way to guard against these students entering the workforce with a high school diploma that does not match their skill set and to encourage higher achievement, many jurisdictions

have enacted graduation examinations that students must pass to earn a high school diploma. In the 2009–2010 school year, 28 states had graduation examinations, and two additional states were developing them (Center on Education Policy, 2010).

These tests can either be the same test used for state accountability reporting or a specialized test tailored to skills required for graduation. If one test is co-opted as a graduation examination, it is essential that the minimum score required for graduation be set separately from the existing cut scores for proficiency. Panelists in such a standard setting should be instructed to envision the minimum skills required to deserve a high school diploma rather than relying on performance definitions designed for accountability purposes.

Historically, there has been controversy over the fairness of high school graduation tests. It is generally not considered fair or appropriate to withhold a student's right to a diploma if he or she did not receive instruction in the material covered on the graduation test. This is a notable distinction from test content in tests used for accountability, which measure what should have been taught and learned but typically without direct consequences for students (Geisinger & McCormick, 2010). Graduation exams can be particularly subject to lawsuits, especially to the extent that some minority-group students, on average, perform lower on achievement tests in the jurisdiction or pass the graduation exam at a lower rate. Eighty-three percent of minority students in the United States live in states that have graduation exams (Center on Education Policy, 2010). If implementing a graduation examination, the jurisdiction should carefully document the design and implementation processes to provide stronger defense against such accusations (Phillips & Camara, 2006).

A landmark case relating to this point is *Debra P. v. Turlington* (Phillips, 1991), in which the judge ruled that the students in the case had been inappropriately denied their rights to graduation when they were held accountable for material they had not been taught. A key point in this case was that the exam requirement was added after the students were already well into their education. The judge also ruled, however, that the test could be used after a transition period in which content coverage in the classroom was expanded. A related case defending high school graduation exams was *GI Forum v. Texas Education Agency* (2000), in which the Texas Education Agency successfully demonstrated the test's validity for the purpose of determining graduation. Ward (2000) wrote that the case highlighted that the need for expert staff, a clear purpose and plan, close coordination with contractors, firm support for teacher and administrative professional development, ongoing research, equitable distribution of teaching, and rigorous documentation are essential for a defensible high school graduation exam.

Graduation exams on one hand seem to protect qualified students by ensuring their high school diploma tells potential employers they have met minimum expectations for high school learning. On the other hand, students in most severe need of special education services for cognitive disabilities may be unlikely to meet such requirements. In effect, universally required graduation exams strip these students of the ability to receive a high school diploma as a result of their 12 years of education (Guy, Shin, Lee, & Thurlow, 2000). Individuals may argue whether students who cannot meet the requirements deserve a diploma—whether or not they have disabilities—but like the previously raised issue for minority students, assessments adversely affecting students with disabilities have a tendency to lead to confrontation, if not to court. Once again, responsible parties should be prepared for such complaints and consider test fairness when implementing such a test.

Group-Score Assessments (National and International Assessment)

Many of the questions individuals may wish to answer are not currently addressed by general large-scale assessment but can be answered through specially developed group-score assessments. How has student achievement changed over the past 30 years? Is there a larger difference in achievement between Black and White students in California or Mississippi? Does teacher certification have a significant effect on student achievement? Do students in America or India spend more time on mathematics

homework? What country has the highest student reading achievement in the world? Around the world, do boys or girls perform better on mathematics?

Group-score assessments refer to tests that use special procedures to measure achievement of groups such as states or countries rather than individual students (Mazzeo, Lazer, & Zieky, 2006). The major group-level assessment used to measure student achievement in the United States is NAEP. In addition, achievement of children in America can be compared with that of students in other countries using TIMSS, Progress in International Reading Literacy Study (PIRLS), and the Programme for International Assessment (PISA). Unlike tests used for accountability in which all students are currently assessed, a representative sample is drawn from a population in these survey assessments. Because the goal of these assessments is to assess student achievement at the group level, whether state or country, it is not necessary to administer all items to all students. A unique feature of this sort of assessment is that individual students do not receive scores but rather are assigned plausible values, which take into account both sampling error and additional error arising from the matrix sampling test design (Mislevy, Johnson, & Muraki, 1992). This feature leads to a need for specialized analysis procedures.

These testing programs are not simply achievement tests but larger research studies that allow researchers to investigate a wide range of topics. Each tested student completes an achievement test and also a background questionnaire. In addition, their family, teacher, and school leaders may complete separate background questionnaires. The answers to these questionnaires are available for researchers to use in their analyses, either as predictors of student achievement or as dependent variables themselves. Student demographic variables are especially important in such analyses.

A major goal of NAEP in the United States is to track the achievement gap over time to determine whether it is growing or shrinking. The achievement gap refers to differences in achievement test scores between White students and students from traditionally underrepresented minority groups as well as test score differences between affluent and low-income students. These differences represent a pervasive problem in U.S. education reflected in a wide range of achievement tests. In general, the term is used to refer to differences thought to exist in actual academic achievement rather than to differences attributed to test bias. Comparing the 2007 national results with the 2005 results (Vanneman, Hamilton, Anderson, & Rahman, 2009), the Black–White achievement gap narrowed in eighth-grade mathematics and fourth-grade reading but not in fourth-grade mathematics or eighth-grade reading. In all four grade and subject combinations, however, White students performed significantly better than Black students in every state for which data were available, except for eighth-grade reading in Hawaii. The Black–White achievement gap in eighth-grade reading remains a persistent problem, with no states significantly narrowing it between 2005 and 2007.

There are major concerns about the educational preparation of students in U.S. schools compared with those in other countries. Results of international assessments in some ways provide a basis for this concern as the United States lags behind peers and even certain developing economies. In particular, in 2007 Singapore topped the list of TIMSS scores for mathematics and science and outperformed the United States in mathematics to such an extent that scores of Singapore's 75th percentile are higher than the United States' 95th percentile (Mullis et al., 2008). Results such as these are valuable for informing educational policy within the United States.

Alternate Assessments

The goal of large-scale assessment is to measure the skills of a wide range of students efficiently and accurately. However, some students receiving special education services cannot interact with the assessment appropriately, even with certain accommodations. As such, current federal law allows for states to meet reporting requirements for up to 1% of students using an alternate assessment better suited for students with severe cognitive disabilities. A key feature of these tests is that they must measure academic content standards, not simply life skills. Through demanding assessment of academic

content, the hope is that students with severe disabilities will benefit from increasingly academic instruction.

Alternate assessments typically have different formats than the regular assessment and usually are administered individually rather than in groups. Alternate assessments need to demonstrate high-quality development procedures, but due to the unique student population, the tests, including their development and scoring, are expected to differ from the regular assessment. Alternate assessment is an important component of accountability policy in that such tests work with the regular assessment to measure whether all students are learning the academic content knowledge and skills expected of them. For more information about alternate assessments and the challenges associated with them, see Schafer and Lissitz (2009); see also Chapter 18 in this volume.

HIGH STAKES FOR WHOM? MOTIVATION IN TESTING

There are many types of tests with many different uses and potential consequences. A defining characteristics of achievement testing in K–12 education is that consequences may fall on individuals other than the test takers. For example, under ESEA, schools have been evaluated and forced to undergo reforms based on student test performance. Individual teachers increasingly are being evaluated on the basis of their students' test scores, as described with the Race to the Top. Plake (2011) defined *high-stakes testing* as "when a test is used to make important decisions, whether those decisions are about the student who took the test or about others involved in the education process" (p. 11). If the consequences a student would face for low performance are less severe than those a teacher would face for low student performance, then the stakes are higher for the teacher than the students.

When students do not receive direct consequences for their test performance, they may not be motivated to try their best. Studies have shown that student test performance can be increased by providing incentivizing rewards for performance (e.g., Braun, Kirsch, & Yamamoto, 2011), which may be missing from many tests used for accountability. Factors that influence student test performance other than their mastery of the content can harm the validity of scores. This is particularly worrisome when the scores are used to make important decisions about schools and teachers.

INSTRUCTIONAL SENSITIVITY

At first, the idea of using tests to measure how much students have been taught seems relatively straightforward. Upon deeper consideration, however, it becomes clearer that rather than measuring how much students *have been taught*, tests measure how much students *know and can do*. Test users often are left to assume that students' performance is based on the quality and content of instruction for the year of learning the test is meant to evaluate. To equate measurement of learning with measurement of skills and knowledge is actually a large logical leap that assumes (a) students would have gotten the questions wrong before they were taught the material, (b) students would get the questions right if they were adequately taught the material, and (c) change in performance is not simply the result of maturation. To attribute this learning to the current year's teacher also assumes that (d) students would have gotten the questions wrong before the current year and (e) students learned the material from this classroom teacher. Although tests are used for these purposes, these issues are not raised in the NCLB guidelines. According to Braun (2009), "the fact that the test score obtained one spring day is an outcome that depends on a student's entire history and not just on the school-based inputs of the past year apparently did not trouble NCLB's proponents" (p. 52).

Following Polikoff (2010), instructional sensitivity is "the extent to which student performance on a test or item reflects the instruction received" (p. 3). One of the most vocal advocates for tests that are instructionally sensitive is James Popham. Popham (2007) has argued that much of student test performance is not based on whether the student learned the material in the classroom but rather on socioeconomic status and innate aptitude, and he urged greater use of procedures to evaluate tests' instructional sensitivity. (For current procedures used to

evaluate the impact of student background such as race and family income as they relate to test fairness, see Volume 1, Chapters 8 and 17, this handbook, and Chapter 27, this volume.) In short, tests should measure "what students were *taught* in school," not "what students *brought* to school" (Popham, 2008, p. 126).

In Popham's (2007) proposed evaluation of items, responsiveness to instruction is identified as follows: "If a teacher has provided reasonably effective instruction related to what's measured by this item, is it likely that a substantial majority of the teacher's students will respond correctly to the item?" (p. 150). Instead, traditional test construction procedures often select items with high variation in performance between students and thus actually may be an obstacle to selecting items most responsive to instruction (Popham, 2008).

Polikoff (2010) reviewed various methods of evaluating instructional sensitivity, including empirical statistics, such as those creating contingency tables of pretest and posttest item responses, those focusing on instructional methods, and those based on expert judgment, such as Popham (2007) proposed. Limitations of pretest and posttest methods are that they typically require extra data collection and do not control for the effect of maturation rather than classroom learning. Furthermore, if items do not show postinstruction gains, it is not clear whether the item or the instruction is at fault. They, however, can be useful indicators of sensitivity.

Among methods focusing on instructional methods, Polikoff (2010) concluded that it is more effective to measure how much time was spent covering different topics rather than only asking whether a topic was covered. This teacher-reported data can then be correlated with student achievement gains, provided that appropriate comparison data are available. D'Agostino, Welsh, and Corson (2007) found that there was also an effect for the correspondence between the way content was taught and the way it was assessed. Polikoff questioned the ability of judges to evaluate instructional sensitivity but encouraged the establishment of clear methodologies for judgment-based measures of instructional sensitivity. Although various methods to analyze instructional sensitivity are available, Popham (2007) and Polikoff have agreed that instructional sensitivity is not evaluated nearly as consistently as it ought to be:

> There is ample evidence that instructional sensitivity is an important facet of any criterion-referenced assessment. It is even more apparent that sensitivity has largely been ignored as a feature of tests or items worth studying. This seems a grievous oversight, one that threatens the validity of the thousands of decisions that are made annually based on the results from state assessments under NCLB. (Polikoff, 2010, pp. 12–13)

MEASURING GROWTH

It has long been recognized that simply counting the percentage of students classified as proficient does not do enough to evaluate the teacher's contribution to student learning. Unfortunately, NCLB legislation as originally written focused nearly solely on the percent of student proficient, that is, a status model. In such a system, teachers whose students start the year with advanced skills could make less than a year's growth and still meet the proficiency target, whereas teachers whose students start the year far behind expectations could produce 2 years' worth of student growth in 1 academic year but still be labeled as not meeting the standard. Clearly, such a system puts particular pressure on those teachers whose students are most disadvantaged and farther behind, and it fails to recognize these teachers' contributions to bringing students closer to proficiency. Betebenner (2009) concluded that status models "are inappropriate for judgments about educational effectiveness" (p. 42) because they do not show the effect schools and teachers have on learning over time.

Some states have supplemented federal status requirements with their own models that track student achievement over time, often called growth models. Despite the status-only model prescribed in NCLB, states, testing professionals, elected officials, and parents saw the importance of measuring growth in addition to status. For example, if administrators

want to know how much this year's fourth graders have progressed since last year, they need scores that can show such growth. Measurement professionals put different tests on the same scale to allow comparability between scores on different tests. A specialized technique, often called vertical scaling, allows scores to be compared between grades. When attempting to measure change over time, it is preferable if a vertical score scale has been created to track growth in a quantifiable way, rather than simply examining changes in proficiency status. Although there are some costs to completing such between-test scaling, comparable scores are especially useful toward informing improvement efforts because they allow teachers, administrators, parents, and the public to know how student achievement has changed. Volume 1, Chapter 11, this handbook, more thoroughly addresses scaling.

RELIABILITY AT CUT SCORES

When tests are designed to differentiate between students who are proficient, compared with not proficient, many items will and should focus on the part of the ability spectrum near that cut score. Doing so, however, also results in less information about students with very high or very low ability. The amount of information the test provides for each student is closely tied to the number of items close to the student's ability level. With less information about these extreme students' abilities, scores for these students will be less reliable. Psychometricians can calculate how much error is likely in students' scores at different points along the ability spectrum. As described, most tests will show less error near the middle and more error at the high and low ends of the scale, unless those tests are adaptive. When making decisions about students, it is especially valuable to have information about error in scores, not just overall, but at different score points.

A test that could yield very accurate decisions for students near the mean might have too much error at more extreme ability levels to be used properly for intended purposes. Particularly if the testing program uses multiple proficiency levels—as current state tests do—it is important that reliability is high near each of these cut scores to yield more accurate classifications. This information should be available for most large-scale achievement testing programs, although it may not be readily accessed by the public. If the testing program publishes a technical manual on its website, look for this information there.

COMPUTERIZED ADAPTIVE TESTING

There is a constant desire to make assessments shorter while maintaining their high reliability or to make assessments more reliable without making them longer. These objectives can be achieved using the much-researched technique of computerized adaptive testing. As more K–12 achievement testing programs transition to computers instead of paper-and-pencil tests, the option to make the test adaptive becomes more appealing. There are, however, some additional costs and risks of doing so.

The basic premise of computerized adaptive testing is that in fixed tests where all students answer all items, many items are much too difficult or much too easy for any given student. When students answer questions inappropriately matched to their ability level, these items provide little information that can help pinpoint what a student's score should be. Instead, if students answer items closer to their ability level, then each item becomes more valuable in achieving an accurate score. In large-scale assessment, this is not possible with a fixed test because students vary greatly in their skills and knowledge. Instead, by dynamically tailoring the items administered to each student, fewer items need to be administered to reach the same level of reliability as a fixed test (Wainer, 2000; Weiss & Kingsbury, 1984). This item selection is based on the student's performance on previous items. Although the principles of computerized adaptive testing are not new, the nearly universal availability of a large number of computers in schools was a prerequisite to its use in large-scale K–12 assessment. The SMARTER Balanced Assessment Consortium has repeated that it expects to use computerized adaptive assessment (Center for K–12 Assessment Management and Performance at ETS, 2011).

On the basis of the principles of these computerized adaptive testing, one might wonder why all tests are not adaptive. One major obstacle is that

adaptive tests require much larger pools of items to provide all students with items better suited for them (Wainer, 2000). For testing programs already fighting for budget dollars, the cost of these additional items can add up quickly. Many states require that items counting toward student scores be released after each administration. Releasing entire item pools for adaptive tests each year would be prohibitive. Furthermore, current federal requirements stipulate that the test content must assess grade-level standards, limiting the flexibility of the adaptive process to measure all students' ability levels. Less technically, there is a very real concern about explaining to parents that students were given different items but somehow their scores can be compared. A full answer to that question would require a lesson in item response theory and automated item selection, and although some eager psychometricians might gladly provide it, the listener is likely to be less eager. For more information about computerized adaptive testing, see Volume 1, Chapter 10, this handbook.

CONCLUSION

This chapter has addressed many of the important issues in K–12 achievement testing and when combined with related chapters throughout the handbook provides a rich understanding of the practices and challenges. Although achievement tests are defined as tests to indicate student knowledge of and skills in academic content areas, a key feature of achievement tests in K–12 education is their intended use to inform decisions and actions to improve education for individual students and groups of students. Education assessment in the K–12 setting is at a crucial transition point, with the opportunity to advance the use of research-based assessment practices and answer more questions through additional research. Even though the future of educational assessment is unclear, it appears poised to continue to play a critical role in education evaluation and reform. With this continued high-stakes use comes an increasing need to ensure that tests are used appropriately. It is essential that those involved in achievement testing understand the complex issues that accompany their use.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Bandalos, D. L., Ferster, A. E., Davis, S. L., & Samuelsen, K. M. (2011). Validity arguments for high-stakes testing and accountability systems. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High stakes testing in education: Science and practice in K–12 settings* (pp. 155–175). Washington, DC: American Psychological Association. doi:10.1037/12330-010

Bennett, R. E., & Gitomer, D. H. (2008). *Transforming K–12 assessment: Integrating accountability testing, formative assessment, and professional support*. Princeton, NJ: Educational Testing Service.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51. doi:10.1111/j.1745-3992.2009.00161.x

Braun, H. (2009). Discussion: With choices come consequences. *Educational Measurement: Issues and Practice*, *28*(4), 52–55. doi:10.1111/j.1745-3992.2009.00162.x

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record*, *113*, 2309–2344.

Center for K–12 Assessment Management and Performance at ETS. (2011, July). *PARCC and SMARTER balanced assessment designs—Approved by consortia*. Princeton, NJ: Author.

Center on Education Policy. (2010, December). *State high school tests: Exit exams and other assessments*. Washington, DC: Author

Common Core State Standards Initiative. (2010). *Common core state standards for English language arts and literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

Common Core State Standards Initiative (2012). *In the states*. Retrieved from http://www.corestandards.org/in-the-states

CTB-McGraw Hill. (2011). *TerraNova, Complete Battery* (3rd ed.). Retrieved from http://www.ctb.com/ctb.com/control/ctbProductViewAction?productFamilyId=449&productId=733&p=products

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state standards-based assessment. *Educational Assessment*, *12*, 1–22.

Dillon, S. (2011, August 8). Overriding a key education law. *New York Times*. Retrieved from http://www.nytimes.com/2011/08/08/education/08educ.html

Duncan, A. (2010, April 9). Overview information; Race to the Top fund assessment program; notice inviting applications for new awards for fiscal year (FY) 2010. *Federal Register*, *75*(68), 18171–18185.

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice-Hall.

Foley, B. F. (2011, April). *Realistic expectations: State-level changes in the percentage of proficient students 2002–2008*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Forsyth, R. A., Ansley, T. N., Feldt, L. S., & Alnot, S. D. (2006). *Iowa Tests of Educational Development*. Rolling Meadows, IL: Riverside.

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, *29*(1), 38–44. doi:10.1111/j.1745-3992.2009.00168.x

GI Forum v. Texas Education Agency, No. SA-97-CA-1278-EP (GI Forum filed January 7, 2000).

Guy, B., Shin, H., Lee, S., & Thurlow, M. (2000). State graduation requirements for students with and without disabilities. In D. R. Johnson & E. J. Emanuel (Eds.), *Issues influencing the future of transition programs and services in the United States* (pp. 85–110). Minneapolis: University of Minnesota.

Harcourt Assessment, Inc. (2003). *Stanford Achievement Test* (10th ed.). San Antonio, TX: Pearson.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*, 351–360. doi:10.3102/0013189X08323842

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Bray, G. B., Naylor, R. J., . . . Qualls, A. L. (2003). *Iowa Test of Basic Skills*. Rolling Meadows, IL: Riverside.

Jerald, C. D., Doorey, N. A., & Forgione, P. D. (2011). *Putting the pieces together: Summary report of the invitational research symposium on through-course summative assessments*. Retrieved from http://www.k12center.org/rsc/pdf/TCSA_Symposium_Final_Summary.pdf

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*(1), 31–41. doi:10.1111/j.1745-3992.2002.tb00083.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–46). Washington, DC: National Council on Measurement in Education and the American Council on Education.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Washington, DC: National Council on Measurement in Education and the American Council on Education.

Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives*, *13*(33), 1–17.

Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 681–700). Washington, DC: National Council on Measurement in Education and the American Council on Education.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*, 131–154. doi:10.2307/1165166

Mullis, I. V. S., Martin, M. O., & Foy, P., with Olson, J. F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grade*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

National Governors Association & Council of Chief State School Officers. (2010, June.) *Reaching higher: The common core state standards validation committee*. Washington, DC: National Governors Association & Council of Chief State School Officers. Retrieved from http://www.corestandards.org/assets/CommonCoreReport_6.10.pdf

Nichols, P. D., Meyers, J. L., & Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice*, *28*(3), 14–23. doi:10.1111/j.1745-3992.2009.00150.x

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Nunnally, J. C. (1972). *Educational measurement and evaluation* (2nd ed.). New York, NY: McGraw-Hill.

Pearson Education, Inc. (2011). *Stanford Achievement Test Series* (10th ed.). Retrieved from http://www.pearsonassessments.com/haiweb/cultures/en-us/productdetail.htm?pid=SAT10C&Community=EA_PreK–12_API_Achievement

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, *28*(3), 5–13. doi:10.1111/j.1745-3992.2009.00149.x

Phillips, S. E. (1991). Diploma sanction tests revisited: New problems from old solutions. *Journal of Law and Education,* 20, 175–199.

Phillips, S. E., & Camara, W. J. (2006). Legal and ethical issues. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 733–755). Washington, DC: National Council on Measurement in Education and the American Council on Education.

Plake, B. S. (2011). Current state of high-stakes testing in education. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High stakes testing in education: Science and practice in K–12 settings* (pp. 11–26). Washington, DC: American Psychological Association. doi:10.1037/12330-001

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29*(4), 3–14. doi:10.1111/j.1745-3992.2010.00189.x

Popham, J. W. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, *89*(2), 146–155.

Popham, J. W. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards: The new U.S. intended curriculum. *Educational Researcher*, *40*, 103–116. doi:10.3102/0013189X11405038

Schafer, W., & Lissitz, R. W. (2009). *Alternate assessments based on alternate achievement standards: Policy, practice, and potential*. Baltimore, MD: Paul H. Brookes.

Schmidt, W., Houang, R., & Cogan, L. (2002). A coherent curriculum: The case of mathematics. *American Educator*, *26*(2), 1–17.

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*, 553–573. doi:10.1080/00273171.2010.483382

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson.

U.S. Department of Education. (2009, November). *Race to the Top program executive summary*. Washington, DC: U.S. Department of Education. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education. (2010a, March 29). *Delaware and Tennessee win first Race to the Top grants*. Retrieved from http://www.ed.gov/news/press-releases/delaware-and-tennessee-win-first-race-top-grants

U.S. Department of Education. (2010b, August 24). *Nine states and the District of Columbia win second round Race to the Top grants*. Retrieved from http://www.ed.gov/news/press-releases/nine-states-and-district-columbia-win-second-round-race-top-grants

University of Iowa. (2011). *Iowa testing programs*. Retrieved from http://itp.education.uiowa.edu

Vanneman, A., Hamilton, L., Anderson, J. B., & Rahman, T. (2009). *Achievement gaps: How Black and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress*. Washington, DC: National Center on Education Statistics, Institute for Education Sciences, U.S. Department of Education.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Erlbaum.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361–375. doi:10.1111/j.1745-3984.1984.tb01040.x

Wise, L. L. (2011, February). *Picking up the pieces: Aggregating results from through-course assessments*. Paper presented at the Invitational Research Symposium on Through-Course Summative Assessments, Atlanta, GA.

CHAPTER 17

TESTING OF ENGLISH LANGUAGE LEARNER STUDENTS

Jamal Abedi

English language learner (ELL) students face dual challenges in their academic lives: (a) the challenge of learning a new language and adjusting to a new culture and (b) the more serious challenge of learning content knowledge in a language that they are still struggling to acquire. In the process of learning a new language and adjusting to a new culture, they may be faced with many equity and fairness issues. Thus, attention to the academic needs of ELL students is of great importance as the number of these students in the nation increases rapidly. Nationally, ELL K–12 enrollment has grown 57% since 1995, whereas the growth rate for all students has been less than 4% (Flannery, 2009).

Assessment outcomes play a vital role in the academic careers of ELL students. Normally, assessment of students' academic achievement is conducted after students are instructed in academic content. For ELL students, however, the situation is different. They have to be assessed for their level of English language proficiency (ELP) before receiving any academic instruction. On the basis of the outcome of ELP assessments and other related criteria, students are identified as ELLs or non-ELLs. ELP assessments typically include four domains of English proficiency: reading, writing, speaking, and listening. The ELL category includes students that are not fully proficient in one or more of the four domains of English. The non-ELL category includes native speakers of English and nonnative English speakers who are identified as initially fluent in English or reclassified as fluent in English based on ELL assessment outcomes and other relevant criteria. Those in the ELL category will likely receive ELL services. The type and level of such services depends on the state's policy and resources. If ELL students are improperly classified, they may receive inappropriate instruction and assessments.

From this brief introduction, it is clear that assessment outcomes shape ELL students' academic lives and should be given the utmost attention. Major or even minor threats to the validity of assessments could have grave consequences on ELL students' academic progress. In this chapter, we will discuss the principles of assessment for ELL students, the current status of assessments for these students and issues that could jeopardize the validity and authenticity of assessments for ELL students.

As noted, ELL students are assessed before they are instructed and required to go through two assessment systems and take two batteries of tests: (a) ELP assessments, referred to as Title III in the No Child Left Behind Act of 2001 (NCLB, 2002) to determine ELL students ELP levels; and (b) assessments of content knowledge, referred to as Title I assessment in NCLB. ELP assessment outcomes should inform participation of ELL students in content-based assessments. For example, it is important to know whether an ELL student is at a sufficient level of proficiency in English to be able to meaningfully participate in English-only instruction and assessment. We discuss the two assessment systems separately and then elaborate on how these two assessments can inform each other.

I thank Kimberly Mundhenk and Nancy Ewers for their research assistance with this topic and chapter.

DOI: 10.1037/14049-017
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

ENGLISH LANGUAGE PROFICIENCY ASSESSMENT

The assessment of an ELL student's ELP level is a major milestone in their academic career. These measures are the basis for classification, reclassification, curriculum planning, instruction, and fair assessment for ELL students. Without enough knowledge of ELL students' ELP levels, any decision regarding their instruction and assessment could be unproductive.

Many different ELP assessments have been used in the past and currently are being used in measuring ELL students' ELP levels. These assessments can be grouped into two major categories: those that were developed and used before the implementation of NCLB (2002) and those developed after the implementation of NCLB based on the NCLB Title III guidelines. Reviews of the pre-NCLB ELP assessments expressed concerns over the content coverage and validity of these assessments (Abedi, 2007; Zehler, Hopstock, Fleischman, & Greniuk, 1994). Many of the pre-NCLB assessments were not based on an operationally defined concept of English proficiency, had limited academic content coverage, were not consistent with states' content standards, and had psychometric flaws (Del Vecchio & Guerrero, 1995).

Title III of NCLB provided guidelines and instructions for creating a new generation of ELP assessments based on a set of predetermined state standards and cover a comprehensive set of ELP domains. Specifically, Title III of NCLB indicates that states must develop a set of clearly defined ELP standards in the four domains of reading, writing, speaking, and listening; and develop reliable and valid ELP tests in each of the four domains that are aligned to the state ELP standards. In addition, these new assessments must be aligned to states' content standards across three academic content areas (i.e., English language arts, mathematics, and science) and one nonacademic topic area related to the school environment (Fast, Ferrara, & Conrad, 2004).

Different consortia of states with the support from the U.S. Department of Education (USDE) under the Enhanced Assessment Grant opportunities developed ELP assessments that were based on NCLB Title III assessment guidelines. They measure proficiency in the four domains of reading, writing, speaking, and listening; they are aligned with the state ELP and content standards (in varying degrees); and they measure academic English (for a more detailed discussion of these consortia and their assessments, see Abedi, 2007). Although the new generation of ELP assessments shows much improvement over the existing ELP assessments (pre-NCLB), many different issues regarding these assessments remain to be resolved. Among these issues are problems in scoring and reporting the test outcomes, issues concerning inconsistencies between scores of the four domains, and inconsistencies among the post-NCLB assessment outcomes used by different consortia of states.

Issues concerning inconsistencies among outcomes of the four proficiency domains are among the most serious issues to consider in this chapter. The new generation of NCLB assessments reports four domain scores (reading, writing, speaking, and listening), a composite score of oral (speaking and listening) and a composite score of comprehension (reading and listening) in addition to an overall composite score of all four domains (Abedi, 2007). In computing this overall composite score, different consortia of states (or different individual states) weigh the four domains differently. Often the four domains are weighted equally in the composite scores but some consortia of states adopt differential weighting. For example, the World-Class Instructional Design and Assessment consortium places more weight on reading and writing (35% each) and less emphasis on listening and speaking (15% each; e.g., see Bauman, Boals, Cranley, Gottlieb, & Kenyon, 2007).

Major decisions are often made based on the composite scores in spite of discrepancies in arriving at these scores. For example, the decision on reclassification from an ELL category to "fluent in English" is mostly based on the ELP composite score. The major problem in the assessment of the level of English proficiency is the possible inconsistency between proficiency levels in the four domains and the proficiency level identified by the composite score. Literature has clearly shown that ELL students who have been reclassified as "proficient" on the basis of their composite ELP scores may not be proficient in one or more of the domains (e.g., see Abedi, 2008).

To elaborate on this issue, ELP composite scores can be based on two different models, a compensatory model and a conjunctive model (Abedi, 2004). In the compensatory model, the composite score is the sum of the four domain scores (weighted or unweighted). That is, a higher score in one domain can compensate for a lower score in the other domains. For example, a student who speaks quite well but is far below proficiency in reading may be reclassified as "proficient" in English because of her high speaking score. This can put such students at risk of failure because reading is one of the most important components in the academic performance of students (Hakuta, Butler, & Witt, 2000; Parker, Louie, & O'Dwyer, 2009). In a conjunctive model, however, students have to reach a proficiency level in each of the four domains regardless of how high their composite score is.

Because many consortia and individual states have adopted the compensatory model, ELL students can exit from an ELL category and the corresponding ELL services without being fluent enough in English, going on to participate in English-only instruction and assessments before they may be ready. This may put these students at risk of failure because of a language barrier.

CONTENT-BASED ASSESSMENT

Language factors can have a great impact on ELL students' academic performances. In this section we elaborate on the impact of language factors on student performance in content-based areas such as mathematics, science, and English language arts. Studies have clearly shown a major performance gap between ELL and non-ELL students, with ELL students performing substantially lower than non-ELL students in almost all content areas, but more so in areas with higher levels of language demands. With the rapid increase in the number of these students, attention to the nature and causes of the performance gap between ELL and non-ELL students is urgently needed before many of these students face the risk of academic failure.

Analyses of data from several locations nationwide consistently showed substantially lower performance by ELL students when compared with non-ELL students (non-ELL student groups consisted of native speakers of English and those bilingual students who were identified as initially fluent in English; Abedi, 2008). As indicated earlier, however, the performance gap between ELL and non-ELL students was mainly dependent on the assessment questions' level of language demand. For example, the performance gap between ELL and non-ELL was the highest in English language arts for which language is the focal construct. The performance gap was lower in such content areas as science and mathematics for which the science and mathematics content, not language, is the focal construct. The performance gap was lowest or even nonexistent in content areas with minimal levels of language demands, such as math computations (e.g., see Abedi, 2008; Abedi & Herman, 2010; Abedi, Leon, & Mirocha, 2003).

Research on the assessment of ELL students suggests that many different factors could contribute to performance gaps. These factors include parental education level and poverty (Abedi et al., 2003), the challenge of second-language acquisition (Hakuta et al., 2000; Moore & Redd, 2002), and a host of inequitable schooling conditions (Abedi & Herman, 2010; Gándara, Rumberger, Maxwell-Jolly, & Callahan, 2003) in addition to measurement tools that are ill equipped to assess ELL students' skills and abilities.

In a recent study, Abedi and Herman (2010) found that ELL students have less opportunity to learn (OTL) than non-ELL students in the classroom. Results of this study showed that when instructional materials are linguistically complex, ELL students report significantly lower opportunity to learn. Once again, many different factors explain equity in OTL for ELLs, including cultural differences, poverty, mobility, and, more important, language factors. The outcome of the OTL study clearly indicated that ELL students may have difficulty understanding teachers' instruction because of a possible lack of proficiency in English as well as various acculturation issues, including those related to the U.S. system of education.

IMPACT OF LANGUAGE FACTORS ON THE ASSESSMENT OF ELL STUDENTS

ELL students are from different cultural and family backgrounds. Although many of them have immigrant

parents, many were born in the United States and have been granted the same rights to education as their native English-speaking peers (Garcia, Lawton, & Diniz de Figueiredo, 2010). In spite of differences in their personal, family, and academic backgrounds, however, they may share one common characteristics—they have difficulty with the language of instructional and assessment materials to varying degrees. Therefore, content-based assessments that are linguistically complex may render invalid outcomes for these students. Research on the assessment of ELL students clearly shows that assessments that are developed and field tested mainly for native speakers of English may contain unnecessary linguistic complexity that could be problematic for ELL students and could threaten the validity of assessments for these students (e.g., see Abedi, 2010).

To make assessments more accessible to ELL students, the impact of language factors that are unrelated to content should be controlled to the extent possible. Research has identified sources of linguistic complexity that may slow down readers, increasing the likelihood of misinterpretation or adding to the reader's cognitive load, thus making assessment tasks inaccessible to ELL students. Specifically, research has identified 14 major categories of linguistic features that cause difficulty in understanding test questions (Abedi, 2006, 2010; Abedi, Lord, & Plummer, 1997; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006). Abedi (2006) provided detailed explanations of these features and presented the methodology for conducting linguistic modification of test items.

Several studies were conducted on ELL and non-ELL students to examine the impact of linguistic modifications on assessment outcomes for ELL students. These studies focused on two aspects of linguistic modification: effectiveness and validity. In effectiveness research, the impact of linguistic modification on ELL students was examined. Specifically, the research question was whether ELL students showed any significant improvements on assessment outcomes when they were given a linguistically modified version of an assessment. In validity research, the impact of linguistic modification was examined on non-ELL students. The specific research question was whether linguistic modification significantly changed the performance of non-ELLs and, if so, whether the change affected or altered the construct being measured. In the next section of this chapter, we briefly discuss the process of linguistic modification and provide a summary of research on both effectiveness and validity of this approach.

LINGUISTIC MODIFICATION OF TEST ITEMS

The concept of linguistic modification is based on the premise that some test questions may be unnecessarily complex particularly for those who are not quite fluent in the language of the test. Thus, the language structure of assessments should be analyzed to differentiate the language that is the essential component of assessments and is needed to communicate assessment content (construct relevant) from language that is unnecessary and not related to the content being assessed (construct irrelevant; Abedi, 2006, 2010; Abedi et al., 1997; Shaftel et al., 2006). Although differentiating between language relevant and language irrelevant to an assessment question might be a challenging task, the idea can be implemented with help from content and linguistic experts.

By targeting linguistic structures that are unrelated to the focal construct, a linguistic modification approach can help provide more accessible assessments for ELL students as well as for the general student population. This approach involves modifying the linguistic features on any item of an existing assessment that may have unnecessary linguistic complexity. For example, something as simple as replacing unfamiliar vocabulary with familiar vocabulary or switching from passive voice to active voice is considered linguistic modification. The approach sometimes is referred to as linguistic simplification, but it is not necessarily simplifying test items, rather it is reducing linguistic complexity that is not related to the focal construct.

The process starts with identification of sources of unnecessary linguistic complexity. A team of content, linguistic, and measurement experts determines which linguistic features, among the 14 major categories that were identified in research (Abedi, 2006),

are present in the text and how to reduce their impact. Test items can be examined for linguistic complexity using two different approaches. The first approach uses an *analytical* rubric in which each individual item can be rated in each of the 14 linguistic features (discussed in the following bulleted list) and then those feature with highest ratings of complexity can be marked for revisions. The test items can also be rated on overall linguistic complexity using a *holistic* approach. A 5-point Likert-type rating rubric (with 1 indicating little or no linguistic complexity of the item and 5 suggesting that the item was linguistically complex) is used to generate ratings on all items. Items that were rated high (rating of 3 or higher) can then be marked for either deletion or revision. Revisions are often guided by the same experts who identified the level of linguistic complexity, with the goal of all items having a linguistic complexity rating of 2 or below.

Following is a list of the 14 linguistic features (Abedi, 2006; Abedi et al., 1997) that slow down readers and make it difficult for ELL students to understand assessment questions. For some of these features, examples were provided from actual test items that have been used in the past (Abedi, Lord, & Plummer, 1996).

- *Word frequency/familiarity:* Unfamiliar words are harder to understand and process, particularly for nonnative speakers of language. For example, words such as *census* that are unrelated to the mathematic concept in the test items but are used in those items are less familiar than frequently used words such as *pencil*.
- *Word length:* Longer words are more likely to be morphologically complex and difficult for ELL students to comprehend.
- *Sentence length:* The longer the sentence the more linguistically demanding it might be and the more difficult it may be for ELL students to process.
- *Voice of verb phrase:* Passive voice constructions are more difficult to process than active constructions. For example, instead of using "If a marble is taken from the bag," the test question could be written as "If you take a marble from the bag."
- *Length of nominals:* Long nominal compounds are more difficult to interpret. For example, "Last year's class vice president" is more complex than "vice president."
- *Complex question phrases:* Longer question phrases occur with lower frequency and low-frequency expressions are harder to read. For example, "Which is the best approximation of the number?" is more difficult to understand than "Approximately how many?"
- *Comparative structures:* ELL students have difficulty with comparative constructions.
- *Prepositional phrases:* Languages may differ in the ways that motion concepts (e.g., action verbs) are encoded using verbs and prepositions.
- *Sentence and discourse structure:* The syntactic structure for some sentences may be more complex than for others.
- *Subordinate clauses:* Subordinate clauses may contribute more to complexity than coordinate clauses.
- *Conditional clauses:* Separate sentences, rather than conditional "if" clauses, may be easier for ELLs to read and understand. For example, instead of using "If Lee delivers x newspapers" use "Lee delivers x newspapers."
- *Relative clauses:* Relative clauses are less frequent in spoken English than in written English; therefore, ELL students may have had limited exposure to them.
- *Concrete versus abstract presentations:* Information presented in narrative structures tends to be understood better than information presented in expository text.
- *Negation:* Sentences containing negations are harder to comprehend than affirmative sentences.

Additional examples are listed in Appendix 17.1. Following is an example of a test item, from the National Assessment of Educational Progress devised by the U.S. Department of Education (1992), that is deemed to be linguistically complex along with a linguistically modified version of the item (Abedi, Lord, & Plummer, 1997):

> Original Test Item
>
> If Y represents the number of newspapers that Lee delivers each day, which of the following represents the total number of newspapers that Lee delivers in 5 days?

A) 5 + Y
B) 5 × Y
C) Y + 5
D) (Y + Y) x 5

Modified Test Item
Lee delivers Y newspapers each day.
How many newspapers does he deliver in 5 days?

Changes to Test Item
Conditional clause changed to separate sentence
Two relative clauses removed and recast
Long nominals shortened
Question phrase changed from "which of the following represents" to "how many . . ."
Item length changed from 26 to 13 words
Average sentence length changed from 26 to 6.5 words
Number of clauses changed from four to two
Average number of clauses per sentence changed from four to one

PSYCHOMETRIC ISSUES IN THE ASSESSMENT OF ELL STUDENTS

Assessments that are developed and field tested for the mainstream student population may not be applicable to ELL students in the same way they are used for mainstream students. For example, unnecessary linguistic complexity of test items that may not be an issue for non-ELL students may have a serious impact on the performance of ELLs. In this section, we discuss the impact of language factors on psychometric properties of assessments for ELL students. We discuss the classical test theory of measurement underlying many assessments currently used in schools and then elaborate on the applicability of classical test theory in assessments of ELLs (see also Volume 1, Chapter 1, this handbook, and Abedi, 2006).

Comparing Reliability of Assessment Outcomes Between ELL and Non-ELL Students

Classical test theory describes how errors of measurement influence observed scores. On the basis of classical test theory, the purpose of measurement is to estimate the true score (T) of an examinee for a focal construct. This true score, which is not directly observable, can be estimated based on the size of measurement error (E). The larger the measurement error the less accurate is the observed score (X) in estimating the true score (T). Therefore, the first principle of classical test theory is that observed score (X) is the sum of true score (T) and error score (E):

$$X = T + E. \tag{17.1}$$

In classical test theory, reliability ($\rho_{XX'}$) is defined as the ratio of true-score variance (σ_T^2) over observed-score variance (σ_X^2; Allen & Yen, 1979). Observed-score variance (σ_X^2) is the sum of two components: true-score variance (σ_T^2) and error variance (σ_E^2; Allen & Yen, 1979, p. 73; see also Chapter 3, this volume).

$$\rho_{XX'} = \sigma_T^2 / \sigma_X^2. \tag{17.2}$$

The major assumption underlying classical test theory is that the error score is a random variable and is not correlated with the true score; therefore, there is no covariance between T and E components in Equation 17.1. For ELLs, however, this assumption of random measurement error may not hold because unnecessary linguistic complexity as a source of measurement error affects ELL students' performance systematically. Therefore, it is correlated with true scores, and a covariance term (σ_{TS}) is added.

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 + \sigma_s^2 + \sigma_{TS}. \tag{17.3}$$

A comparison between Equations 17.2 and 17.3 reveals that the principle of classical test theory, which is based on the assumption of random measurement error, may not apply to ELL students because unnecessary linguistic complexity of assessments is a source of systematic error (for a more detailed discussion of classical test theory, see Allan & Yen, 1979; as it applies to assessment of ELLs, see Abedi, 2006; Thorndike, 2005). In Equation 17.3, two additional sources of variance can be seen—when

compared with Equation 17.2, a source of systematic error (σ_s^2) due to unnecessary linguistic complexity and a covariance term (σ_{TS}) due to the correlation between the true score and the systematic error source. Thus, assessments may be less reliable for ELL students than for non-ELL students. Studies on the psychometric properties of assessments for ELLs support this hypothesis of lower reliability for ELL students. Results of some of these studies are summarized in the next section.

Comparing Reliability of Assessment Outcomes Between ELL and Non-ELL Students

As discussed, unnecessary linguistic complexity of assessments is a source of systematic error affecting the reliability of tests; therefore, assessments for ELL students suffer from significantly lower reliability coefficients than those for non-ELLs. Research on the assessment of ELL students has clearly demonstrated the trend of lower reliability coefficients for ELL students.

In several studies on assessments and accommodations for ELL students, reliability (internal consistency) coefficients for ELLs and non-ELLs were compared (e.g., see Abedi, 2008; Abedi et al., 2003). Results indicated that the gap in reliability coefficients between ELL and non-ELL students increased as the level of language demand of the assessment increased. For example, studies (Abedi, 2006; Abedi et al., 2003) have found that in math computation, for which language factors might not have as great an influence on performance, the reliability coefficient (alpha) for ELL (.802) was only slightly lower than the alpha for non-ELL students (.898). In English language arts, science, and social science, however, reliability coefficients for ELL assessment outcomes were substantially lower than reliability coefficients for non-ELL students. For non-ELLs, the average reliability coefficient over English language arts, science, and social science was .808 as compared with an average alpha of .603 for ELL students. Once again, as the level of linguistic complexity in assessment content increases, the gap in reliability coefficients between ELL and non-ELL students also increases. The results of analyses based on other data sets nationally were quite consistent with those reported in this chapter and cross-validated our findings.

Comparing Validity of Assessment Outcomes Between ELL and Non-ELL Students

Language factors have a greater impact on the validity of assessments for ELL students (see Volume 1, Chapter 4, this handbook). When content-based assessments such as a mathematics test use complex linguistic structures, it is harder for ELL students to understand the language of the test, and they may not perform well—not because they lack the content knowledge but because of the underlying problem of the language of the question.

As indicated earlier, however, we must distinguish between the language related to the focal construct of assessments (construct relevant), which is "mathematics" in the previous example, and language unrelated to the focal construct, which may interfere in the process of conducting a valid and reliable assessment. The unnecessary linguistic complexity is referred to as a construct-irrelevant source in the previous example. Thus, the higher the level of unnecessary linguistic complexity, the higher the level of impact of construct-irrelevant variance resulting from language factors, and the lower the validity of assessment for ELL students.

FORMATIVE VERSUS SUMMATIVE ASSESSMENTS FOR ELL STUDENTS

Knowledge of ELL student performance would be valuable for teachers to provide more effective instruction for ELL students based on their level of English proficiency. Although end-of-year assessment outcomes, often referred to as summative assessments, could be used to help design assessments and instructions, for individual students, the outcomes of such assessments may be too little too late. Formative assessments, which include a wide range of assessments, including portfolio assessments, teacher-made assessments, and even school- and districtwide assessments that are used formatively, can help teachers of ELL students to understand their academic needs. These assessments, when, administered in a timely manner with applicable instruction,

may be more useful than summative assessments. The two uses of assessments have different separate goals and objectives. For example, Shepard (2000) indicated that formative assessments should be used to improve learning and called for a change in culture for this to effectively happen. Shepard suggested that the social meaning of evaluation should be revised to allow for more interaction between instruction and assessment, a change from the current perception that a single summative yearly test can adequately identify unique student needs.

A major strength of formative assessments that makes them particularly useful in the assessment and instruction of ELLs is that they immediately inform teachers and provide them with valuable information to plan a more productive instructional strategy for ELL students. For example, the results of state assessments may show that ELL students had more difficulty with the extended constructive-response items in mathematics for which a substantial level of writing is expected. Low performance of ELL students in the mathematics test may be due to low levels of the proficiency needed to write their explanations in math problem solving not due to a lack of content knowledge. Thus, information on a student's level of writing would help teachers to focus on areas in which ELL students need help to succeed academically. Thus, the most important characteristic of formative assessments is their immediate ability to inform instruction particularly for ELL students.

Because formative assessments are not part of the state accountability system, little attention has been given to the content and psychometric qualities of such assessments. These assessments are often offered in the form of classroom quizzes and portfolio assessments that are not based on the principles of test development theory and practice. To provide more effective formative assessment outcomes for ELL students, teachers must be provided with the needed resources, including psychometric advice and professional training.

ACCOMMODATIONS USED IN THE ASSESSMENT OF ELL STUDENTS

ELL students are faced with many challenges in their academic careers. The challenge of learning English and at the same time competing with their native English-speaking peers in learning academic concepts in English is enormous. The linguistic complexity of assessment can strongly affect the assessment outcomes of these students. For example, studies suggest that ELL students lag behind their non-ELL peers in almost all areas, but the difference is particularly striking in those areas with a high level of language demand (e.g., see Abedi et al., 2003; Abedi & Lord, 2001).

To provide equal education opportunities for these students, accommodations are provided to offset these challenges without giving ELL students an advantage over students who do not receive accommodations. Test accommodations refer to changes in the test process, in the test itself, or in the test response format.

Most accommodations that are currently offered to ELL students have a limited empirical research base and often use a common-sense approach to choosing accommodations. The concept of accommodations has been introduced in the literature mainly for students with disabilities (see Chapter 18, this volume). To provide a fair assessment for all students, these students need to be accommodated because of their disabilities (see Chapter 27, this volume). For example, students with hearing disabilities should be provided with hearing aids. The situation for ELL students may not be as clear as the case for students with disabilities.

For ELL students, the goal of accommodations is to help with their language needs. Although there are many accommodations used for ELL students across the nation (e.g., see Abedi, Hofstetter, & Lord, 2004; Acosta, Rivera, & Willner, 2008; Elliott, Kratochwill, McKevitt, & Malecki, 2009; Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006; Kieffer, Lesaux, Rivera, & Francis, 2009; Rivera, 2003; Sireci, Li, & Scarpati, 2003; Willner, Rivera, & Acosta, 2008; Wolf et al., 2008), a majority of them do not seem to be relevant for these students (Abedi et al., 2004) and there is no research support for many of these accommodations.

To offer appropriate accommodations for ELL students that could help give a valid picture of what they know and are able to do, several criteria must be considered. These criteria include the following:

(a) effectiveness of accommodations, (b) validity of accommodated outcomes, (c) relevance of accommodations to students' backgrounds, and (d) feasibility.

Effectiveness

Accommodations used for ELL students must be effective in making assessments more accessible for these students. Because the main issues regarding assessment of ELL students are language-related factors, effective accommodations for ELL students should address linguistic issues. There are, however, many accommodations currently used for ELL students that may not be effective or many not even have any relevance to their assessments. For example, of the 73 accommodations that were reported used by states (Rivera, Stansfield, Scaildone, & Sharkey, 2000), only 11 (15%) of them were deemed to be relevant (Cormier, Altman, Shyyan, & Thurlow, 2010; Willner et al., 2008).

Other examples include *one-on-one* and *small-group* testing, which commonly are used for students with disabilities but also have been used for ELLs by the National Assessment of Educational Progress and the states (Abedi & Hejri, 2004). Such accommodations may not help ELL students because they may not be able to provide ELLs with the linguistic assistance they need. Other accommodations such as a customized dictionary and linguistic modifications of assessments could be more relevant and more effective for ELL students. Willner et al. (2008) recommended accommodations that provide direct linguistic support for ELL students rather than those that may not help English learners.

To examine the effectiveness of accommodations for ELL students, these students should be assigned randomly to accommodated (the treatment group) and nonaccommodated (the control group) conditions (Kirk, 1995). Accommodations will be considered as effective if the ELL recipients show improved performance on the assessment when compared with the performance of ELL students in the control group.

Validity of Accommodated Outcomes

Accommodations that are shown to be effective in making assessments more accessible for ELL students should not provide an unfair advantage for these students. That is, a major requirement of accommodations used for ELL students is to "not alter the focal construct." If accommodations do more than what they are supposed to do, then the validity of accommodated assessments would be in jeopardy. Validity of accommodated assessments can be examined using two different approaches: (a) randomized trial experiments and (b) a criterion validation approach.

In the randomized trial approach, non-ELL students (for whom accommodations are not intended) can be assigned to either a treatment group in which they receive an accommodation or to a control group in which students are tested under the standard condition with no accommodations provided. A significantly higher performance by non-ELL students under the accommodated condition may be an indication that the accommodation altered the construct being measured. Thus, accommodation should improve performance of ELL students but not non-ELL students. However, a slight improvement on the performance of non-ELL students may not be sufficient to judge the validity of accommodated assessments. This information must be complemented by data from valid external criteria. Under this approach, scores from both accommodated and nonaccommodated conditions can be compared with external criteria. Major differences in the structural relationship between the performance outcomes and the external criteria across accommodated and nonaccommodated assessments could be an indication of changes in the construct resulting from the provision of accommodations.

Relevance of Accommodations to Students' Backgrounds

ELL students are quite different in many aspects, including their personal, family, and academic backgrounds. Therefore, it would be unrealistic to expect that the same accommodation would work for all or most of these students. Literature has clearly demonstrated that accommodations may work differently for ELL students from varying backgrounds (Abedi et al., 2004). An accommodation that may provide accessible assessment for some students may not do the same for other students; therefore, it is extremely important to

consider students' background variables in assigning accommodations. For example, for ELL students with strong academic backgrounds in their native language who recently enrolled in U.S. schools, the native language testing might be a relevant accommodation, whereas for ELL students who are quite proficient in their native language but have been instructed mostly in English, a linguistically modified version of the assessment in English might produce more valid outcomes. Therefore, variables such as students' levels of proficiency in English and their native language, number of English-only classes in which students participated, number of years in the United States and other similar variables may contribute to more informed decisions about the provision of accommodations.

Feasibility

Accommodations that are used for ELL students must be easy to apply with a minimum level of burden to students and classrooms. This is particularly important in large-scale state and national assessments in which a great number of students are tested and any logistical issues could be a deterrent to administration and scoring of the assessments. For example providing "extended time" is one of the most commonly used accommodations for ELL students. This accommodation, however, could create an undue burden on test administrators, teachers, and schools, requiring additional testing areas and qualified individuals to be present during this extended time. Another accommodation example is "reading test items aloud." If this accommodation has to be provided for a subgroup of students within a classroom, then it could be a burden or a distraction for those students not receiving this accommodation, or it would require additional qualified testing personnel and space to administer this accommodation separately from the rest of the students in the class.

SUMMARY AND RECOMMENDATIONS

Assessment plays a major role in ELL students' academic lives, perhaps more so than for native speakers of English. For ELL students, assessment starts before any classroom instruction. ELL students are tested for their levels of proficiency in English before enrolling in any classes. On the basis of their levels of proficiency in English, a particular program of instruction can be recommended. Thus, ELL students have to go through a formal assessment program twice: once to measure their level of proficiency in English and once to measure their mastery of content knowledge, as is done for all students.

Thus, because of the high-stakes nature of assessments in ELL students' academic lives, these assessments must be carefully examined for any possible sources of threat to their validity. Unfortunately, assessments for ELL students are more affected by extraneous variables, such as cultural and linguistic biases, than those for the mainstream students.

This chapter has presented data that showed a large performance gap between ELL and non-ELL students. The trend of such a performance gap, however, clearly suggests that language factors play a major role in explaining this performance gap. It was demonstrated that the higher the level of language demands in assessments, the larger the performance gap between ELL and non-ELL students.

Although the impact of language factors on the assessment of ELL students is clear and well documented in the literature, controlling for these factors may not be as clear as their impact. Language is a major component of assessments. Except in a few cases (e.g., math computation), it is hard to imagine assessments presented in U.S. schools with no English language context. ELL students must understand the English language to succeed in their academic careers under English-only instruction and assessment settings. Therefore, it is essential to have a good measure of students' levels of English proficiency.

There are, however, major concerns with the validity of ELP assessments that were developed and used before the implementation of NCLB (2002). Assessments that were developed after the implementation of NCLB show major improvements over pre-NCLB assessments, but there are still some major validity concerns regarding these assessments as we elaborated earlier in this chapter. Therefore, initial identification and reclassification of ELL students were affected by the validity issues related to

the ELP assessments. Research literature suggests that some ELL students who were not proficient in English in some of the domains of English proficiency (reading, writing, speaking, and listening) were identified as proficient in English based on the overall composite score of an ELP test. Similarly, ELL students that were prematurely identified as fluent speakers of English were below proficiency in some of the four domains. The problem in initial identification and premature exit has major effects on ELL students' performances in a content-based state assessment and accountability system. When ELL students are not proficient enough in English to participate in English-only instruction and take assessments in English, then they may not be able to present what they know and are able to do.

Other major areas of concern for ELL students include content-based assessments and the challenges that ELL students face with regards to cultural and linguistic aspects of these assessments. State summative (end-of-year) assessments are often created and field tested for mainstream students. ELL students may not be at the level of English proficiency to clearly understand assessment questions and may not have enough writing skills to be able to write their responses to open-ended questions. Therefore, language is an important aspect of these assessments. In this chapter a distinction was made between language factors that are related (construct relevant) and language that is unrelated to the assessment (construct irrelevant). Assessments for ELL students can be greatly improved by identifying language factors that are unnecessary in an assessment and by removing or revising such factors.

One approach to control unnecessary linguistic complexity of assessments for ELL students is to directly target sources of unnecessary linguistic complexity and to revise test items by removing or reducing such sources. This chapter has introduced the concept and application of linguistic modification for such purpose. It also has presented a summary of research that indicated ELL students benefited from linguistically accessible versions of an assessment. When the level of unnecessary linguistic complexity of test items was reduced, the performance gap between ELL and non-ELL students was reduced as well.

The impact of language factors can also be controlled by providing language-based accommodations. Examples of language-based accommodations include English and bilingual dictionaries or glossaries, native language testing, and customized dictionaries. However, literature on accommodations for ELL students suggests that other accommodations that are created and used for students with disabilities (such as writing responses on the test booklet or one-on-one or small-group testing) are being used for ELL students. These accommodations may not be effective in making assessments more accessible for ELL students and may even provide invalid assessment outcomes by altering the focal construct.

In general, to make assessments more accessible for ELL students it is important to first identify sources of threat to the validity of these assessments and then to control for the impact of such sources. This can be accomplished by carefully designing assessments that are valid and fair for ELLs and that at the same time do not provide them any unfair advantages. There are many sources of threat to the validity of content-based, standards-based assessments for all students. Among them are lack of alignment to the state content standards, psychometric issues (low reliability and low validity), item difficulty and bias, and test format. In addition to the sources discussed thus far, a major source of threat to the validity of assessments for ELL students is the linguistic complexity of items that are not related to the construct being measured. Therefore, I recommend identifying and reducing to the extent possible these sources of threat to the validity of assessments.

GUIDELINES FOR CREATING ASSESSMENTS THAT ARE ACCESSIBLE FOR ELL STUDENTS

Because linguistic and cultural factors are most likely to affect the performances of ELL students, these factors must be carefully examined to see whether they are relevant to the purpose of assessments. If they are considered as sources of bias or construct irrelevant, then these factors should be controlled. For example, as explained in this chapter, the first step in examining the impact of language on the assessment of ELL students is to

identify whether the linguistic structure is relevant to assessment questions or is unnecessary or irrelevant to the purpose of the assessment. The next step will be to eliminate or reduce those sources of linguistic complexities that are not related to the focal construct. Following is a set of guidelines to provide fair and valid assessments to ELL students:

1. Examine the linguistic structure of test items by a team of at least three experts (content, measurement, and linguistic) and identify unnecessary linguistic complexity of the items. For this review, use linguistic modification rubrics that are introduced in the literature and briefly discussed in this chapter that are validated for this purpose.
2. Revise test items that are judged to be unnecessarily complex in linguistic structure to reduce the level of linguistic complexity.
3. Ask a team of content and linguistic experts to review the original test items and the revised versions to ensure that the revisions did not alter the construct being measured. That is, revisions on the linguistic structure of the items should be done only to the language unrelated to the content or unnecessary to the assessment.
4. Conduct a series of focus groups and cognitive labs with ELL students and ask them to read test items and discuss sections of the assessments in which they may have difficulty understanding the language.
5. Ask a team of experts in cultural issues and biases to review test items for any sign of cultural issues and revise items accordingly. However, ensure that the cultural revision process does not affect the focal construct.
6. If resources allow, conduct a study by randomly assigning students to a treatment group in which they receive a linguistically modified version of the assessment and a comparison group in which they receive the original version. A performance improvement for ELLs without an impact on the performance of non-ELL may suggest that the assessment has become more linguistically accessible for ELL students.
7. Select accommodations that are relevant to ELL students, that directly address their linguistic needs, and that do not alter the focal construct.

APPENDIX 17.1: EXAMPLES OF TEST ITEMS

PASSIVE VOICE

In active voice, the subject is the one performing an action. In passive voice, the one receiving the action is in the subject position. Often the "actor" is not stated.

> He was given a ticket. vs. The officer gave him a ticket.
> Girls' ears were pierced in infancy. vs. Parents pierced infant girls' ears.

RELATIVE CLAUSES

A relative clause is an embedded clause that provides additional information about the subject or object it follows. Words that lead a relative clause include *that, who*, and *which*. Note: Often *that* is omitted from a relative clause.

> A bag that contains 25 marbles . . . (vs. One bag has 25 marbles.)

When possible, relative clauses should be removed or recast.

PREPOSITIONAL PHRASES

Prepositional phrases work as adjectives or adverbs to modify nouns, pronouns, verbs, adverbs, or adjectives. When they occur before question words, between the subject and the verb, or in strings, they can be especially confusing to English language learners.

> Which of the following is the best approximation of the area of the shaded rectangle in the figure above if the shaded square represents one unit of area?

COMPARATIVE CONSTRUCTION

Comparisons are made using *greater than*, *less than*, *n times as much as*, and *as . . . as* constructions as well as by using certain verbs.

> Jesse saw more mountains than he'd ever seen.
> Who has more marbles than Carlos?

Who has the most?
From which bag is he more likely to pull out a green marble?
If Bill runs 100 yards per hour faster than Peter . . .

Certain verbs imply comparison:

Joan underbid her hand.
Compared to Keith, Jen is short.

Note the reduced clauses that can cause confusion.

John is taller than Mary. (than Mary is.)
The flour doesn't cost as much as the sugar. (as the sugar does.)
Mr. Jones' account is greater than that of Mr. Johnson. (than the account of . . .)

NEGATION

Several types of negative forms are confusing to English language learners.
Proper double negative:

Not all the workers at the factory are not male.
It's not true that all the workers at the factory are not male.

Negative question:

Which student will not finish in time?

Negative terms:

Ted can no longer drive over 40 mph in his truck.

References

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, *33*(1), 4–14. doi:10.3102/0013189X033001004

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Mahwah, NJ: Erlbaum.

Abedi, J. (Ed.). (2007). *English language proficiency assessment in the nation: Current status and future practice*. Davis: University of California.

Abedi, J. (2008). Measuring students' level of English proficiency: Educational significance and assessment requirements. *Educational Assessment*, *13*, 193–214. doi:10.1080/10627190802394404

Abedi, J. (2010). English language learners with disabilities: Classification, assessment, and accommodation issues. *Journal of Applied Testing Technology, 10*(2, article no. 2), 1–30.

Abedi, J., & Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education*, *17*, 371–392. doi:10.1207/s15324818ame1704_3

Abedi, J., & Herman, J. L. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. *Teachers College Record*, *112*, 723–746.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*, 1–28. doi:10.3102/00346543074001001

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of students' language background on content-based assessment: Analyses of extant data.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Hofstetter, C. (1997). *The impact of students' language background variables on their NAEP mathematics performance*. Los Angeles: University of California, Center for the Study of Evaluation.

Acosta, B. D., Rivera, C., & Willner, L. S. (2008). *Best practices in state assessment policies for accommodating English language learners: A Delphi study*. Arlington, VA: George Washington University Center for Equity and Excellence in Education.

Bauman, J., Boals, T., Cranley, E., Gottlieb, M., & Kenyon, D. (2007). Assessing comprehension and communication in English state to state for English language learners (ACCESS for ELLs). In J. Abedi (Ed.), *English language proficiency assessment in the nation: Current status and future practice* (pp. 81–91). Davis: University of California.

Cormier, D. C., Altman, J. R., Shyyan, V., & Thurlow, M. L. (2010). *A summary of the research on the effects of test accommodations: 2007–2008* (Tech. Rep. No. 56). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Del Vecchio, A., & Guerrero, M. (1995). *Handbook of English language proficiency tests*. Albuquerque: New Mexico Highlands University, Evaluation Assistance Center—Western Region.

Elliott, S., Kratochwill, T., McKevitt, B., & Malecki, C. (2009). The effects and perceived consequences of testing accommodations on math and science performance assessments. *School Psychology Quarterly*, *24*, 224–239. doi:10.1037/a0018000

Fast, M., Ferrara, S., & Conrad, D. (2004). *Current efforts in developing English language proficiency measures as required by NCLB: Description of an 18-state collaboration.* Washington, DC: American Institute for Research.

Flannery, M. E. (2009). A new look at America's English language learners. *NEA Today*. Retrieved from http://www.nea.org/home/29160.htm

Francis, D., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments.* Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Gándara, P., Rumberger, R., Maxwell-Jolly, J., & Callahan, R. (2003). English learners in California schools: Unequal resources, unequal outcomes. *Education Policy Analysis Archives*, *11*(36), 1–54.

Garcia, E. E., Lawton, K., & Diniz de Figueiredo (2010). *Assessment of young English language learners in Arizona: Questioning the validity of the state measure of English proficiency*. Retrieved from http://civilrightsproject.ucla.edu/research/k-12-education/language-minority-students/assessment-of-young-english-language-learners-in-arizona-questioning-the-validity-of-the-state-measure-of-english-proficiency/garcia-az-azella-assessment-2010.pdf

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* Santa Barbara: University of California, Linguistic Minority Research Institute.

Kieffer, M., Lesaux, N., Rivera, M., & Francis, D. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *79*, 1168–1201. doi:10.3102/0034654309332490

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Moore, K. A., & Redd, Z. (2002). *Children in poverty: Trends, consequences, and policy options*. Washington, DC: Child Trends.

No Child Left Behind Act of 2001, 20 U.S.C. § 6319 (2002).

Parker, C. E., Louie, J., & O'Dwyer, L. (2009). *New measures of English language proficiency and their relationship to performance on large-scale content assessments* (Issues & Answers Report, REL 2009–No. 066). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

Rivera, C. (2003, June). *State assessment policies for English language learners.* Paper presented at the National Conference on Large-Scale Assessment, San Antonio, TX.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment*, *11*, 105–126. doi:10.1207/s15326977ea1102_2

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14.

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature* (Center for Educational Assessment Research Report No. 485). Amherst: School of Education, University of Massachusetts.

U.S. Department of Education. (1992). *National Assessment of Educational Progress (NAEP)—Mathematics Assessment.* Washington, DC: Author.

Willner, L. S., Rivera, C., & Acosta, B. D. (2008). *Descriptive study of state assessment policies for accommodating English language learners.* Arlington, VA: George Washington University Center for Equity and Excellence in Education.

Wolf, M. K., Herman, J. L., Kim, J., Abedi, J., Leon, S., Griffin, N., & Shin, H. (2008). *Providing validity evidence to improve the assessment of English language learners* (CSE Technical Report No. 738). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Zehler, A. M., Hopstock, P. J., Fleischman, H. L., & Greniuk, C. (1994). *An examination of assessment of limited English proficient students.* Arlington, VA: Development Associates, Special Issues Analysis Center.

CHAPTER 18

CONSIDERATIONS FOR ACHIEVEMENT TESTING OF STUDENTS WITH INDIVIDUAL NEEDS

Rebecca Kopriva and Craig A. Albers

This chapter summarizes the primary issues and topics germane to constructing defensible large-scale academic assessment systems that are accessible for two groups of students, English learners (ELs) and students with disabilities (SwDs). Most of the research and policy to date has occurred within the context of large-scale statewide summative achievement tests designed to be administered yearly to students in Grades 3 to high school, and this material is the bulk of what is summarized in this chapter. Much of what is discussed can be generalized to college-level tests and psychological assessments as well, and to other types of K–12 achievement assessments such as benchmark tests (tests that occur over schools several times during the academic year and are meant to gauge the partial-year performance of students) and standardized formative assessments (these can be course-embedded assessments or other stand-alone assessments designed to support instruction). Although the purposes of the summative, benchmark, and formative achievement assessments differ (they range from evaluation of status knowledge and skills to providing more fine-grain information teachers can use to adapt instruction or clarify misunderstandings), issues of access are usually similar. Because access might be adjusted for the different types of achievement testing, this is noted when possible. Furthermore, of course, the notions about how to properly evaluate the learning of young K–2 ELs and children with special needs is examined to a limited extent. To date, this is an important area that unfortunately has received limited attention.

Although the chapter does not address the assessment of English language proficiency for ELs, emergent skills in the development of English certainly affect how ELs need to be asked questions about academic content and the proficiency and strategies they need to have in demonstrating their understandings using appropriate communication supports (see Chapters 10 and 17, this volume). For SwDs, this chapter does not examine cognitive or emotional assessment considerations (see Chapters 3 and 6, this volume). Also not specifically addressed is the issue of ELs who have also been identified as having a disability (see Chapter 9, this volume).

What does it mean to make academic assessments and especially academic assessment systems accessible? This is a complex question, addressed in some depth by Rigney, Wiley, and Kopriva (2008) and then examined in detail throughout Kopriva (2008). Recently, Winter (2010) has used the lens of test score comparability to highlight various considerations that need to be resolved when access in achievement systems means deciding under what conditions and with what evidence can scores be considered interchangeable when students are taking the same or similar tests under different conditions.

In essence, the goal of achievement tests is to be able to appropriately capture the knowledge and skills intended by the test, and more specifically, by each item, question, or task. Accessibility within this

The contents of this chapter were developed in part by Enhanced Assessment Grants (S368A090029) and (S368A080004) from the U.S. Department of Education and awarded to Rebecca J. Kopriva and Craig A. Albers, respectively.

DOI: 10.1037/14049-018
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

context means that students will be able to properly hear and respond to the intent or target of each question. Any time a question is presented to the student, methods of communication are employed—these methods are meant to facilitate the interaction of the targeted question and response between test maker and test taker. These methods are ancillary to the intended meaning of the task, and sometimes, instead of facilitating the interaction, they act as barriers that wholly or partially prevent the transference of the question or the intended knowledge and skills between the test taker and the test maker. There seem to be three aspects of access that occur within each achievement task. First, the student must have sufficient access to how the meaning and the requirements are conveyed in the task—in other words, what is the task or question asking? Second, for the student to initiate and sustain problem-solving activities relative to the task requirements, students must be able to access their procedural skills and other content assumed by the task and must have the tools necessary to implement the activities. Third, students must be able to access their representation skills commensurate with the representation constraints in the particular task. This means the task or question must be set up in such a way that the student can adequately convey their skills or knowledge to the test maker.

To be able to address accessibility adequately, an understanding of the two populations is crucial. A brief demographic summary and policy overview and an outline of some assessment-related challenges of each of these groups are described in the following sections.

STUDENTS WITH DISABILITIES

Demographics and Assessment-Related Challenges

More than 6.5 million infants, toddlers, children, and youth have been identified as exhibiting specific developmental delays or meet criteria for at least one of the designated disability categories under the Individuals With Disabilities Education Improvement Act of 2004 (IDEA, 2004), depending on their age (U.S. Department of Education, n.d.). Part B of IDEA pertains specifically to schoolchildren and youth. Under Part B of IDEA (2004), children and youth between the ages of 3 and 21, along with their families, are afforded special education and related services upon meeting the criteria of at least one of the following disability categories: mental retardation, hearing impairments (including deafness), speech or language impairments, visual impairments (including blindness), emotional disturbance, orthopedic impairments, autism, traumatic brain injury, other health impairments, and specific learning disabilities. The majority of children and youth between the ages of 6 and 21 who are served under Part B of IDEA are classified as having specific learning disabilities

TABLE 18.1

Disability Categories of Children and Youth, Ages 6 to 21 Years, Served Under IDEA Part B

Disability category	Percentage of students
Specific learning disabilities	43.6
Speech or language impairments	19.2
Other health impairments	10.5
Mental retardation	8.3
Emotional disturbance	7.3
Autism	4.3
Multiple disabilities	2.2
Developmental delay	1.5
Hearing impairments	1.2
Orthopedic Impairments	1.0
Traumatic brain injury	0.4
Visual impairments	0.4
Deaf-blindness	near 0

(43.6%), followed by speech or language impairments and other heath impairments (19.2% and 10.5%, respectively; Data Accountability Center, 2007). The remaining disability categories, along with the percentages of children and youth between the ages of 6 and 21 who are served under IDEA Part B and classified as having such disabilities, appear in Table 18.1 (Data Accountability Center, 2007).

In general, SwDs are included in achievement assessments as they are written and administered, by using accommodations along with the general test forms, through modifications of the general test forms or testing conditions, or through using alternate assessments. The goal of any adaptations is to provide more valid and accurate information about the constructs being measured than would be the case when these students take the general assessments under typical conditions. Test accommodations usually fall under the following categories: presentation accommodations, equipment and materials accommodations, response accommodations, scheduling and timing accommodations, setting accommodations, and linguistic accommodations. In the nomenclature of the educational content testing industry, the term *modifications* of the general test denotes that the modifications affect how the constructs are measured through making changes to test modality, complexity, space, time, language, and possibly other aspects (Poteet, 1990). Alternate assessments are intended to facilitate inclusive assessment for students with significant disabilities and must yield information about students' achievement for purposes of statewide accountability. Ideally, alternate assessments should also provide instructional utility. Each of these is discussed more later in the chapter.

Policy Overview

Two significant pieces of federal legislation require that SwDs be included in standardized assessment programs: the IDEA, for SwDs only; and the current authorization of the federal Elementary and Secondary Education Act (ESEA; No Child Left Behind [NCLB], 2001) legislation, for all students in public schools. Both of these laws were designed to improve the academic achievement of all students through high expectations and high-quality education programs.

ENGLISH LEARNERS

Demographics and Assessment-Related Challenges

Estimates suggest that approximately 25% of all U.S. students currently in schools are ELs (Hernandez, Denton, & Macartney, 2008). Of these, children of immigrants now constitute one fifth of all U.S. school-age children, for which a large majority of the households may be described as *linguistically isolated*, which means that no one in the household age 14 or older speaks English exclusively or very well (Capps et al., 2005). Furthermore, ELs consistently perform below grade level in all content areas. For instance, on the 2005 National Assessment of Educational Progress (NAEP), 46% of EL fourth graders scored "below basic" in mathematics as compared with only 18% of non-ELs; for eighth graders, 71% of ELs scored below basic as compared with 30% of non-ELs (Perie, Grigg, & Dion, 2005); achievement gaps between EL and non-Hispanic White students were 35% in Grade 4 and 50% in Grade 8 (Fry, 2007). ELs are also nearly twice as likely as their native English-speaking peers to drop out of high school (Rumberger, 2006; Silver, Saunders, & Zarate, 2008). Gándara and Rumberger (2009) attributed the higher dropout rate to schools' lack of academic and social supports for ELs beginning well before high school. Callahan and Gándara (2004), among others, have argued that because many ELs and their families are unfamiliar with the U.S. educational system, and because ELs tend to score poorly on language-heavy exams, ELs are often placed in classes that are remedial or do not prepare them for college. As a result many of them fall further and further behind native English-speaking peers with the same academic capacity. All in all, this snapshot begins to reflect why school districts and states feel enormous pressure and often lack of readiness to provide viable schooling for their student bodies (García, Jensen, & Scribner, 2009).

In particular, there appear to be two overarching challenges to appropriately measuring the academic achievement of ELs: (a) proper exposure to challenging content in school and (b) proper evaluations and assessments that minimize their English language limitations and cultural misunderstandings while being

able to still effectively measure their knowledge and skills in subjects such as mathematics and science.

Policy Overview

The Civil Rights Act of 1964 advanced the federal commitment to equity in education and in 1974 *Lau v. Nichols* spelled out the educational rights of language minority students. Up until the 1994 reauthorization of the federal ESEA, however, a great percentage of ELs were exempted from most state and local standardized achievement testing regimens and, with little accompanying accountability oversight, were often schooled separately from their native English-speaking peers. This exclusion changed in 1994 and again in 2001 when the NCLB reauthorization was passed, and states and schools were held accountable for ELs in such a way that teachers were expected to teach, and ELs were expected to learn, the same academic content as their native English speakers. Once this change occurred, researchers and practitioners began to investigate how to make challenging content and assessments accessible for this population.

To design and build accessible achievement assessments several interwoven steps are essential. This is particularly the case when the assessment systems are constructed to measure the same concepts and skills of all test takers, including but not limited to ELs and SwDs. The rest of the chapter outlines and discusses some of the primary issues and solutions that have been found to be effective to date. These and other considerations are discussed in more detail in Kopriva (2008).

BUILDING ACCESSIBLE SYSTEMS: SETTING THE STAGE

Before test construction begins, it is important to put into place procedures associated with participation in test development and methods to ensure that items and forms are accessible.

Participation in Test Development

Adequate participation of EL and SwD experts as well as adequate representation of EL and SwD students should to be built into the development process. Typically, experts with substantive knowledge of these populations have been used primarily in bias reviews, where the charge has been narrow. They have not been included in the planning, item development, and decision-making processes to the same extent that mainstream teachers and content experts have been in recent years. This participation includes involvement throughout the design, construction, and technical phases of development (for a general discussion of test development, see Volume 1, Chapter 9, this handbook). Tasks in which it would be appropriate for them to actively participate can be found in Exhibit 18.1.

Experts who bring the most to the test development process have a deep understanding of content standards, experience with adapting academic teaching environments for these students, and knowledge of their students' strengths and challenges. Examples of relevant expertise of EL experts can be found in Exhibit 18.2.

Just as the diverse perspectives of multiple experts should be included during test development, a full range of SwDs and ELs should be involved in all item and test data collections. It is well known that ELs respond differently based on their proficiency levels and adequate accommodations, and so participating students should range from new arrivals through former English language learners that have successfully transitioned. The same is true for SwDs, whose diverse set of challenges make this broad category extremely heterogeneous. To ensure validity of inferences across all tested students, it will be important

Exhibit 18.1
Expert Participation

- Designing the comprehensive testing system
- Developing test specifications
- Writing and reviewing content items and rubrics that are appropriate for the students with disability and English learner populations
- Providing training to other item writers and developers
- Trying out items in classes
- Evaluating forms for coverage and accessibility
- Making decisions about inclusion or exclusion of items, all testing materials, and administration and response options based on data from pilots, field tests, and other technical data collections
- Scoring, reporting, and making decisions about test use for accountability and program evaluation

Exhibit 18.2
Types of Expertise for English Learner Experts

- Educators from classrooms in which students are learning English as well as grade-level academic content
- Educators from mainstream academic classrooms in which English learners are placed after they have reached a certain level of English proficiency
- Educators working with students who are newly arrived to the United States
- Educators working in classrooms in which the students' primary language (also known as their first language or L1) is the language of instruction or in bilingual (L1 and English) classrooms
- Educators with urban experience and educators with rural experience
- Educators working with migrant students
- Educators who come from the primary language and cultural backgrounds of the students they teach

to determine that all subgroups are responding in similar fashion. As such, enough ELs and SwDs from preidentified strata should be included during piloting to be able to analyze the data by these subgroups as well as the mainstream population. Sireci and Wells (2010) and DePascale (2010a), among others, recommend that the analyses should control for academic ability, and they have demonstrated several ways this might be accomplished.

Building in Procedures to Ensure an Accessible Product

Kopriva (2008) argued that ensuring access is not just a post hoc project. Rather, in addition to including SwD and EL experts and students in range of development, it is important to explicitly consider during planning if general items and forms are accessible, and, if so, for whom. When accommodations will be used, have the proper accommodations been selected and for which EL and SwD student profiles, and is there an oversight mechanism in place to ensure that each student is receiving what they need during the test administration? Are their translations of any forms, are they of high quality, which ELs will they benefit, precisely, and who is still not accommodated adequately? For which SwDs, precisely, are modifications or alternate assessments being considered, and are the plans adequate to satisfy their accessibility to the academic content? Procedures, such as conducting bias reviews and analyzing differential functioning of some items in some subgroups, are seen as ways to address accessibility but are not sufficient by themselves. Finally, to ensure that all questions such as these are adequately addressed, Kopriva maintained that test publishers and consumers should develop a systematic system for checking that the needs of all students are properly considered in test development. This system is briefly outlined in the last section of the chapter.

PRINCIPLED ITEM AND FORM CONSTRUCTION

For many students (e.g., many ELs and some SwDs with literacy, language, or attention or other disabilities), how items in standardized testing systems are typically presented and communicated to the students represent barriers to either accessing what the item is asking or barriers to how the student can show what they know. In these cases, accessible forms with item adaptations need to be created to minimize the barriers and measure intended content at specified cognitive complexity levels.

For most SwDs and all ELs, item adaptations in standardized content testing systems are purposefully designed to measure the same content and cognitive skills as the general test that is given to a majority of the student population. In these cases, if properly constructed, adapted forms and formats and general test forms are intended to yield the same score inferences about the same knowledge and skills. In some cases, however, some SwDs are assessed in large-scale statewide content assessment systems (and some other systems) using modifications and alternate assessment forms and formats that are known to result in different score inferences. The decisions to measure content with modifications or alternate assessments are driven by the nature of the students' disabilities. Both adaptations built to be interchangeable with general test forms and those considered to not be interchangeable are briefly discussed.

Item and Form Adaptations Built to Be Interchangeable

Accessible forms with item adaptations measure the same content, at the same cognitive complexity,

as the items used in the general test, and provide, as necessary, alternative ways for students with particular needs to meaningfully respond. Form and item adaptations may include braille or large print, translations into languages other than English, and plain-language edits in English, with supports as visual aids or access to such tools as manipulatives or picture glossaries. Harnessing computer capabilities increases how meaning might be successfully conveyed, for instance, through animations and interactive aspects, and this methodology also allows for greater flexibility in how students can respond, for instance, by demonstrating their skills, assembling, or modeling (e.g., Kopriva, Gabel, & Cameron, 2011).

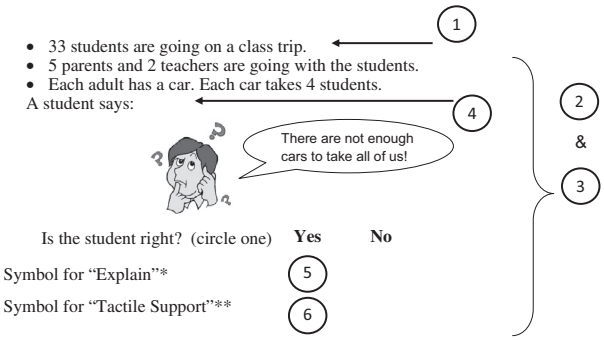
Form and item adaptations designed to measure the same content and cognitive complexity and lead to the same score inferences as the original or base items and forms share certain key development processes, regardless of the nature of the adaptations. First, using a model such as evidence centered design (ECD; Mislevy, 1996), a clear understanding of what the intended inferences are at the item level is essential. Note that explanations of the target content and complexity at the item level are at a finer grain size than is typically required in general tests but are necessary if student scores on the adapted forms are going to be considered interchangeable. Second, particular barriers and then particular item adaptation elements intended to ameliorate or minimize each barrier need to be identified. Third, using techniques that have been found to be successful item adaptations can be designed and built to address one or more particular barrier purpose while still measuring the same content and processes, at the same levels of cognitive complexity. Contextual concerns, formatting, layout of text and nontext elements, attention to language and linguistic structural factors, and continuing adherence to meaning in the base item are always considered.

Adapted forms may have a similar look and feel as the general forms, albeit in a different language, in large print, or with more visuals. With enough documentation, however, it can be argued that forms such as portfolio systems (Barton & Winter, 2010) or computer-interactive animated forms and tasks (see Kopriva, Gabel, et al., 2011) may be used to yield similar score inferences as general forms (on paper or on computer) with multiple-choice and constructed-response items. It remains an ongoing discussion in education assessment as to what kinds of evidence and documentation are needed to successfully make these cross-forms and format claims. For instance, how comparability issues are resolved can influence these adaptations that do not adhere to the given structure of the general test. Readers are encouraged to read the last section in this chapter for a summary of some of these issues as well as Volume 1, Chapter 4, this handbook.

Although readers are directed to other resources for details on how to properly design adapted items (e.g., see Kettler, Elliott, & Beddow, 2009; Kopriva, 2008; Thurlow, Thompson, & Lazarus, 2006), one example of a plain-language edited item in English with formatting and visual supports, and response adaptations, can be seen in Figure 18.1). This item and its base (Figure 18.2) were used in a randomized study of elementary students (Kopriva & Mislevy, 2005) and have been discussed in Kopriva (2008, Chapter 5). Independent ratings found that both items measured the same content and targeted cognitive complexity. The study found that the adapted item measured the targeted mathematics ability significantly better than the base item for many ELs and for struggling native English-speaking readers with no individualized education programs (IEPs), whereas both items similarly measured the targeted ability for more robust native English-speaking readers. These findings suggest that adaptations such as what are shown in Figure 18.1 can be effective in measuring the intended inferences for students that present certain barriers to communication similar to what are illustrated in Figure 18.2. Explanations of some of these adaptations are shown in Figure 18.1.

Item and Form Adaptations Built to Lead to Different Score Inferences

For SwDs, the student's IEP team makes the decision as to how the student will participate in large-scale academic assessments systems. For a relatively small number of these students, these recommendations involve the use of modifications or one of two types of alternate assessment forms.



*This symbol is introduced before the test and is common across all tests this state uses. It means students need to provide an answer and they can do so using words, algorithms, pictures, or other diagrams.

**This symbol is introduced before the test and is common across all tests this state uses. It means that there is an available tool set students can tactilely manipulate to help them solve the problem.

1. Information that is not needed to set the context for the problem has been eliminated, reducing the amount of text.
2. Plain language principles have been applied to the item to reduce the semantic and syntactic complexity of the item. The sentences are shorter and straightforward, using present tense and active voice and reducing the use of prepositional phrases and dependent clauses. A visual is used to illustrate the item. Note that numerals have been used consistently throughout. The translation between a verbal and symbolic representation of a number was considered construct-irrelevant mathematics.
3. The formatting has been arranged to provide maximum access to the problem requirements. Each complete piece of information is presented separately, since, for this item, selecting the appropriate information from among relevant and irrelevant pieces of information was not part of the measurement target. The question is clearly separated from the rest of the text, and the two-stage character of the item, answering the question and explaining the response, is evident.
4. While both the base and the variation assume students are familiar with class trips, which may not be the case in all schools, potential cultural schooling bias has been reduced in the variation by having a student's statement the focus of the question. In some cultures, children are not used to questioning teacher judgments and decisions .
5. Students are given options for how they represent their response.
6. Students are allowed to use manipulative tools to help them represent and solve the problem. The targeted content knowledge and skills do not preclude allowing various methods of representation or solution. The manipulatives provide students who are ELs a way to represent the text that may help them understand the problem situation.

FIGURE 18.1. Adapted item.

At Jefferson Midlands Middle School, the sixth grade students and their teacher are planning a field trip to the state capital at the end of the year. In the morning they will visit the state legislature, and in the afternoon they will go to the zoo.
There are 33 students in sixth grade. Five parents and two teachers will be coming with the students on the trip. Each of the adults has a car that can hold four students. One of the teachers says: "There are not enough cars to take all of us!" Do you agree with the teacher? Explain your answer.

FIGURE 18.2. Grade 4 mathematics item (base).

Modifications. Like the presentation adaptations, modifications provide alternatives to the standardized way test forms are presented to some SwDs to allow some students to better demonstrate their knowledge and skills in learning and testing situations. These adaptations, however, change the testing situation in a way that changes the construct being measured, and hence because of purpose and use, they are generally defined as modifications versus accommodations (Thurlow et al., 2006). Modifications can be made with respect to test modality, complexity, space, time, language, and possibly other aspects (Poteet, 1990). This means that some modifications are form related, whereas some involve other accommodations, which are discussed in the next section. Form-related modifications may involve substituting some of the general test items with modified items that are less cognitively complex, using fewer option choices in multiple-choice questions, or scaffolding constructed-response items that may change the nature of the targeted construct if the constructs involves skills associated with how to approach and conceptualize the problem-solving process.

Specifications for modifications are considered part of NCLB's federal peer review guidance associated with statewide content K–12 assessments used for accountability purposes. Modification constraints may be identified by test publishers, or by users of district or other tests, such as the SAT (formerly known as the Scholastic Achievement Test) or

Graduate Record Examinations (GREs). Consequences are varied and sometimes convoluted, but some consequences include "flagged" test scores—that is, scores that are not allowed for accountability purposes—and scores that count for some purposes but not others (Thurlow et al., 2006).

Alternate assessments. These assessments are intended to facilitate evaluations of academic content knowledge and skills for students with significant disabilities. This type of content assessment was specifically defined within the original NCLB legislation, additional regulations, and nonregulatory guidance; tends to focus on particular purposes and uses; and addresses certain criteria. These assessments are not assumed to be interchangeable with general test forms, but the scores from alternate assessments are used as evidence of statewide accountability for federal purposes. Two forms of alternate assessments for academic achievement tests currently exist. Additionally, Albers (2011) recently developed an alternate assessment form of the ACCESS for ELLs to measure the English language proficiency of ELs who also have significant cognitive disabilities.

Ideally, alternate assessments should provide instructional utility, guiding the development of future instructional goals and learning. Thus, alternate assessments should meet needs for both *required* information (i.e., for accountability) and *desirable* information (i.e., for instructional utility). The assessments must meet the same standards of high technical quality—validity, reliability, accessibility, objectivity, and consistency—expected of other educational tests. In addition, alternate assessments of academic content must have an explicit structure, guidelines for determining which students may participate, clearly defined scoring criteria and procedures, and a report format that communicates student performance in terms of academic achievement standards.

Alternate content achievement assessments for students with the most significant cognitive disabilities. Students with the most significant cognitive disabilities are individuals who

> (a) have disabilities within one or more of the existing categories of disability under the IDEA (e.g., autism, multiple disabilities, traumatic brain injury, etc.), and
>
> (b) whose cognitive impairments may prevent them from attaining grade-level achievement standards, even with the very best instruction. (U.S. Department of Education, 2005, p. 23)

This type of alternate assessment is based on *alternate* achievement standards. These standards are required to be aligned with grade-level content standards, but they are allowed to be reduced in depth, breadth, and complexity. The U.S. Department of Education allows up to 1% of a school district's total number of students to be rated as "proficient" or "advanced" using alternate assessments that are based on alternate achievement standards.

Alternate content achievement assessments for students with other significant disabilities. Additional regulations were established in April 2007 that allow states to report proficient or advanced scores for up to 2% of the total student population using alternate assessments based on *modified* achievement standards. Alternate assessments based on modified achievement standards are directed toward a small group of SwDs who have been determined to be capable of making significant academic progress, but who nonetheless may have significant difficulties in reaching grade-level achievement. In contrast to the *alternate* achievement standards on which students with the most significant cognitive disabilities may be assessed, modified achievement standards are not based on a restricted range of grade-level content. They are based on the same range of grade-level content as the general achievement standards, although the expectations for mastering the grade-level content standards may be less rigorous. An alternate assessment based on modified achievement standards, for example, may include less difficult items based on the same content as the general assessment, include fewer distractors on multiple-choice questions (e.g., three response choices rather than four), or have shorter reading passages than the general assessment.

ADDITIONAL TEST ACCOMMODATIONS

Test accommodations sometimes refer to adaptations to standard testing conditions that fall outside

of what is presented to students. Like the adaptations in items and forms, these changes are based on minimizing particular barriers associated with how this test is taken and are used most often for ELs or SwDs. In the language of current educational testing, the term *accommodations* refers to changes in conditions that do not alter the construct being measured; *modifications* refer to changes in conditions that do alter the construct. For ELs, in addition to adaptations to forms and items, accommodations include tools, administration, and response accommodations (Abedi, 2007). For SwDs, additional assessment accommodations usually fall under one of the following categories: equipment and materials, scheduling and timing, setting, linguistic, and response accommodations (Christensen, Lazarus, Crone, & Thurlow, 2008). It is well known that the permissibility of specific accommodations varies across content area and state or other users. These differences have led to a great deal of confusion particularly when cross-educational agency comparisons are made (Fields, 2008).

The following sections outline some of the most relevant accommodations for both SwDs and ELs. Readers are directed to Kopriva (2008) for a fuller explanation of and research base for EL accommodations and to Thurlow et al. (2006) for more details and additional resources about accommodations for SwDs.

Tools

For SwDs, common equipment and materials accommodations alter the test setting to include certain types of tools and assistive devices, including magnification equipment, amplification equipment, templates, and lighting or acoustics. For ELs, tools often include bilingual, English and picture glossaries, and sometimes manipulatives and other content relevant materials used by students to demonstrate what they know without using much language.

Administration Accommodations

Primary administration accommodations for ELs involve oral English or oral administration of the assessment in their home academic language. Secondary administration accommodations for this population are specified to facilitate the oral administrations or response demonstrations or to deal with extended time requirements, anxiety, or fatigue. They generally include extra time, small group or individual administration, and more frequent breaks. For SwDs, administration accommodations include signing and interpreting directions and reading questions aloud. They might also include scheduling and timing accommodations such as change of time or scheduling of a test, incorporating breaks, testing at a time that is beneficial to the student, and allowing extended time. Administration accommodations might also involve setting—for instance, changing the test location or environment (including individual or small-group administration or administration in a separate room or carrel) and changing the proximity of the student's seat to the test administrator.

Response Accommodations

Response accommodations, as they are defined most often, change the standard conditions around how students can respond to the items presented to them, including the parts of the items presented to them that frame the response environments. Like administration accommodations, these post hoc adaptations do not change the response options or forms of response themselves—any substantive variations that alter the kinds of responses students can reply to are item adaptations. For SwDs, examples of post hoc response accommodations include using a Brailler, writing in test booklets, and using a computer or machine to communicate what the students know (including not only disability-specific technology such as recording puffing or visual cues and then translating these data into a form that can be scored but also using a tape recorder or voice recognition that records the students' audio responses). Communicating responses to a proctor or scribe and allowing this person to bubble, complete, or write the response is another accommodation used for some SwDs.

For ELs, response accommodations have typically involved students responding orally or in text using their home language or code-switching (using both English and their home language). Although these methods seem to be effective for constructed-response

items for ELs (e.g., see Kopriva & Mislevy, 2005), they do not affect multiple-choice or other close-ended items that make up the vast majority of standardized tests. Item variations that use better editing, plain-language text, and visual supports help students with higher English proficiency respond meaningfully to these types of questions, but these methods are often not enough for students with lower English proficiency. Although the multiple-choice questions do not require any additional language to respond, the language of the options is often problematic. For students with little English and for those with little literacy in their home language or first language (L1; in cases in which the test is in L1), correct response to these questions hovers around the guessing level, making this type of item a bad fit for these students (Emick, Wiley, & Kopriva, 2007). An adequate accommodation would be to allow these students to communicate by demonstrating or modeling their knowledge and skills rather than using only English language, but this approach is usually not feasible in high-volume testing. Recent large-scale prototypes of computer-interactive test questions that allow these students to demonstrate, assemble, and model what they know have been found to be very effective (Kopriva & Carr, 2009). Efforts are under way to integrate these advances into large-scale summative and formative testing systems.

ASSIGNMENT OF TEST, FORMS, AND ACCOMMODATION OPTIONS

Even as large-scale content tests may be developed and accommodated to specifically address the needs of ELs and SwDs, if there is no technically rigorous mechanism in place to get the specific methods to the specific students who need them, it is argued that these efforts have little effect. Several researchers who investigate accommodation effectiveness for these populations point out that consistent and appropriate accommodations decision making is critical to the validity of standardized academic testing programs and to the ability to properly use scores to compare student performance across states and districts (e.g., Fuchs, Fuchs, Eaton, Hamlett, Binkley, & Crouch, 2000; Hollenbeck, Tindal, & Almond, 1998; Kopriva, 2008). At the individual level when accommodations decisions are not appropriate to meet the needs of individual students, test results misrepresent their knowledge and skills (Hipolito-Delgado & Kopriva, 2006). At the aggregate level, when accommodations decisions are inconsistent from classroom to classroom or district to district, comparisons across classrooms, districts, and states may be unfair and meaningless (Abedi, 2007; Fields, 2008; Solomon, Jerry, & Lutkus, 2001).

Current guidelines for selecting large-scale and classroom-based accommodations for content testing of SwDs primarily stems from authorizations of federal legislation in IDEA. Regulations or instructions for assigning accommodations to individual ELs, on the other hand, are generally policy based, most often at the state level. The practice for assigning large-scale accommodations for SwDs typically focuses on the role of the IEP. In addition to developing and evaluating each student's learning goals and instructional plans, the IEP addresses the proper test accommodations appropriate for each student at both the classroom and standardized testing levels. Current practices typically used to assign large-scale test accommodations to individual ELs reflect that decisions generally are made by a single person (commonly the student's teacher or the school EL specialist), although some education agencies are beginning to use teams.

In both situations, guidelines tend to offer broad parameters rather than specific guidance for those who must make accommodations decisions. Both individual teachers and teams making accommodations decisions attempt to work within the policies given to them by the federal, state, or local education agency, but these policies generally do not contain specific recommendations for how to address the needs of specific students. Koran, Kopriva, Emick, Monroe, and Garavaglia (2006) found that teacher recommendations, unfortunately, were not statistically different from random assignment of large-scale content testing accommodations to EL students. In the past few years, there have been efforts to tighten the criteria for accommodating SwDs and ELs (e.g., Fields, 2008), but large inconsistencies remain at all levels of schooling.

Research over the past 10 years has continued to confirm that one cannot validly assign accommodations to groups of students based on some broad classification or status (Sireci, Li, & Scarpati, 2003). How then should educators intelligently and reasonably make decisions about accommodations for particular SwDs and ELs when competing tensions of time and accountability are combined with the complexity of needs associated with the heterogeneous populations?

Emerging work suggests that systematic methods of assignment may work better than relying on current policy approaches to assign accommodations for both SwDs and ELs (Fuchs, Fuchs, Eaton, Hamlett, Binkley, & Crouch, 2000; Helwig & Tindal, 2003; Kopriva, Emick, Hidalgo-Delgado, & Cameron, 2007; Russell, 2010; Weston, 2003). Furthermore, researchers present evidence that using systematic methods to match the particular needs and strengths of individual students to specific accommodations may increase validity and be superior to using educator-directed decision making alone.

Elliott and others (e.g., Elliott, Kratochwill, & Gilbertson-Schulte, 1999; Roach & Elliott, 2006) have continued to provide guidance to IEP teams about how to wisely assign large-scale accommodations for SwDs. These researchers have identified key information and student needs that teams should know, critical access skills that are particularly salient for this population, and process factors that influence accommodation decision making. The *Assessment Accommodations Guide* (Elliot et al., 1999) and associated guidance direct IEP team members through the accommodation selection, implementation planning, and documentation processes. The authors encourage members to link any of the 16 key access skills they have identified as being problematic for an individual student to one or more accommodations that specifically minimize interference between conditions and measurement of target skills. These skills represent elements of typical large-scale standardized testing conditions that could pose a problem for SwDs.

The work of Fuchs and colleagues (Fuchs, Fuchs, Eaton, & Hamlett, 2005; Fuchs, Fuchs, Eaton, Hamlett, Binkley, & Crouch, 2000) that is discussed briefly in the next section provides a good example of an empirically grounded systematic method for matching SwDs to particular accommodations based on specific needs. These researchers have found that the test scores of SwDs who receive appropriate accommodations reflect more accurately what others think these students know and that their method is far superior to other methods of accommodation matching. This method is time intensive, however, as determinations are made individually using a trial-and-error process.

More precise, systematic, guidance to identifying needs of SwDs and then recommending some classes of large-scale accommodations for content tests was completed in the past few years by a consortia of states (Christensen, Thurlow, & Wang, 2009). To date, however, there is little research to support that it is consistently better than state-level guidance manuals that several researchers, including Thurlow and colleagues (e.g., Thurlow, Lazarus, Thompson, & Robey, 2002; Thurlow, Moen, & Wiley, 2004; Thurlow et al., 2006) and Rivera and Collum (2006) have argued lead to notoriously inconsistent assignments over locales and students with similar profiles. A recent program that guides IEP teams through a long series of student needs and prior accommodations questions and then leads to suggestions for particular accommodations may be more successful (South Carolina Department of Education, 2010).

Recently, Abedi (2007) and Rivera and Collum (2006) have introduced a hierarchy of choices to the large-scale accommodations for ELs. As discussed in the previous section, the researchers divided relevant accommodations into primary and secondary. The primary accommodations refer to language adaptations, whereas the secondary accommodations refer to conditions that can facilitate or at least not discourage the ability of students to receive the primary accommodations. Rivera and others in her center have compiled guidance to encourage teachers to properly choose accommodations based on guidelines around these primary and secondary accommodation sets, but research about the effectiveness of this advice is as yet unpublished.

To date, only one systematic accommodation matching system for ELs has been published. STELLA, the Selection Taxonomy for English Language Learner Accommodations, is a newly

developed informant system designed to assign individual accommodations for K–12 ELs (e.g., Carr, 2009). It identifies critical variables, collects data, combines the data with standard information regarding how accommodations perform, and then uses a standard series of computerized algorithms. These algorithms have been successfully built, revised, and vetted by experts (Kopriva & Hedgspeth, 2005) and by a team of state specialists (Carr, 2008). The system is designed to utilize the latest information about students that appear to be the most relevant for making accommodation decisions about this population, and it is designed to be customized to accommodate the policies of different states or districts. One of two validation studies found that this system seems to be producing decisions for individual students that better match the data than teacher methods do (Koran et al., 2006), whereas the second study found that ELs who received proper accommodations scored significantly higher than ELs who received incorrect or no accommodations (Kopriva, Emick, Hidalgo-Delgado, & Cameron, 2007).

Guidance manuals that leave the decisions to teachers or IEP teams do not seem to be sufficient to ensure ongoing consistency across locales. The STELLA computer-based matching method looks promising for ELs, whereas trial-and-error methods, or taxonomies of pointed questions for guiding decision makers, may be more relevant for SwDs. Whatever effective processes are used, Solano-Flores and Trumbull (2008) have argued that they must be coupled with consistent implementation procedures and systematic oversight or else the benefits of appropriate matching may be lost. In one hint of how this might be accomplished, Russell (2010) reported that he is currently working with a test publisher to link recommended accommodations directly with computer-based tests for SwDs in such a way that students would receive some of their accommodations electronically as they take their content assessments. Going forward, it will be important to continue to focus on refining consistent data-collection methodologies that isolate the most relevant information for decision making, to attend to the algorithms that are used to convert and combine data, and to attend to the decision-making rules to ensure that they sensitively yield the most salient accommodations for the students who need them.

TECHNICAL CONSIDERATIONS

This section focuses on three interrelated issues that need to be considered to defend the scores from content tests that include variations that address the testing needs of SwDs and ELs: (a) defining and building content assessment systems with proper adaptations, (b) conducting more rigorous research, and (c) constructing empirically based comparability arguments to support when scores should and should not be considered interchangeable.

Defining Content Assessment Adaptations

Kopriva (2008, Chapter 12) described an adapted evidence-centered design model and procedures for test developers to use when building their assessment systems to include variations for these populations. The approach is designed to identify which item and form, tools, administration, and response adaptations to make and use in assessment systems when interchangeable score inferences are intended. Specifically, beginning with (a) identifying intended inferences, the approach recommends methods for (b) identifying the assessment barriers for various profiles of ELs and SwDs, (c) identifying the variations to address the specific barriers for specific profiles, (d) constructing the tests and additional accommodation options to include the variations, and then (e) employing oversight procedures to ensure that all intended adaptations are included. Two additional goals of the approach are to improve the odds that appropriate students receive the proper adaptations and that the proper analyses are completed to support the common inferential claims. As noted, attention should be paid at the item as well as form and post hoc accommodation levels to support construct validity and comparability arguments for students who take the large-scale tests under nonstandard conditions. Often, in our rush to build content assessment systems, the design procedures step is shortchanged, putting the framework for the entire assessment system at risk. Going forward, others as well as Kopriva (e.g., Barton & Winter, 2010; DePascale, 2009; Winter & Gong,

2009) have asked for thoughtful and organized a priori designs to determine which adaptations are to be integrated into or recommended for content testing systems, and how these systematic specifications, implemented properly, might interact with and help defend notions of comparability.

Conducting More Rigorous Research

Several authors, including Abedi (2007), Rivera and Collum (2006), Thurlow et al. (2006), and Tindal and Fuchs (2000), have described the types of item and form, tools, and post hoc accommodations that seem to be useful for SwDs and ELs who have particular profiles of needs and strengths. Yet, research findings that underpin effective links between student profiles and accommodations are often mixed (Kopriva & Lara, 2009), particularly for ELs. To some extent, the fault lies in studies conducted without the proper robust research controls as consumers and test developers rushed to implement accessible agendas quickly. Furthermore, lack of research funding, the heterogeneity of the SwD and EL populations, and the small numbers of many students who fit certain profiles at specific grades or content areas makes research difficult. But methodological flaws, funding constraints, small populations in some cases, and small sample sizes in many of the studies are only part of the story.

Students with disabilities. Tindal and Fuchs (2000) asserted that for accommodation effectiveness to be considered defensible, these accommodations should be based on individual need. The accommodations should benefit only or mainly the students who need the change and not other students. To address the first part of this challenge for SwDs, Fuchs and colleagues (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000) set out to conduct a series of single-subject investigations in which students with certain profiles were given no accommodations and then one or more in sequence, checking at each point to see when the students responded in a manner that the researchers thought was closer to what the student actually knew. If a boost in response rate was evident with one set of accommodations versus another, the researchers concluded that this set was the proper adaptations for this type of student. Over time, many of these SwD profile and accommodation choices were assembled and Fuchs et al. (2005) published the *Dynamic Assessment of Test Accommodations*. These profiles, however, do not nearly cover the range of SwDs.

What has substantially lagged for this population are focal and control group experimental investigations (groups that receive and do not receive accommodations) to address the second point made by Tindal and Fuchs (2000). This type of research is long overdue. Most of the group accommodation studies to date are post hoc and outside well-designed and systematic research agendas, both of which makes the findings less tenable. Although small populations and sample sizes of low-incidence groups make implementing the research challenging, there are only a few investigations of this type with high-incidence populations. It is suggested that experimental studies for low-incidence populations be conducted over years or over sites, using the same experimental design and comparable controls. Even then, most aggregate samples will be small in nature.

Albeit with a single-subject research design, Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000) have laid the groundwork for research on effective accommodations for SwDs. It seems reasonable that with such a heterogeneous population, streams of group experimental research need to be organized to take advantage of the work that has been done and to build a directory of findings that can defend the kinds of test adaptations consumers and advocacy groups are expecting for students with disabilities.

English learners. The accommodations research for ELs has followed a different path. Consistently, meta-analyses (Kieffer, Lesaux, Rivera, & Francis, 2009; Pennock-Roman & Rivera, 2006; Sireci, Li, & Scarpati, 2003) find many questionably designed studies and only a small number of experimental investigations. Results are mixed, even for well-controlled studies. Why is this? Kieffer et al. (2009) have argued that perhaps accurate content inferences from large-scale testing (even with accommodations) are not possible for students with lower language proficiency. Others, however, have argued against this (Boals, Lundberg, & Kopriva, 2012; Grant &

Cook, 2010). Instead, they have asserted that much of the confusion is a casualty of how the groups are defined and how the studies are designed. Most centrally, almost all the studies to date have studied ELs as a monolithic group even though researchers are aware of the diversity of their needs and strengths (Abedi, 2011; Kopriva, 2008; Solano-Flores, 2010). Not surprisingly, focal accommodations that may be effective for one subgroup of ELs is often not useful for another. Several researchers (e.g., Abedi, 2007; Emick & Kopriva, 2006; Emick et al., 2007; Kopriva, Emick, Hidalgo-Delgado, & Cameron, 2007) have argued that level of English language proficiency, at the very least, is a group criterion—students with low English proficiency often need different accommodations than those with higher proficiency. Other characteristics appear to be important as well, such as literacy in their home academic language and how they have been schooled to date (Carr, 2009). Kopriva, Emick, Hidalgo-Delgado, & Cameron (2007) illustrated that English language proficiency and L1 literacy were salient factors in choosing proper accommodation sets.

Furthermore, Kopriva, Cameron, and Gabel (2010) found that providing adequate nontext language rollovers, some L1, and a broader set of response avenues were effective in measuring the science knowledge and skills of ELs with the lowest English proficiency, to such an extent that they scored on par with their native English-speaking peers on this adapted form. Examples of nontext language rollovers include static or animated visuals or halo-highlighting of relevant areas on the screen. To address Tindal and Fuch's (2000) second point, it is interesting to note that the English-speaking peers in this study did not score significantly differently on adapted items than they did on the general test form, whereas EL scores were substantially higher on the variation as compared with the general test. This study is significant because it suggests that large-scale testing *can* be properly accommodated for even students with very little English or literacy skills.

All in all, only a few studies with proper grouping have been completed. Until there is a critical mass, it will be difficult to make definitive judgments about the usefulness of specific accommodations or accommodation sets for ELs with particular profiles.

Comparability Evidence That Supports Decisions About Scores

Advances in cognitive learning theory in the 1990s led to the identification of an expanded set of measurement approaches that seemed to be promising for use in large-scale content assessment. The focus was on comparability of responses *within* approaches—for instance, when rubrics allowed for various ways for students to demonstrate their content knowledge and skills at, say, a Level 3 out of 4 possible points. Over the past 15 years, federal legislation mandating inclusion challenged the status quo that required all students to take tests under standard conditions. This required considering when scores from tests taken under various conditions by SwDs and ELs might be considered interchangeable. Mislevy (1996) argued that the traditional argument for common inferences was made on procedural grounds, leading to the requirement for common products and testing conditions. It is the common inferences, however, that test developers are interested in holding constant, not the procedures per se. As such, Mislevy and others (e.g., see Mislevy, Steinberg, & Almond, 2003) suggested that this conceptual argument should be built on providing adequate evidence about the knowledge and skills of interest, necessary observations, properties of tasks or items designed to elicit the observations, and assessment situations in which students interact with assessment requests. This approach suggests that data may be collected under alternate conditions, as long as there is proper documentation and evidence.

A number of issues relating to validity and comparability are discussed in Kopriva (2008, Chapter 12). Readers are encouraged to review this chapter for more detailed information relevant to making decisions about comparability when different testing conditions are used for different students or when variations in forms are considered. Additionally, related chapters include Volume 1, Chapters 4 and 17, this handbook, and Chapter 17, this volume. What follows is a brief summary of some of the primary points associated with comparability of scores in K–12 academic content testing for ELs and SwDs.

For the purposes of topics discussed in this chapter, comparability of score inferences suggests that

the meaning of the scores is the same or "similar enough," whether students take form A or form B. Comparability includes two steps: First, development methods and empirical evidence need to demonstrate that the forms are measuring equivalent knowledge and skills in the content domain of interest. Second, assuming content equivalence across forms, statistical methods place scores from the forms on a common scale so that comparisons can be made across forms. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) specifies that comparability is easiest to establish when procedures for test development and implementation are identical. In cases in which formats or other conditions vary, other evidence may be required. In all cases, the requisite degree of comparability is defined in terms of context and purpose for which the tests (or forms) are being used. For example, to compare the performance of individual students over time or with other students, a more precise definition of comparability would most likely be required than when equivalence is determined at the prespecified achievement standards level (e.g., basic, proficient, and advanced) with school-level data (i.e., how adequate yearly progress is reported under the NCLB legislation).

In their discussion of steps for achieving comparability when accommodations or nonstandard forms or conditions are used by some students, Haertel and Wiley (2003) focused on the necessity of determining construct target equivalence before any consideration of statistical equivalence is considered. Haertel (2003) differentiated between comparability of objectives for norm-referenced and criterion-referenced tests (like the standards-based assessments being used today) and the implications for these tests when standard and nonstandard forms are considered. He also specified comparability of test administrations under student-to-student comparisons, student-to-achievement level standards comparisons, and student-to-quantifiable criterion measures. Although he suggested that judgments may be the primary avenue when student score inferences are desired at the level of achievement standards, he did not explain how comparability might be accomplished across nonidentical forms when more precision within achievement levels is desired. Winter and Rabinowitz (2006) defined two conditions, both of which they believe are necessary to evaluate comparability. The first of their two conditions is content consistency: At the targeted level of comparison (i.e., to other students or to content standards), do the forms measure the same content information? The second condition is that of score consistency: At the appropriate level of comparison, do the same scores or same performance levels, across forms, reflect the same level of abilities? They emphasized that adequate evidence is essential to document the equivalence at each level.

Winter and Rabinowitz (2006) argued that only after an adequate level of content equivalence has been established, should score equivalence methodologies be implemented. Mislevy (1993) differentiated three levels of linking academic forms—equating, calibration, and social moderation. Feuer, Holland, Green, Bertenthal, and Hemphill (1999) extended these methods to include equating, calibration, projection, and moderation. In both taxonomies, the methods are hierarchically arranged in terms of assumptions and precision of inferences arising from the results. That is, assumptions and precision are relaxed as approaches move from equating to moderation. Mislevy's top level, equating, is the typical approach developers and researchers use to produce comparable forms. This level supports the finest distinctions in ability gradations. The methods evaluate test comparability through the use of statistical procedures in which comparisons are made directly between performances across forms. In addition to building forms from the same blueprints, the goal of content equivalence has typically been achieved by using identical development procedures, materials, and testing conditions. It is not clear whether this method of securing score consistency or equivalence is sufficient for producing forms with comparable inferences when forms include both standard and nonstandard versions. To date, it does not appear that other score equivalence methods have been considered to handle forms from the same blueprints for cases in which presentation or testing conditions are not identical.

Calibration, Mislevy's (1993) second level of linking, assumes that a well-specified content domain is the common frame of reference (e.g., content standards), and it evaluates the degree to which each form reflects that referent. The forms are compared with one another only indirectly. In development, calibration seems to assume that the forms do not use the same test specifications but substantively refer to the same referent throughout construction. As such, part of demonstrating adequate calibration will revolve around a quantified criterion estimate of the referent or detailed judgments from expert raters about the degree of alignment of the items on forms with the corresponding aspects of the target reference domain. Depending on the precision of analysis, comparisons may be made at the level of achievement standards, and possibly at some designations within the standards as well. Social moderation is the third level of linking in which the referent is levels of performance (e.g., the academic achievement levels). Here, forms are not designed to be parallel, and a looser form of expert judgment than calibration is utilized to evaluate how well the combined cognitive demand or other aspects of the content domain on each form supports comparability of performances. Empirical evaluations of linking in this case could compare the judgments about the forms, the subscore or total score performance of students, and perhaps some other independent judgments about the target abilities of the students. This level produces the least specific degree of comparability.

In 2006, federal funding was provided to continue to wrestle with comparability issues in statewide K–12 content testing when tests are given under varying conditions (Bazemore, 2006). This project used three general guiding questions to focus its work:

1. What do we want when we want score comparability?
2. What do we mean when we say comparability for a given purpose?
3. How can we evaluate comparability?

As the project unfolded, the questions were interpreted as follows (Winter, 2010a): The first question focused on the inferential achievement claims the test evidence can support. Documentation of the design of test development and subsequent procedures used to produce the evidence will need to pass scrutiny and should be evaluated through the lenses of appropriateness for capturing the knowledge and skills of particular students in particular situations. In other words, the evidence is viable if the logic of the overall design and individual procedural expectations can be argued through precedence to address and minimize alternative explanations, the implementation of the design and expectations are consistent with what is intended, and the implementation of the procedures themselves are implemented systematically and in a defensible manner. It is probable that test score evidence will come, to a reasonably large degree, from viable evidence at the item level, including systematic protocols and procedures associated with how some items responses are scored.

The second question addresses the level of comparability that is desired. For instance, is comparability focused at the achievement standards level (a series of about four school performance levels required under ESEA legislation for public school accountability), individual scale score level, or single cut-point level? This level of comparability makes a difference for the kinds of evidence that need to be collected, with the overall expectation that scores from both the general test and variation should be considered interchangeable enough and without flags. If the focus is one cut-point score (as in pass–fail), the whole assessment exercise should be focused on producing performances correctly identified on one side of the cutoff or the other. If more than one but a discrete number of scores are of interest, then interchangeability documentation needs to address the same question at each of the relevant scores. When raw or scale scores are the focus, then evidence needs to demonstrate that multiple scores along a continuous range are measuring similar enough knowledge or skills for the students taking each form. The third question focuses on how to analyze the evidence and make decisions about whether the evidence is good enough. Winter (2010a) has argued that there must be sufficient evidence of both content and construct equivalence and score equivalence, and that sufficient evidence along these lines form the basis of how one might judge the comparability of given materials for a given purpose.

Content and construct equivalence. The definition of content and construct equivalence as Winter (2010a) has applied this term focuses on grounding the score inferences across all variations considered to be interchangeable, in documented judgments and empirical evidence of the intended constructs being measured. Content and construct equivalence also involves ensuring that the user can have confidence that the meanings are the same (or the same enough). This aspect of equivalence reflects the analysis of evidence produced to defend the first question. Kopriva (2008) has argued that for equating, both adequate judgments and sufficiently rigorous empirical validation of the content and construct target equivalence need to undergird claims of score equivalence. Some elements of empirical support should supplement the judgments of content and construct equivalence at the other linking levels as well.

To make judgments about content and construct equivalence for ELs and SwDs, development methods designed to promote correspondence across items are referred to in earlier sections of this chapter. For instance, backtranslation and simultaneous (across languages) test development methods are important for ELs when the focus is content and construct equivalence between English and translated forms (e.g., see Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; see also Chapter 26, this volume). Alignment analysis and other types of independent expert evaluations are examples of judgments that are also needed. For instance, judgment review procedures of item variations targeted to the same test specifications include those used by Gierl and Khaliq (2001), and alignment reviews such as those utilized by Webb, Alt, Ely, and Vesperman (2005) could be used to evaluate forms. Some researchers have used judgment techniques to evaluate the content similarity and comparability of cognitive complexity levels in items across forms (e.g., Kopriva, Wiley, & Winter, 2007; O'Neil, Sireci, & Huff, 2003–2004). Williamson, Bejar, and Sax (2004) explored how and when comparability might be affected when open-ended responses were scored using human and automated graders. After analyzing the judges' criteria for assigning scores and how the judges appeared to draw conclusions, they discussed how internal discrepancies might be handled to mitigate differences that arise.

For forms not built to be parallel, content experts may review the bodies of knowledge and skills assessed across forms and determine whether the same level of content complexity exists in both. Quality of judgments can be evaluated using statistics such as the confidence-interval approach proposed by Penfield and Miller (2004) or those used in standard setting. Approaches defined in multidimensional scaling or other similar content validation methods may also be appropriate to use in some situations (e.g., Haertel, 2003; Sireci, 1998).

Score equivalence. Score equivalence focuses on documenting that the scores from the variation and the general forms are behaving in the same way (or the same enough) for students with similar abilities. Evidence that will be analyzed for this aspect of equivalence comes from data that are appropriate to address the second question—that is, to defend the claims of interchangeability at the level of purpose. Examples of construct equivalent evidence that need to be evaluated include same-standards coverage, similar criteria for inclusion, similar judgments about relevant cognitive demands, and similar internal structure. Given evidence that data are drawn from samples for which similarity of student groups on important variables can be documented (e.g., through random assignment, control for differentiated ability using recognized methods, or evidence of similar distributions on relevant background variables), score equivalent evidence includes similar enough proficiency percentages, similar enough score distributions, similar enough structure of forms, and similar enough rank order. How "enough" is defined is a key part of determining score equivalence for particular uses and purposes.

When standard and nonstandard forms are designed to be parallel, statistical equating is the preferred approach to obtaining score equivalence because of the precision with which the equated scores can differentiate performance. Explanations of equating methods are outside the scope of this chapter. Basically, texts such as Kolen and Brennan (1995) have summarized a number of methods that collect test data from either equivalent or nonequivalent groups. When the distributions of groups are

considered to be equivalent (i.e., through random selection), linear equating and equipercentile techniques have been derived and similar techniques have been developed to handle nonequivalent groups as well. For most of these methods, data are collected on different forms or tests for the different groups. Most companies have moved to using item response theory techniques with nonequivalent groups to produce equated scores. This approach specifies that a subset of common items are given to the different groups as well as items that vary across groups. Item parameters for the common items are set across groups and maximum likelihood techniques are used to estimate the parameters for the rest of the items.

A number of different types of calibration and social moderation procedures have been identified in the past few years. Most often these look like modified standard-setting procedures, such as the Modified-Angoff and Bookmark methods (e.g., see Brennan, 2006; Cizek & Burg, 2005). See Feuer, Holland, Green, Bertenthal, and Hemphill (1999); Kopriva (2008); and Mislevy (1993), among others, for a more detailed discussion of this topic.

The Bazemore project yielded reports on a number of different studies that addressed methodological aspects of both content and construct equivalence and score equivalence (Winter, 2010b). These included findings about propensity score matching as an option to repeated measures methodology (Lottridge, Nicewander, & Mitzel, 2010); video versus paper-and-pencil forms for ELs of different languages, and factor analyses and multidimensional scaling to evaluate form differences (Sireci & Wells, 2010); simultaneous item codevelopment and anchor item methods (DePascale, 2010a); an evaluation model for a modified achievement test method for selected SwDs (DePascale, 2010b); and qualitative comparisons and judgment-based methods for students with significant cognitive disabilities and some ELs (Barton & Winter, 2010). Literature reviews reported in Winter (2010b) also summarized papers about paper-based versus computer-based modes of administration, translations and English forms for ELs, and plain-language editing.

CONCLUSION

Within the past 20 years, fields of cognitive psychology, educational practice, and accountability policy have each emphasized the diversity of the U.S. student population. One implication of this insight is that variable content testing methods predicated on making the same inferences about the content abilities of students tested under conditions that are designed to minimize challenges that are irrelevant to the academic knowledge and skills under scrutiny are most likely here to stay. To date, however, the assessment specialists, including academics and researchers, practitioners in state and local educational agencies, and test publishers, who design, implement, and interpret results from these testing systems, are far from unanimous as to how to construct, use, and defend content testing systems that include these variations. This chapter has summarized how several of the relevant aspects of this important topic have been conceptualized to date and has reviewed empirical work that has been completed to investigate the issues. Together this literature forms a body of work that defines the complexity of the topic and points to considerations for the future. The strength of the work is that it points to a multidimensional framework associated with effective academic measurement when systematic construct-irrelevant needs of certain populations would otherwise confuse the inferences about achievement that can be defensibly drawn for these groups. Although several next steps have been identified throughout the chapter, central to most of these is a focus on increasing the number of well-designed and thoughtfully implemented empirical studies to confirm or dispute hypothesized solutions. The findings from these investigations would, in turn, provide additional nuance, rigor, and direction to the discussion considered here.

References

Abedi, J. (2007, April). *Research on the validity of accommodations to help states with providing accommodations for ELL students.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Abedi, J. (2011). Assessing English language learners: Critical issues. In M. del Rosario Basterra, E.

Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment* (pp. 49–71). New York, NY: Routledge.

Albers, C. A. (2011). *The alternate ACCESS for ELLs with significant cognitive disabilities*. Madison: University of Wisconsin, Wisconsin Center for Education Research.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Barton, K. E., & Winter, P. C. (2010). Evaluating the comparability of scores from an alternative format. In P. C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations* (pp. 95–104). Washington, DC: Council of Chief State School Officers.

Bazemore, M. (2006). *Funded proposal to the U.S. Department of Education, Office of Elementary and Secondary Education. Strengthening the comparability and technical quality of test variations.* Raleigh, NC: Department of Education.

Boals, T., Lundberg, T., & Kopriva, R. J. (2012). *Performance, policy and politics: The place for well-reasoned accommodations in large-scale testing and accountability practices.* Manuscript submitted for publication.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: Praeger.

Callahan, R. M., & Gándara, P. (2004). On nobody's agenda: Improving English-language learners' access to higher education. In S. Michael (Ed.), *Teaching immigrant and second-language students: Strategies for success* (pp. 107–127). Cambridge, MA: Harvard Education Press.

Capps, R., Fix, M., Murray, J., Ost, J., Passel, J., & Herwantoro, S. (2005). *The new demography of America's schools: Immigration and the No Child Left Behind Act*. Washington, DC: Urban Institute.

Carr, T. G. (2008). Report of state specialists' review of STELLA decision-making algorithms. In T. Siskind (Ed.), *Final report for the AVAD grant* (pp. 2–6). Columbia: South Carolina Department of Education.

Carr, T. G. (2009, April). *It's about time: Matching English learners and the ways they take tests by using an online tool to properly address individual needs.* Paper presented at the National Council on Measurement in Education, San Diego, CA.

Christensen, L., Lazarus, S., Crone, M., & Thurlow, M. (2008). *2007 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 69). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Christensen, L. L., Thurlow, M. L., & Wang, T. (2009). *Improving accommodations outcomes: Monitoring instructional and assessment accommodations for students with disabilities*. Washington, DC: Council of Chief State School Officers.

Cizek, G., & Burg, S. (2005). *Addressing test anxiety in a high-stakes environment: Strategies for classrooms and schools.* Thousand Oaks, CA: Sage.

Data Accountability Center. (2007). *Table 1–11: Children and students served under IDEA, part B, in the U.S. and outlying areas, by age group, year, and disability category-Fall 1998 through fall 2007.* Retrieved from https://www.ideadata.org/arc_toc9.asp#partbCC

DePascale, C. (2009). *Formative reform: Purposeful planning for the next generation of assessment and accountability systems*. Retrieved from http://www.nciea.org/publications/RILS_FormativeReform_CD09.pdf

DePascale, C. A. (2010a). Evaluating linguistic modifications: An examination of the comparability of a plain English mathematics assessment. In P. C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations* (pp. 69–94). Washington, DC: Council of Chief State School Officers.

DePascale, C. A. (2010b). Modified tests for modified achievement standards: Examining the comparability of scores to the general test. In P. C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations* (pp. 105–118). Washington, DC: Council of Chief State School Officers.

Elliott, S. N., Kratochwill, T. R., & Gilbertson-Schulte, A. (1999). *Assessment accommodations guide*. Monterey, CA: CTB McGraw-Hill.

Emick, J., & Kopriva, R. J. (2006, April). *Access-enhanced item development: A summary*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Emick, J., Wiley, D. E., & Kopriva, R. J. (2007, April). *The validity of large-scale assessment scores for ELLs under optimal testing conditions: Does validity vary by language proficiency?* Paper presented at the American Educational Research Association Annual Meeting, Chicago, IL.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*, 301–321. doi:10.1207/s15324818ame1703_4

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures:*

Equivalence and linkage among educational tests. Washington, DC: National Academy Press.

Fields, R. (2008). *Inclusion of special populations in the national assessment: A review of relevant laws and regulations*. Washington, DC: Report to National Assessment Governing Board.

Fry, R. (2007). *How far behind in math and reading are English language learners?* Washington, DC: Pew Hispanic Center.

Fuchs, L. S., Fuchs, D., Eaton, S. B., & Hamlett, C. B. (2005). *Dynamic assessment of test accommodations*. San Antonio, TX: Psychological Corporation.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. B., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children*, *67*, 67–81.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. B., & Karns, K. M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review*, *29*, 65–85.

Gándara, P., & Rumberger, R. W. (2009). Immigration, language, and education: How does language policy structure opportunity? *Teachers College Record*, *111*, 750–782.

García, E. E., Jensen, B. T., & Scribner, K. P. (2009). The demographic imperative. *Educational Leadership, 66*(7), 8–13.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement, 38*, 164–187. doi:10.1111/j.1745-3984.2001.tb01121.x

Grant, R., & Cook, H. G. (2010). *Relating English language proficiency to mathematics performance through structural equation modeling*. Unpublished manuscript.

Haertel, E. H. (2003, April). *Evidentiary argument and comparability of scores from standard versus nonstandard test administrations*. Paper presented at the National Council of Measurement in Education, Chicago, IL.

Haertel, E. H., & Wiley, D. E. (2003, August). *Comparability issues when scores are produced under varying testing conditions*. Paper presented at the psychometric Conference on Validity and Accommodations, University of Maryland, College Park.

Helwig, R., & Tindal, G. (2003). An experimental analysis of accommodation decisions on large-scale mathematics tests. *Exceptional Children*, *69*, 211–225.

Hernandez, D. J., Denton, N. A., & Macartney, S. E. (2008). Children in immigrant families: Looking to America's future. *Social Policy Report, 22*(3), 3–22.

Hipolito-Delgado, C., & Kopriva, R. J. (2006, April). *Assessing the Selection Taxonomy for English Language Learner Accommodations (STELLA)*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *Journal of Special Education*, *32*, 175–183. doi:10.1177/002246699803200304

Individuals With Disabilities Education Improvement Act of 2004, Pub. L. 108–446, 118 Stat. 2647 (codified at 20 U.S. C. § 1400 *et seq.*).

Kieffer, M. J., Lesaux, N. R., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *79*, 1168–1201. doi:10.3102/0034654309332490

Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modified achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education*, *84*, 529–551.

Kolen, M. K., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York, NY: Routledge.

Kopriva, R. J. (2008). *Improving testing for English language learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments*. New York, NY: Routledge.

Kopriva, R. J., Cameron, C., & Gabel, D. (2010). *ONPAR findings from the 2008 science study for 4th and 8th graders*. Report from the University of Wisconsin Research Series. Retrieved from http://www.ONPAR.US

Kopriva, R. J., & Carr, T. G. (2009, June). *Building comparable computer-based science items for English learners: Results and insights from the ONPAR Project*. Paper presented at the National Conference on Student Assessment Annual Meeting, Los Angeles, CA.

Kopriva, R. J., Emick, J., Hildago-Delgado, C. P., & Cameron, C. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision-making on scores for ELLs. *Educational Measurement: Issues and Practice*, *26*(3), 11–20. doi:10.1111/j.1745-3992.2007.00097.x

Kopriva, R. J., Gabel, D., & Cameron, C. (2011, April). *Designing dynamic and interactive assessments for English learners which directly measure targeted science constructs*. Presentation at the American Education Research Association Annual Meeting, New Orleans, LA.

Kopriva, R. J., & Hedgspeth, C. (2005). *Technical manual, Selection Taxonomy for English Language Learner Accommodation (STELLA) decision-making systems*. Madison: University of Wisconsin,

Center for the Study of Assessment Validity and Evaluation.

Kopriva, R. J., & Lara, J. (2009). *Looking back and looking forward: Inclusion of all students in NAEP, U.S.'s National Assessment of Educational Progress* [commissioned paper]. Washington, DC: National Assessment Governing Board.

Kopriva, R. J., & Mislevy, R. (2005). *Final research report of the Valid Assessment of English Language Learners Project* (C-SAVE Rep. No. 259). Madison: University of Wisconsin, Center for the Study of Assessment Validity and Evaluation.

Kopriva, R. J., Wiley, D. E., & Winter, P. C. (2007). *Analyzing skill complexity using specially constructed test scores.* Unpublished manuscript.

Koran, J., Kopriva, R. J., Emick, J., Monroe, J. R., & Garavaglia, D. (2006, April). *Teacher and multisource computerized approaches for making individualized test accommodation decisions for English language learners.* Paper presented at the National Council of Measurement in Education, San Francisco, CA.

Lottridge, S. M., Nicewander, W. A., & Mitzel, H. C. (2010). Summary of the online comparability studies for one state's end-of-course program. In P. C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations* (pp. 13–32). Washington, DC: Council of Chief State School Officers.

Mislevy, R. J. (1993). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379–416. doi:10.1111/j.1745-3984.1996.tb00498.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–62. doi:10.1207/S15366359MEA0101_02

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002). Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/107–110.pdf

O'Neil, T., Sireci, S. G., & Huff, K. L. (2003–2004). Evaluating the consistency of test content across two successive administrations of a state-mandated assessment. *Educational Assessment, 9*, 129–151.

Penfield, R., & Miller, J. M. (2004). Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. *Applied Measurement in Education, 17*, 359–370. doi:10.1207/s15324818ame1704_2

Pennock-Roman, M., & Rivera, C. (2006, April). *A review of test accommodations for ELLs: Effect sizes in reducing the mean achievement gap.* Paper presented at the American Educational Research Association, San Francisco, CA.

Perie, M., Grigg, W., & Dion, G. (2005). *The nation's report card: Mathematics 2005* (NCES 2006-453). Washington, DC: U.S. Government Printing Office.

Poteet, J. (1990). The what and how of modified assessment techniques. *Diagnostique*, *16*, 58–60.

Rigney, S., Wiley, D. E., & Kopriva, R. J. (2008). The past as preparation: Measurement, public policy, and implications for access. In R. J. Kopriva (Ed.), *Improving testing for English language learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments* (pp. 37–64). New York, NY: Routledge.

Rivera, C., & Collum, E. (2006). *A national review of state assessment policy and practice for English language learners.* Hillsdale, NJ: Erlbaum.

Roach, A. T., & Elliott, S. N. (2006). The influence of access to the general education curriculum on the alternate assessment performance of students with significant cognitive disabilities. *Educational Evaluation and Policy Analysis*, *28*, 181–194.

Rumberger, R. W. (2006). *The growth of the linguistic minority population in the U.S. and California, 1980–2005.* Santa Barbara: University of California Linguistic Minority Research Institute.

Russell, M. (2010, October). *APIP.* Paper presented at the Reidy Interactive Lecture Series, Cambridge, MA.

Silver, D., Saunders, M., & Zarate, E. (2008). *What factors predict high school graduation in the Los Angeles Unified School District?* Santa Barbara: University of California.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299–321. doi:10.1207/s15326977ea0504_2

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature* (Center for Educational Assessment Research Report No. 485). Amherst: University of Massachusetts, School of Education.

Sireci, S. G., & Wells, C. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations* (pp. 33–68). Washington, DC: Council of Chief State School Officers.

Solano-Flores, G. (2010). Assessing the cultural validity of assessment practices: An introduction. In M. del Rosario Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment* (pp. 3–21). New York, NY: Routledge.

Solano-Flores, G., & Trumbull, E. (2008). In which language should English language learners be tested? In R. J. Kopriva (Ed.), *Improving testing for English lan-*

guage learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments (pp. 169–200). New York, NY: Routledge.

Solomon, C., Jerry, L., & Lutkus, A. (2001). *The nation's report card: State mathematics 2000* (NCES 2001-519). Washington, DC: National Assessment of Educational Progress.

South Carolina Department of Education. (2010). *AccSelPro.* Retrieved from http://www.accselpro.org

Thurlow, M., Lazarus, S., Thompson, S., & Robey, S. (2002). *State participation and accommodation policies for students with disabilities: 2001 update.* Minneapolis, MN: National Center on Educational Outcomes.

Thurlow, M., Moen, R., & Wiley, H. I. (2004). *Biennial performance reports: 2002–2003 state assessment data.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., Thompson, S., & Lazarus, S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 653–673). Mahwah, NJ: Erlbaum.

Tindal, G., & Fuchs, L. (2000). *A summary of research on test changes: An empirical basis for defining accommodations.* Lexington, KY: Mid-South Regional Resource Center.

U.S. Department of Education. (2005). *Alternate achievement standards for students with the most significant cognitive disabilities: Nonregulatory guidance.* Washington, DC: Author.

U.S. Department of Education. (n.d.). *Building the legacy: IDEA 2004.* Retrieved from http://idea.ed.gov

Webb, N. L., Alt, M., Ely, R., & Vesperman, B. (2005, September). *The WEB alignment tool: Development, refinement, and dissemination.* Paper presented at the Council of Chief State School Officers' State Collaborative on Assessment and Student Standards, Technical Issues in Large-Scale Assessment Collaborative.

Weston, T. J. (2003). *NAEP validity studies: The validity of oral accommodations in testing* (NCES 2003-06). Washington, DC: National Center for Education Statistics.

Williamson, D. M., Bejar, I. I., & Sax, A. (2004). Automated tools for subject matter expert evaluation of automated scoring. *Applied Measurement in Education*, *17*, 323–357. doi:10.1207/s15324818ame1704_1

Winter, P., & Rabinowitz, S. (2006, May). *Two conditions necessary for comparability of achievement levels from test variations.* Paper presented at the Title I Peer Review Technical Assistance Meeting, U.S. Department of Education, Washington DC.

Winter, P. C. (2010a). Comparability and test variations. In P. C. Winter (Ed.), *Evaluating the comparability of results from educational achievement test variations* (pp. 1–12). Washington, DC: Council of Chief State School Officers.

Winter, P. C. (Ed.). (2010b). *Evaluating the comparability of results from educational achievement test variations.* Washington, DC: Council of Chief State School Officers.

Winter, P. C., & Gong, B. (2009, August). *Setting the stage.* Paper presented at the Test Score Comparability and Validity: Preparing for the Future of Assessment Dissemination Meeting, Washington, DC.

CHAPTER 19

LICENSURE AND CERTIFICATION TESTING

Mark R. Raymond and Richard M. Luecht

The purpose of credentialing is to assure the public that individuals who practice an occupation or profession[1] have met certain standards (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). The two most common forms of credentialing are licensure and certification. *Licensure* is the "process by which an agency of the government grants permission to an individual to engage in a given occupation" (U.S. Department of Health, Education, and Welfare, Public Health Services, 1977, p. 4). Licensure laws define the scope of activities a licensed individual may perform and those activities that an unlicensed individual is prohibited from performing. The purpose of licensure is to protect the public from unqualified practitioners. In contrast, *certification* usually refers to the process by which an organization grants recognition to an individual who has voluntarily met the requirements established or adopted by the granting organization (Shimberg, 1981, 1982). Individuals who meet the stated requirements are issued a credential, such as Certified Emergency Nurse (CEN) or Cisco Certified Network Professional (CCNP).

The classic distinction between licensure and certification is that licensure is mandated by law with the intent of controlling what a person can and cannot do, whereas certification is often voluntary and regulates the title a person uses. This distinction is evident in law and medicine for which it is common for individuals to first acquire a license for general practice and later earn a certificate attesting to their qualifications in an area of specialization. In practice, however, the term *certification* has taken on numerous meanings and does not always conform to this classic distinction. Sometimes certification carries the legal authority of licensure, as it does in states where only certified teachers are permitted by law to teach in a public school. In other instances, certification tests may be used by organizations for personnel selection or promotion. In yet other instances, certificates may be issued to those who pass a several-hour course addressing some narrow range of skills, such as project management or basic cardiac life support. Indeed, credentials are issued for a variety of purposes to acknowledge various levels of accomplishment. This chapter focuses on licensure and certification in their classic meaning—to assure the public that credentialed individuals have met certain standards. This is important because it has implications for test design and validation strategies.

Credentialing exams date back to at least 1100 B.C.E., when civil servants of the Chan dynasty were required to demonstrate competence in a variety of endeavors ranging from arithmetic to horsemanship (DuBois, 1970; Wainer, 1990). The use of examinations to separate qualified from unqualified individuals continues in the 21st cen-

[1]The *Standards for Educational and Psychological Testing* and other publications use the term *credentialing* to refer to both licensure and certification. This chapter follows that convention except in cases in which the distinction is important. Also, the term *profession* as used in this chapter refers to occupation, trade, or profession.

DOI: 10.1037/14049-019
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

tury for such fields as accounting, automotive service, information technology, law, medicine, and teaching. Many less well-known professions that also require qualifying examinations include mastectomy fitting, crane operation, retinal angiography, and irrigation systems installation. Although some credentialing programs periodically test a handful of examinees, others, such as the national exam for nurse licensure, administer well over 100,000 exams each year. The outcomes of these tests are not inconsequential for the examinees who take them or for the public that such tests are intended to protect. In spite of its long history and social significance, credentialing is one of the newer and less differentiated specialties within psychological assessment. It did not receive notable attention in the *Standards for Educational and Psychological Testing* until the 1985 edition, which devoted a chapter to it. The 1999 edition of the *Standards* then subsumed the topic within its chapter on employment testing. Similarly, the influential handbook, *Educational Measurement* (Brennan, 2006) did not take up the topic until its fourth edition with a chapter by Clauser, Margolis, and Case (2006). Most contributions to credentialing appear to have been made by those with training in industrial–organizational psychology or educational measurement, but mostly the latter. Even though the recognition of credentialing as a measurement specialty is relatively recent, it has made significant contributions to the larger field of assessment, including large-scale computer-adaptive testing, small-sample equating solutions, and the development of simulated performance assessments, to name a few accomplishments.

Credentialing examinations are built on the same principles covered elsewhere in this handbook that provide the foundation for other types of examinations. Therefore, this chapter focuses primarily on issues that are either unique to or particularly challenging for credentialing and on testing practices to which credentialing has made notable contributions. In the next section, approaches to determining test content with an emphasis on job analysis and systematic processes for building detailed test specifications are presented. The item types and assessment formats commonly used by credentialing programs are next described. In the final section, psychometric and practical challenges of particular significance to credentialing are identified, such as validation, retesting, and security issues.

DECIDING WHAT TO TEST

Early in the life of a credentialing program some authority needs to articulate the purpose of the credential. The stated purpose of the credential has implications for exam content and format. Consider the two following statements:

(a) to verify that licensed individuals have acquired the knowledge and cognitive skills required to effectively carry out the job activities typically performed by entry-level personnel.
(b) to verify that licensed individuals can competently perform the job activities typically required of entry-level personnel.

These two purposes should lead to very different credentialing exams. Statement (a) indicates an emphasis on knowledge and cognitive skills, and a written examination consisting of multiple-choice questions (MCQs) may be sufficient for verifying acquisition of the required proficiencies. The presence of the term *competently* in statement (b) implies the assessment of a broader range of skills encompassing the psychomotor and affective domains. Assessment of such skills may require a practical exam or some other type of performance test. Armed with a statement of purpose, the credentialing agency then determines the specific content and format of the assessment by carrying out a job analysis and related activities (AERA et al., 1999).

Job and Practice Analysis

Given that the purpose of credentialing is to determine an individual's readiness to work in a profession, it is important that such tests sample the knowledge, skills, and abilities (KSAs) actually required in typical work settings. This first requires conducting a job or practice analysis to identify the responsibilities of those who work in a profession (AERA et al., 1999; Kane, 1982; Raymond & Neustel, 2006; Shimberg, 1981; see also Volume 1, Chapter 23, this handbook). Job analysis for credentialing

differs in many ways from job analysis in industrial–organizational psychology (Raymond, 2001). Job analysis for credentialing requires a broader sample of respondents because credentials are often national in scope and cover a multitude of work settings and employment positions. Registered nurses, for example, are employed in small community hospitals, large teaching hospitals, health maintenance organization clinics, public schools, private practice, or any number of different settings, and the job analysis must capture this diversity. The fact that professionals often function autonomously further complicates the design of job analysis projects for credentialing exams (e.g., questionnaire content, sampling). As one example, the report describing the monumental job analysis of licensed psychologists completed in the early 1980s devotes several pages just to the sampling plan (Rosenfeld, Shimberg, & Thornton, 1983). The differences in test content between personnel selection tests and credentialing exams also gives rise to differences in job analysis procedures. Selection tests legitimately address any KSA that predicts job success, whereas credentialing exams emphasize those KSAs with direct implications for public protection (Kane, 1982; Shimberg, 1981). Furthermore, whereas personnel selection tests typically address personality and cognitive ability constructs familiar to measurement specialists (e.g., extroversion, spatial orientation, verbal ability), the content of credentialing exams is job specific and may not be familiar to measurement personnel. Consequently, it is necessary to rely extensively on subject matter experts from within the credentialed profession. These experts contribute to the project by translating job activities into detailed KSA requirements. As discussed later in the chapter, criterion-related validation strategies are seldom practical in the credentialing environment, placing increased emphasis on content-oriented strategies. Therefore, a job analysis report may be the single most important piece of evidence supporting the validity of scores on credentialing exams. Given the elevated role of the job analysis, it is all the more important that it be rigorously conducted and well documented.

The terms *practice analysis* and *practice studies* appear frequently in the literature on credentialing (AERA et al., 1999; Kane, 1997) and can be viewed as specific instances of the more general term *job analysis* (Raymond, 2001). Although it is still common for agencies to conduct an informal practice analysis by holding a 2-day meeting of subject matter experts to produce a test blueprint, over time, more credentialing agencies have recognized the benefits of employing formal methods of practice analysis. Although numerous approaches to practice analysis exist, the most common method is the task inventory questionnaire and its variations (Raymond & Neustel, 2006). Mail-out and Internet-based questionnaires have several advantages over other methods of job analysis. First, they provide an efficient way to collect large amounts of job-related information from hundreds or thousands of individuals in numerous work settings. This is especially important for credentialing examinations, which are intended to indicate an individual's readiness for a wide range of activities in a variety of settings (Kane, 1982). A second benefit is that responses to a task inventory questionnaire are conducive to many types of useful statistical analyses, including multivariate procedures that might be used to organize tasks into a meaningful model of practice or to identify subspecialties (Raymond, 2001; Rosenfeld et al., 1983). Third, data from task inventories also lend themselves to the development of test plans based on empirical methods (Kane, 1997). A notable limitation of the task inventory method is that some types of complex information are difficult to collect in questionnaire format (e.g., task criticality). Furthermore, most professions require unobservable cognitive skills and professional judgment; task inventories that emphasize discrete, observable tasks may overlook these cognitive skills (LaDuca, 1994). Although it is common for task inventory surveys to include KSA requirements (Tannenbaum & Wesley, 1993), KSA ratings run the risk of exhibiting positive response bias (Morgeson, Klinger, Mayfield, Ferrara, & Campion, 2004; Raymond, 2001). It seems all too easy for incumbents to rate a KSA as being important even though it is not required to perform any specific job activity.

Specifying Test Content

Most practice analyses encompass two separate but related stages (Harvey, 1991). The first stage is

primarily a descriptive activity; simply put, its purpose is to document job responsibilities. The second stage is inferential; its goal is to identify the knowledge and skills required to effectively carry out those responsibilities. Making the link between job responsibilities and KSA requirements necessarily involves the judgments of content experts. The product of these judgments is a detailed test outline or set of content specifications. The traditional way to translate the results of practice analysis into a content outline is for a panel of experts to meet a few times over a period of several months and work with a measurement specialist to specify the KSAs important enough to test. A limitation is that expert panels can represent a biased sample of the professional community—those who are recognized for their significant accomplishments. Such individuals may identify difficult or cutting-edge topics as relevant, when in fact they are not (Morgeson & Campion, 1997). It is helpful to verify the work of expert panels through replication with other panels or by surveying the professional community. Some authors recommend the use of a highly structured linkage activity to help expert panels map knowledge and skill domains onto job activities (Landy, 1988; Raymond & Neustel, 2006; Wang, Schnipke, & Witt, 2005). The linkage activity can help ensure that only job-relevant topics are retained for inclusion in the test specifications, and the results can be useful for deriving topic weights. Knowledge elicitation and similar procedures (e.g., think-aloud protocols) also provide a way to systematically determine knowledge requirements (Cooke, 1999).

One of the more complicated test development activities is to create the necessary link between the intended measurement constructs, the items or assessment tasks, and the ensuing score scales. One could argue that this linkage should be made explicit and carefully integrated into the definition of the constructs, the design of items, and the scoring and scaling practices. This is not typically the case, however. It seems curious that most serious efforts to validate score interpretations are often executed long after the test forms have been designed, administered, and scored. The result is an ad hoc approach to score interpretation and use (Bejar, Braun, & Tannenbaum, 2007). What is needed is an integrated way to conceptualize test design, item development, psychometric analysis, and score interpretation. The call for new approaches to test development certainly is not new. Messick (1994) argued for an approach to what he termed *construct-centered assessment design*. A similar sentiment has been echoed by proponents of *evidenced-centered design* (ECD; Mislevy, 1994). The ECD methodology underscores the central role of evidentiary reasoning. As such, ECD frames test development as a self-validating process in which the desired measurement inferences, called *claims*, are carefully articulated, including statements of observed performance and situations (conditions of measurement) to be used as evidence of the claims. The notion of a "conceptual assessment framework" is central to ECD and guides the development of test items and assessment tasks, assembly of test forms, and the psychometric processing needed to maintain the chain of inferences from claims to scores (Mislevy & Riconscente, 2006). The conceptual assessment framework consists of six models or processes, the first three of which have implications for determining test content (Tannenbaum, Robustelli, & Baron, 2008). The first of these models, called the *proficiency model*, articulates the purpose of the test, identifies the target audience, and establishes the claims that the credentialing agency wishes to make on the basis of test scores. These claims are statements about the examinee's proficiency formulated to suit the purpose and the audience of the assessment. A high-level claim might be that candidates who pass a certification examination for certified food managers have demonstrated the knowledge required to practice safely without supervision. A more specific claim might be that the certified food handler recognizes the maximum storage life of perishable foods under refrigerated and deep-freeze conditions. Such claims are products of the typical practice analysis. The second stage in the process, referred to as the *evidence model*, specifies the types of measurable behaviors to be elicited that support the claims to be made about candidates. A third model, the *task model*, then specifies the item types and exam formats to be used. Assessment tasks are selected with the explicit purpose of providing examinees an opportunity to produce the evidence

that has been defined as required. Tannenbaum et al. (2008) nicely illustrated the use of ECD to inform the transition from practice analysis to test specifications.

Another comprehensive framework for designing, producing, and developing large numbers of assessment tasks in service of specific score interpretations is called *assessment engineering* (AE; Luecht, 2006a). Assessment engineering is about devising replicable and scalable assessments by linking construct definitions to test specification, and item design to scoring and score–scale interpretations. Assessment engineering differs from more traditional approaches to test design and development in four fundamental ways. First, cognitively oriented task models guide task design and item development, rather than conventional content blueprints. These task models directly integrate the content aspects of the task and the cognitive task requirements with psychometric test design objectives such as measurement information targets. Second, empirically validated assessment task templates, data models, and scoring evaluators are created for each task model to control factors that contribute to item difficulty, dimensionality, and undesirable sources of measurement variance. This iterative task design process has the ultimate goal of generating many items that yield consistent and appropriate measurement properties. Third, automated test assembly (ATA) procedures are employed to build assessments to exacting specifications. Because the content and cognitive task components and psychometric measurement targets are integrated by design into the test development process, the supply (the items) can more easily match the demand (the test specifications). Finally, pursuant to scoring and reporting, psychometric models are employed in a confirmatory—versus exploratory—way to assess fit to the test data to an intended underlying structure of traits or proficiency classes.

Assessment engineering incorporates five integrated processes aimed at generating consistent and interpretable measurement scales: (a) construct mapping and the design of evidence models, (b) design of task models and measurement blueprints, (c) development of task templates, (d) item production and test assembly, and (e) scaling and reporting. The discussion here focuses on the first three steps.

Construct mapping and design of evidence models. A salient feature of AE is that the assessment design and task development process is front-loaded to ensure that the interpretations of the score scale are generated long before test items or assessment tasks are designed and piloted. An important step is the development of performance level descriptors, which are a set of behavioral objectives or claims that the credentialing agencies wishes to make on the basis of test performance. Although performance-level descriptors have gained wide use in educational testing, they are seldom employed by credentialing agencies. Their use helps ensure that assessment tasks focus on the construct and claims of interest and further allows the credentialing agency to articulate its performance expectations before exam assembly and administration. Although traditional standard setting can still occur later, its role is merely to assign a numerical score to the previously defined expectations (see Chapter 22, this volume).

Construct design is an iterative process that articulates the ordered, proficiency-related claims that we wish to make at different levels of one or more construct-based scales. Wilson (2005) described this process as *construct mapping* and noted that

> its most important features are that there is (a) a coherent and substantive definition of the content for the construct; and (b) an idea that the construct is composed of an underlying continuum—this can be manifested in two ways—an ordering of the respondents and/or an ordering of item responses. (p. 26)

A construct map is therefore a design specification for the ordered knowledge and skill claims that we wish to make about some proficiency of interest. Defining a skill domain such as "competence at accounting principles" is not a construct definition because it does not specify any claims in terms of cognitive skills and knowledge structures, nor does

it imply any type of ordering of skills along some scale. Within the AE framework, a construct is typically defined as a cognitively oriented, measurable proficiency scale that provides useful information about some collection of attributes. It is certainly possible also to describe noncognitive constructs, however, such as personality traits, interest traits, or attitudinal traits. Construct maps are comparable to, but more explicit than, the ability requirements scales (Fleishman & Quaintance, 1984), which are behaviorally anchored rating scales sometimes used for test development in personnel selection.

Designing the task model. A task model is a precise specification of what examinees are expected to do, including the knowledge objects, properties, and relations with which they are supposed to work as well as the context and any auxiliary tools or resources available to complete the task. There are at least as many task models as there would typically be items on the test. In other words, each task model is a complete measurement specification for a class of items that behave similarly in a psychometric sense—a behaviorally anchored prescription for what we want examinees to do or demonstrate. There are three salient components of a task model: (a) the function or activity to be performed (i.e., the procedural knowledge we wish to measure); (b) the context for the assessment task, including specifications for the number of informational knowledge objects, the properties of those objects (qualifiers, etc.), the types of relations among the objects, specifications for the semantic difficulty, salient symbolic abstractions, and verbal load; and (c) auxiliary resources (e.g., software; assistive technologies such as dictionaries or calculators; restrictions such as time limits). Task models take the place of vague test outlines with a more complete and integrated set of specifications that operationally define the construct. They serve as a specification for the subsequent development of item banks and replace the usual content codes found in most test blueprints (Luecht, 2006a; Luecht, Dallas, & Steed, 2010).

Developing task templates. One feature of AE rests in the representation of each task model by multiple templates, each of which is capable of generating many items with similar operating characteristics. If the templates are well engineered, they will provide principled measurement that facilitates test assembly and psychometric processing. Figure 19.1 illustrates a construct map, task model map, multiple associated templates for each task model, and multiple items for each template. Everything is aligned to the construct claims, consistent with ECD and other notions of evidence-based validation. The surface plot graphics at the far right further suggest a hierarchical quality control and calibration system that allows any uncertainty about the item operating characteristics to be explicitly considered. It is apparent that task templates not only inform item writing and test assembly but also guide standard setting, scaling, and score reporting. Although AE is a recent development, some credentialing organizations have produced useful research in this area (Luecht et al., 2010).

HOW TO TEST IT: ASSESSMENT TASKS AND EXAM FORMATS

Early credentialing exams for civil servants and other professionals often consisted of an oral exam in combination with a work sample or some other type of performance-based exam (Clauser et al., 2006; Wainer, 1990). Currently, the MCQ appears to be by far the most common format. A recent survey of 125 certification agencies representing a range of trades, occupations, and professions indicated that 97% of all agencies use multiple-choice examinations (Knapp & Knapp, 2007). However, 38% of credentialing agencies also relied on some type of performance assessment ranging from oral exams to computer-based simulations.

Although MCQs provide an efficient way to assess the cognitive knowledge and skills and have desirable psychometric properties, they typically lack the fidelity to directly assess the competencies required in practice, such as communication skills or hair-cutting technique (Leigh et al., 2007). Performance tests are intended to address this limitation by presenting the examinee with tasks that more closely approximate those encountered in the real-world work setting (Baker, O'Neil, & Linn, 1993; Swanson, Norman, & Linn, 1995; see also Volume 1, Chapter 20, this handbook). The following

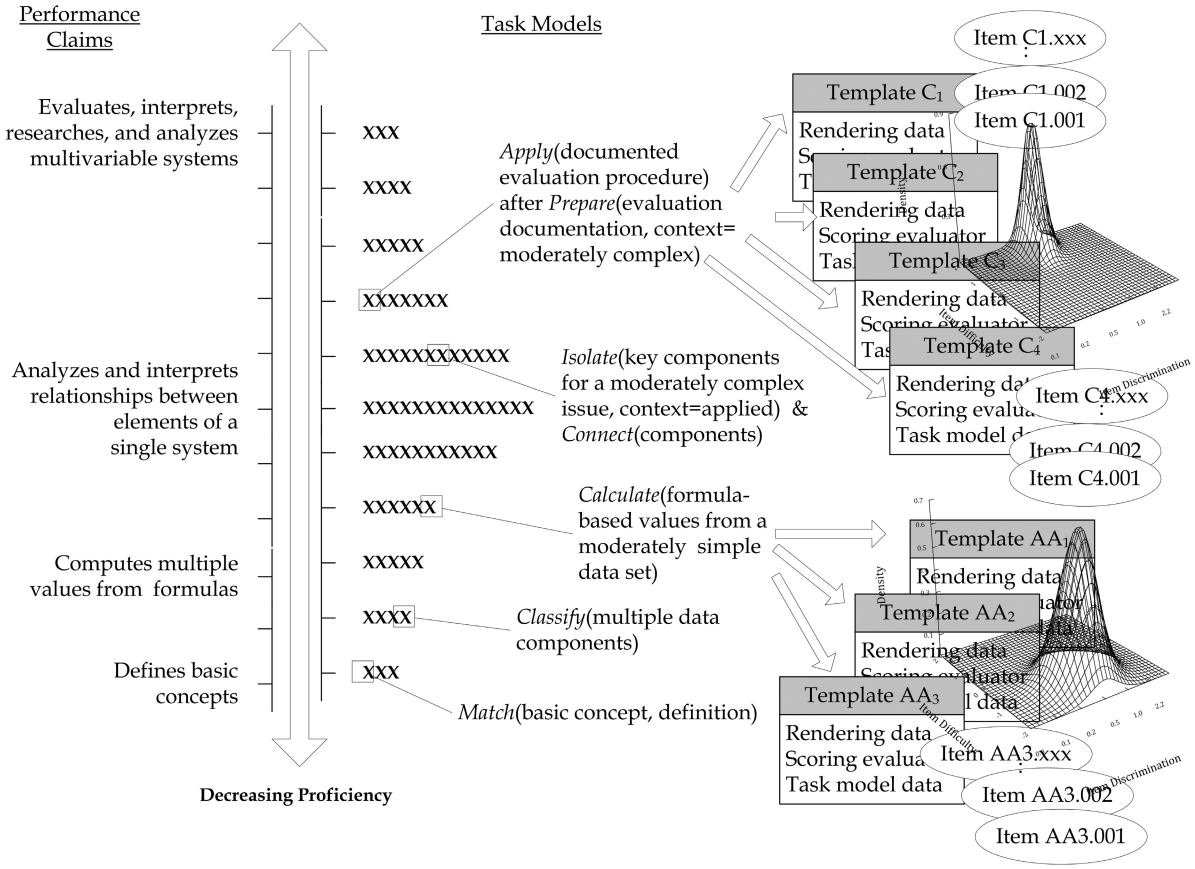


FIGURE 19.1. Assessment engineering overview of a construct map, task model map, templates, and items.

text reviews the more common formats used in credentialing, including MCQs, essay exams, oral exams, work samples, and computer-based simulations.

MCQs and Other Selected-Response Formats

The strengths and limitations of the MCQ format are well documented and have been thoroughly debated over the years (Haladyna, 2004; see also Volume 1, Chapter 18, this handbook). Part of that debate has centered on the optimal number of options or choices. Although an abundance of research demonstrates that three options maximize test reliability for a given amount of testing time (Grier, 1975; Rodriguez, 2005; Tversky, 1964), the MCQs used by most credentialing organizations consist of four or five options (Cizek, Robinson, & O'Day, 1998). That tradition is likely to continue. One factor that contributes to the appeal of the MCQ format for credentialing exams is its efficiency—many items can be administered in a relatively short amount of time. Credentialing exams are typically much longer than tests used in education and personnel selection. An exam consisting of 100 items is considered short, and it is not uncommon for exams to have 300 or more items. The primary reason for this is to allow for broad content coverage and prevent construct underrepresentation (Messick, 1989). The more diverse the content domain, the longer the exam should be. For example, the exams taken by prospective engineers cover chemical, civil, electrical, environmental, industrial, mechanical, and other engineering disciplines (National Conference of Examiners for Engineering and Surveying, 1999). Given the numerous engineering disciplines, and the fact that the test specifications for any single discipline span several pages, it is not surprising that examinees are required to answer 360 items across two exams. Diversity of content also affects reliability, especially if different content domains measure somewhat different traits. Items from aptitude tests

(e.g., verbal analogies) will generally correlate more highly than items on a credentialing exam, and this internal consistency allows even relatively short exams to produce scores with adequate reliability. In contrast, items on a credentialing exam addressing the multiple content domains required of engineers will likely have lower interitem correlations—likely induced by various degrees of multidimensionality—and therefore require additional test items to achieve acceptable levels of reliability. Furthermore, many credentialing exams assess minimal competence and include test items that most examinees will have been expected to master. Therefore, credentialing test items tend to be easier and less discriminating than the types of items that populate personnel selection or college admissions tests. These multiple factors point to the need for longer exams for most credentialing programs.

Computers and Testing

Computer-adaptive testing (CAT) has proven to be a useful tool for limiting the length of credentialing exams (Wainer, 1990; see also Volume 1, Chapter 10, this handbook). Items for a CAT are selected for each examinee based on his or her responses to previous items, with incorrect responses resulting in the presentation of easier items and correct responses resulting in more difficult items. Because the item-selection algorithm creates a test that is tailored to each examinee's level of proficiency, examinees do not spend time responding to items that are too easy or too difficult. Each successive item is chosen in a manner that continually reduces the uncertainty about the examinee's level of proficiency or classification status (i.e., passer or failer). Reliable scores can be obtained after administering a fraction of the items required of a traditional fixed-length test (Drasgow, Luecht, & Bennett, 2006; Wainer, 1990). The item-selection algorithms developed for use in credentialing typically ensure that the test form seen by an examinee is balanced in accordance with the content requirements. This is an important capability, particularly for certification testing for which there is interest in verifying knowledge across multiple content domains.

Although CAT had been the topic of numerous journal articles, books, and conference sessions through the 1970s and 1980s, it did not gain traction on a large-scale basis until the 1990s. At that time the National Council of State Boards of Nursing (NCSBN) initiated an investigation into the operational use of CAT for the licensure examinations taken by registered nurses (Zara, 1999). In 1994 the NCSBN adaptive testing program went live, permitting a substantial reduction in exam length. For example, the NCLEX-RN, the examination for licensed registered nurses, had been reduced from a 300-item fixed-length exam booklet to an adaptive test ranging from 90 to 180 test items. Examinees with scores some distance from the passing score would be done testing after 90 items, whereas examinees near the cut score would require longer exams to achieve a specified level of precision in the pass–fail decision. Several credentialing agencies have since implemented CAT, including the American Institute of Certified Public Accountants, the Council on Certification of Nurse Anesthetists, and Microsoft, to name a few.

A particular type of CAT, called a computerized adaptive multistage test, makes it relatively easy to balance content and still achieve all of the intended measurement efficiencies associated with CAT (Luecht & Nungester, 1998). With a multistage adaptive testing, short multi-item modules called testlets are constructed at different levels of difficulty and assembled into units called panels. The panels can be constructed to control test content, item exposure, and measurement precision. Because the panels are constructed beforehand, this framework also allows for subject matter expert review before test administration, which is not possible with CAT. Computers have revolutionized the development and administration of nonadaptive, fixed-length tests. As noted, most test specifications include a content component that represents the required content, item types, and statistical specifications that emphasize target properties for each test form. Having human beings build test forms that satisfy both the content and statistical specifications is usually labor intensive and still may fail to produce forms that exhibit parallel content and statistical properties. The use of ATA has changed that practice by using mathematical algorithms and heuristics to almost perfectly reconcile content and

statistical specifications, building numerous parallel test forms in seconds (Luecht, 2006b). Although large-scale testing programs routinely apply ATA, they typically rely on editors and subject matter experts to review and approve each test form.

Perhaps most important, computers have improved the variety of item formats available and the realism of the stimuli presented. The traditional MCQ has been supplemented by other types of selected response formats, including drag-and-drop items that require examinees to move options into correct locations or proper sequence, hot-spot items that require examinees to indicate a region of interest on an illustration (e.g., location of left ventricle), and items with multiple correct answers. The use of audio and video presentation of assessment tasks or stimuli is also becoming more commonplace. For example, medical licensure exams now include audio recordings of heart sounds as well as videos of physical exam procedures, physician–patient communications, and other types of patient encounters. Such formats allow for increased fidelity while maintaining the efficiency of the MCQ format. We later describe examinations that use computers to simulate the work environment for selected professions.

Essay Examinations

Survey results reported by Knapp and Knapp (2007) and Shimberg (1981) suggested that 10% to 15% of credentialing agencies use essays or a short-answer format, presumably in conjunction with the MCQ format. The rationale for using essays depends on the profession and the skills targeted for assessment. Although writing proficiency might be viewed as a source of construct-irrelevant variance for some credentialing exams, there are professions for which writing skill is a critical feature of the job. For example, the Multistate Essay Examination is one of three exams taken by aspiring lawyers in many states. The essay portion is intended to test the examinee's ability to identify the legal issues presented in the case, separate relevant from irrelevant information, and identify fundamental legal principles. In addition, the essay component also assesses one's ability to communicate effectively in writing by presenting a clear, concise, well-organized composition (National Conference of Bar Examiners, 2010). Because writing is fundamental to the practice of law, the inclusion of essay questions enhances the job relatedness of the exam and in all likelihood contributes to score validity. The essay format in this instance actually represents a work sample.

In contrast, other boards rely on the essay not to assess writing proficiency but to get at cognitive skills that are difficult to assess with the MCQ format (Day et al., 1990). The certification exam required of radiologist assistants represents an example of an essay exam intended to get at skills such as depth of knowledge, critical thinking, and reasoning (American Registry of Radiologic Technologists, 2009). There are two parts to the exam: a traditional MCQ followed by two in-depth cases. Each case opens with a scenario followed by a series of essay questions that require examinees to determine the types of diagnostic studies needed, explain the utility of those studies, discuss contraindications, interpret results, describe the radiologic procedures to be performed, and so on. Examinees are expected to present a coherent line of reasoning, but they are not penalized for spelling errors or poor writing style. Technical documentation indicates that essay scores based on three raters and two cases (each with four to six essay questions) produce dependability indices based on generalizability theory in the mid .70s. When essay scores are combined with multiple-choice scores, their composite yields an overall score reliability in the low to mid .90s (e.g., American Registry of Radiologic Technologists, 2009).

As previously noted, one limitation of the essay format is that writing skill is a source of construct-irrelevant variance for many professions. Although subjectivity in grading can also be a problem, most essay programs utilize quality-control procedures to ensure that raters remain reasonably consistent over time; in instances in which rater disagreements occur, some type of score resolution process is usually invoked (Johnson, Penny, & Gordon, 2009). Computers are now used to score essays for some large-scale testing programs, such as the Graduate Record Examination, and strong evidence exists to indicate that computers provide scores that are at least as reliable as, if not more reliable than, human raters (Bridgeman, Trapani, & Attali, 2012). To date, however, scoring essays for grammar

and mechanical features of writing has been a more successful endeavor than scoring for the accuracy of large idea units. Perhaps the primary limitation of essays is that they are time consuming for examinees, which severely limits the amount of content that can be assessed. This, in turn, threatens reliability and validity because an examinee's score may depend on the specific questions he or she happens to get on that form of the test.

Oral Examinations

Oral exams have long been regarded as a capstone to professional education in medicine, psychology, and other disciplines. An oral exam is required for psychology licensure in 28 of the 63 jurisdictions listed by the Association of State and Provincial Psychology Boards (2008) and is required for certification in many psychology specialties (Leigh et al., 2007). Similarly, 13 of the 24 member boards of the American Board of Medical Specialties require that candidates pass both a written and an oral examination to be certified (Raymond & Luciw-Dubas, 2010). Like the essay, the oral exam is viewed as a way to measure skills that elude the MCQ format—skills such as clinical reasoning, depth of knowledge, professionalism, and communication skills (Leigh et al., 2007). The case-based oral administered by the American Board of Emergency Medicine is an example of a rigorously developed program. Examinees are presented with case scenarios sampled from actual practice and are required to select and interpret diagnostic studies, prioritize and explain the rationale for various patient care activities, describe how procedures are performed, and monitor simulated patients over time. Seven patient encounters are presented over a 5-hour period; to more realistically portray the demands of an emergency department, some sessions require the simultaneous management of two or more patients. Each case is presented by a pair of examiners, both of whom provide performance ratings, with studies indicating high levels of rater agreement (Bianchi, Gallagher, Korte, & Ham, 2003).

Recent years have witnessed a concentrated effort by many medical boards to enhance their oral exams by standardizing cases and scoring protocols (Tekian & Yudkowsky, 2009), by using psychometric models to statistically control the effects of examiner leniency and case difficulty (McManus, Thompson, & Mollon; 2006; Raymond & Viswesvaran, 1993), and by more carefully sampling examiners and cases (Wass, Wakeford, Neighbour, & Van der Vleuten, 2003). This last point is worthy of elaboration. It was once thought that the best way to improve reliability of orals was to increase the number of examiners (i.e., raters). Studies based in generalizability theory (Brennan, 2001; see also Volume 1, Chapter 3, this handbook), however, clearly demonstrate that the case or topic effect has greater impact on reliability than the rater effect (Baker et al., 1993; Swanson et al., 1995; Turnbull, Danoff, & Norman, 1996; see also Volume 1, Chapter 20, this handbook). This topic is taken up in more detail later in this chapter.

Despite recent advances, oral exams still have their limitations. Reviews of the literature indicate that reliability coefficients typically range from the .30s to the .70s, and that oral ratings correlate in the .20s and .30s with measures of training or work performance (Muzzin & Hart, 1985). Perhaps the most serious challenges stem from the presence of construct-irrelevant variance, with studies suggesting that scores on oral exams are influenced by the candidate's self-confidence, organizational skills, communication style, and appearance (Rowland-Morin, Burchard, Garb, & Coe, 1991; Tekian & Yudkowsky, 2009). Many boards report large score gains for examinees who initially fail and then later repeat oral exams (Raymond & Luciw-Dubas, 2010). Although some of the score gain can be attributed to regression toward the mean because of the unreliability of oral exam ratings, some of the gain can probably be explained as method effect. That is, for many unsuccessful examinees, the first attempt may serve as an opportunity to learn how to take the oral (Roberts, Sarangi, Southgate, Wakeford, & Wass, 2000). These limitations, in combination with their high cost, may explain why the oral exams in psychology have come under increased criticism in some states (Oral Psychologist Exam Judges Personality, 2004).

Work Samples and Practical Exams

Practical exams are used by state and national credentialing programs in such fields as cosmetology,

dental hygiene, and crane operation, to name a few. For example, the National Commission for the Certification of Crane Operators (NCCCO; 2011) certifies operators of different types of cranes, and all exams have a written and practical component. The work-sample test for mobile crane operators requires that examinees operate a crane at a test site resembling a construction zone and demonstrate skill in several predetermined tasks (e.g., follow hand signals; navigate a large weight through a narrow corridor with turns; raise, lower, and relocate a weight to specified positions). The NCCCO takes steps to minimize construct-irrelevant variance associated with a practice effect through the availability of online orientation materials and by providing examinees a practice session just before taking a live exam to operate the equipment used for testing (NCCCO, 2011). Examinees are evaluated by trained examiners in terms of precision and steadiness of movement, and must complete each task within a specified time limit. Performance is generally evaluated in terms of number of errors (e.g., striking a pylon, not responding to a hand command) or the commission of an "unsafe act," which would result in immediate cancellation and failure of the exam. We could not locate any reports documenting the reliability or validity of scores for this program.

One of the more complex and well-researched work-sample tests is the clinical skills examination administered by the National Board of Medical Examiners (NBME). The clinical skills exam, known as Step 2 CS, uses a standardized patient format to assess an examinee's ability to manage patients within the context of a medical clinic. A standardized patient is a person who has been trained to portray a patient with some medical condition. For the Step 2 CS exam, each examinee encounters 12 patients throughout the course of a testing session. The clinic room is equipped with an examination table and other standard equipment (e.g., gowns, blood pressure cuff). For each patient, the examinee has 15 min to complete a patient history and physical examination, and an additional 10 min to document findings in a patient note. Examinees receive scores on skills related to data gathering and documentation, communication and interpersonal skills, and spoken English proficiency. Step 2 CS is taken by nearly 35,000 examinees each year just before starting residency training. Investigations of the factor structure of Step 2 CS support the internal structure of CS scores for both U.S. medical school graduates and international medical graduates (De Champlain, Swygert, Swanson, & Boulet, 2006). Correlations between Step 2 CS and the multiple-choice components of the physician licensure exams are low to moderate, suggesting that it makes a unique contribution to the measurement of physician competence (Harik et al., 2006). Indices of dependability (phi coefficients) for observed ratings based on generalizability theory typically range from the .60s to .90s (Harik et al., 2009). As described later in this chapter, research has shown that score reliability is substantially improved through the use of statistical models to adjust for rater-topic effects. Like oral and essay exams, topic specificity limits the reliability that can be achieved by practical exams and simulated work samples (the latter is described in Volume 1, Chapter 29, this handbook).

Computer-Based Simulations

Computer-based simulations are now used in a wide range of professions, including law, ophthalmic assisting, and dental hygiene. One specific example is the certification examination developed by the American Institute of Certified Public Accountants. Situation-based accounting simulations are developed involving audit settings, financial accounting and reporting settings, and tax and regulation settings, requiring knowledge and skills in research, analysis, and communication. Examinees may use accounting spreadsheets, complete accounting forms, or even author correspondence. Another example is the computer-based simulation given by NBME as part of the U.S. Medical Licensing examination. A simulated case opens with a patient who arrives at a health care facility and requires medical attention. The examinee is required to order and interpret diagnostic studies, initiate consults with other medical staff, and recommend an intervention or course of treatment. Each situation unfolds in real time during which examinees use the keyboard and mouse to manage the case. For example, an examinee can order and obtain results for a CT scan

almost immediately; however, in simulated time 2 hours may have lapsed, during which the patient's health status could have improved or deteriorated. Each action is recorded in a transaction list, which is then compared with a scoring key.

The Architectural Registration Examination (ARE) is another notable program that makes extensive use of computer-based simulations. The ARE consists of MCQs, short-answer questions, and an elaborate computer-administered simulation (National Conference of Architectural Registration Boards [NCARB], 2009). The simulation consists of 11 graphic vignettes addressing architectural tasks, such as site grading and design, stair design, roof plans, structural layout, and mechanical plan. Each vignette presents an architectural problem that requires the examinee to produce a solution in the form of a drawing complete with construction specifications. The instructions for each vignette include necessary requirements (e.g., indicate the elevation of all landings; the maximum slope of a ramp shall be 1:12). The computer software includes drawing and design software and provides access to all necessary references, such as building codes. Given the complexity of the testing software, NCARB has developed extensive practice materials that applicants are encouraged to review before taking the examination.

The ARE utilizes computer algorithms to score each exam. To produce scores, individual elements in the examinee drawings and specifications are compared with elaborate scoring trees generated by subject matter experts (Bejar, 1995). For example, a bathroom design would be evaluated in terms of the presence or absence of certain fixtures (vanity, bathtub), the placement and orientation of fixtures relative to others, appropriate lighting, and other design features judged by experts to be relevant. Each element in the design is graded as acceptable, indeterminate, or unacceptable. Early research on the ARE automated scoring indicated that computer-generated scores agreed with subject matter expert scores almost as well as the two subject matter experts agreed with each other. Cohen's kappa was .77 for two experts, and .70 and .75 between computer-generated scores and each expert (Bejar, 1991). Subsequent research has identified sources of disagreement between human and automated scores (Williamson, Bejar, & Hone, 1999). For example, human graders can make allowances for their perception that a competent examinee may have misinterpreted a design requirement or can make allowances based on perceived examinee intent (Williamson et al., 1999), whereas automated scoring lacks the insight to make such dispensations.

The automated scoring models used for different credentialing exams vary considerably in the way that the scoring keys are developed and implemented. One of the biggest challenges of automated scoring is the development of a clear understanding about the manner in which examinee task, the capture of scorable response information, and scoring rubrics interact to provide valid and reliable measurement information. Ideally, test designers should be cognizant of that interaction and systematically evaluate and modify their task designs, rather than relying on post hoc statistical analysis to discover and then logically label sources of variance or covariance as "valid" measurement information (Luecht & Clauser, 2002). Clauser, Kane, and Swanson (2002) have provided a taxonomy of automated scoring systems and have suggested strategies for assessing the reliability and validity of scores produced by such systems.

Summary of Assessment Methods

A credentialing program's choice of assessment format will be influenced by such factors as the purpose of the credential and the inferences to be made from test scores, the desired level of score reliability, administration costs, scoring logistics, and the perceived fidelity of an examination (i.e., face validity). For credentialing examinations that focus exclusively on assessment of the cognitive domain, any of the selected response formats (e.g., MCQs) are particularly useful for assessing knowledge of a broad range of topics in a limited amount of time. If there is a commitment to measuring higher level cognitive skills, then the examination might include a constructed-response format such as short answer, essay, or oral examination (Leigh et al., 2007). If the purpose of the credential implies assessment of the psychomotor or affective domains, then some type of performance test may be needed. If so, it probably should not be the sole method of assessment, but rather it should be used in conjunction with a

selected response format, assuming that the two formats measure similar constructs. To stand alone and be used for pass–fail decisions, performance tests need to be long—say, 4 to 6 hours of assessment time (Swanson et al., 1995). If they are not, the content domain will be underrepresented, and an examinee's score may be unduly influenced by the particular questions or tasks he or she happened to get. This source of measurement error has been repeatedly documented with various assessment formats and is referred to as content specificity, case specificity, or task specificity (Baker et al., 1993; Elstein, Shulman, & Sprafka, 1978). There is no doubt a trade-off: Enhancements to validity through increased task fidelity come at the expense of reliability and breadth of content. Other limitations of performance testing stem from the amount of time and number of raters required for performance testing. The pursuit of scoring efficiency has led to the successful implementation of large-scale practical exams and computer-based simulations with automated scoring. The use of computers to administer and especially score complex simulations is promising. Just as the popularity of MCQs is largely a consequence of their scoring efficiency, automated scoring may encourage further use of a wide range of complex item formats. Automated scoring also has the potential to compromise score validity, however, if hastily introduced and implemented (Clauser et al., 2002, p. 430).

PSYCHOMETRIC CHALLENGES IN CREDENTIALING

In many respects, credentialing exams are subject to the same practical constraints as educational and personnel tests, as are the psychometric tools for handling those constraints. Some issues, however, are especially challenging for credentialing exams. This final section mentions a few of those challenges, including reliability estimation, validation strategies, standard setting, retest effects, test security, and rater and task effects in performance assessment.

Reliability of Pass–Fail Decisions

No test is perfectly accurate, and any test score contains some degree of measurement error. Indeed, all test theories start with the conceptualization of an observed test score as a random draw from a distribution of possible scores had we repeatedly assessed an examinee using parallel forms of the test (Haertel, 2006). Classical test theory provides the basis for such indices as coefficient alpha, KR-20, and the conventional standard error of measurement (SEM). Indices based on classical test theory, however, are not appropriate for most credentialing examinations. First, traditional reliability indices are intended for norm-referenced score interpretations for which the primary interest lies in comparing the relative ranking of examinees (e.g., college admissions, personnel selection). In contrast, scores on credentialing examinations can be viewed as an estimate of the examinee's true score in the content domain from which the test items or tasks were sampled, and the reliability index should reflect this source of sampling error. Second, traditional measures of reliability based on classical test theory summarize the average measurement error across the score scale even though the magnitude of error is known to vary across the scale such that scores at the cut score can be more or less precise than implied by the overall SEM. Given that scores on credentialing examinations are interpreted with respect to a predetermined cut score, it is important to know the magnitude of measurement error near the cut score. Many would take this line of reasoning a step further and argue that reliability should be conceptualized in terms of decision consistency or the proportion of examinees accurately classified as passers and failers (AERA et al., 1999).

Pass–fail decisions can be characterized by two types of decision errors: false positives and false negatives. False-positive errors occur when examinees with a true score below the cut score are classified as passers because their observed test score meets or exceeds the cut score, whereas false-negative errors arise when examinees who are true passers are classified as nonpassers. Indices of classification consistency quantify the proportion of examinees who would be classified the same on two administrations of the same or parallel test. Cohen's kappa is one common index of classification consistency suitable for credentialing examinations. Brennan and Kane (1977) proposed an index of classification accuracy

based on generalizability theory that is computationally straightforward. Several other indices also exist, with many of them requiring the use of a computer to simulate large numbers of exam administrations. For details regarding the appropriateness of these various indices and their computation, see Clauser et al. (2006) and Haertel (2006).

Numerous approaches to computing measurement error at the cut score—and at any other point across the score distribution—have also been proposed. Generalizability theory provides one framework for computing SEMs conditioned on total score that is particularly useful with performance ratings. The absolute conditional SEM, $\sigma(\Delta_p)$, is given by

$$\sigma\left(\Delta_p\right)=\sqrt{\frac{\sum_r\left(X_{pr}-X_{p\cdot}\right)^2}{n_r\left(n_r-1\right)}}, \qquad (19.1)$$

where X_{pr} is the rating assigned to examinee p by rater r, and $X_{p\cdot}$ is the mean of all ratings for an examinee. This approach to computing conditional SEMs is also suitable for multiple-choice tests. If performance ratings are replaced by dichotomous item scores, then Equation 19.1 reduces to the conditional SEM based on the binomial error model and introduced by Lord more than 50 years ago (Brennan, 2001).

Item response theory (IRT) provides another useful conceptualization of measurement precision. Using the concept of measurement information conveyed in test scores, IRT provides easily computed conditional standard errors of the score estimates (*SEE*) from the test information function (*TIF*):

$$SEE\left(\hat{\theta}|\theta\right)=\left[TIF\left(\theta\right)\right]^{-1/2}=\left[\sum_{i=1}^{n}I_i\left(\theta;\xi_i\right)\right]^{-1/2}, \qquad (19.2)$$

where θ is the IRT proficiency estimate (i.e., score) of interest and ξ is the vector of item parameter estimate for a particular item (Lord, 1980). The *SEE* at the cut score for well-targeted tests typically will be smaller than the SEM based on classical theory because the standard error of estimate is properly computed in the region of the cut score rather than at the mean. For example, Test A may have a higher coefficient alpha than Test B but may actually be less precise than Test B at the cut score. IRT provides a way to determine the precision of scores at different regions of the score scale, which allows one to assemble tests that have maximal precision at the cut score. The use of TIFs to assemble test forms has become a standard practice for many credentialing programs (Luecht, 2006a). A final advantage of having conditional SEEs is that statistically defensible adjustments can be made by a policy board to effectively raise or lower a cut score based on the IRT information-based impact and the importance of the two types of classification errors. Because classification errors are a function of the precision of scores estimated from a particular test form in the region of the cut, it is possible to work out TIF-based adjustments that would minimize one type of decision error or the other.

Validation Strategies

Until the 1980s it was common to interpret scores on credentialing examinations as predictors of performance in practice, which implies the use of a criterion-related strategy for validating the interpretation of scores (Hecht, 1979). Although criterion-related validation is the sine qua non for personnel selection and admissions testing, for various reasons it has limited applicability to credentialing (Kane, 1982, 2006; Shimberg, 1981). Credentialing tests resemble achievement or job-knowledge tests and focus on job-related knowledge. This implies a content-orientation strategy, which places heavy emphasis on the role of job analysis. Furthermore, credentialing exams limit the extent to which general cognitive abilities and personality traits influence test scores (AERA et al., 1999). For example, the credentialing test taken by architects does not explicitly assess verbal ability, extroversion, or business acumen even though such qualities probably contribute to job success. The fundamental difference is that scores on credentialing tests are not interpreted in the same way as scores on selection tests (AERA et al., 1999). Successful performance on a credentialing test does not necessarily indicate that an examinee is likely to perform well at a specific job but only that he or she has acquired the knowledge and skills required for safe and effective practice. Interpretations concerning job success or

higher levels of performance generally are not warranted (Kane, 1982, p. 914).

The use of criterion-related validity studies in credentialing is also hindered by a variety of practical constraints. Problems associated with defining and obtaining reliable measures of criterion performance are well documented (see Volume 1, Chapter 3, this handbook). Job performance in many professions is remarkably multidimensional and context dependent (Richards, Taylor, Price, & Jacobsen, 1965), which compounds challenges when developing measures for use in criterion-related studies (Kane, 1982). Other challenges to validation stem from the interpretation of test scores on credentialing exams as indicators that one either does or does not possess the sufficient knowledge and skill required for effective practice. That is, test scores are used to make pass–fail decisions, and the use of dichotomous outcomes on the predictors or criteria (pass–fail on test, competent–not competent) in practice renders the traditional "validity coefficient" as inappropriate. Another challenge to collecting criterion-related evidence is that such data often are not available for examinees who fail the exam. In licensure, those who fail are not permitted to work, whereas for certification, those who fail are less likely to work in their preferred setting resulting in selection bias. Nonetheless, a few studies have been completed that document positive relationships between scores on credentialing examinations and job performance and patient outcomes in health care settings (Lunz, Castleberry, James, & Stahl, 1987; Norcini, Swanson, Grosso, Shea, & Webster, 1985; Ramsey et al., 1989; Tamblyn et al., 2002).

Validation and Standard Setting

A credentialing examination could have a strong relationship with professional practice and still be interpreted incorrectly if the cut score is too high or too low. Considerable research has evaluated alternative methods to establish passing scores on credentialing examinations. Although norm-referenced approaches to setting performance standards continue to be suitable for college admissions and personnel selection, nearly all credentialing programs now use the same types of domain-referenced standard-setting methods that are common in educational testing. The Angoff method and its modifications, along with various forms of the bookmark method, are commonly used by credentialing programs (Cizek, 2001; see also Chapter 22, this volume). Given the impact of the passing standard on the interpretation and use of credentialing tests, credentialing agencies need to provide evidence that passing scores are appropriate to the stated purpose of the test. Kane (1992, 2006) advocated an argument-based approach to validating credentialing test scores and interpretations, and his approach encompassed standard setting. An interpretative argument represents a chain of inferences starting with observed test scores and ending with the decisions to be made on the basis of those scores. The exact structure of such arguments is flexible and can be expected to vary depending on the assessment context. The types of inferences that might occur as part of the interpretive argument for scores on a credentialing examination would likely include the following: (a) observation and scoring, (b) generalization, (c) extrapolation, and (d) interpretation and decision making. The first stage of inference is the most basic. Inferences at this stage assume that the methods used to assign scores were consistent with appropriately designed measurement procedures (e.g., sound test items, accurate scoring key). The second stage requires assumptions regarding the extent to which scores on a test generalize to the domain of interest, and the third stage involves the extent to which performance on a test can be extrapolated to some context external to the test—that is, performance in some practical setting. A job analysis and rigorous test development procedures typically produce evidence to support extrapolation, as would correlations with important practice-related criteria. The final link of the interpretive argument concerns the decisions to be made on the basis of the test scores. The legitimacy of test use rests on assumptions and evidence regarding the processes used to establish the cut score. It is apparent that as one proceeds through this chain, the supporting evidence becomes more difficult to amass and the assumptions grow stronger. Both standard setting and argument-based validation are described elsewhere (see Volume 1, Chapter 4, this handbook, and Chapter 22, this volume).

Retest Effects

Given the dramatic influence of testing on an individual's life, it is not surprising that most credentialing agencies provide failing examinees with an opportunity to retest. Such policies are consistent with professional standards and government guidelines (AERA et al., 1999; Uniform Guidelines on Employee Selection Procedures, 1978). Retest policies imply that a single assessment may not provide an accurate measurement of an examinee's proficiency on the construct of interest (Lievens, Buyse, & Sackett, 2005) or that examinees otherwise deserve multiple attempts at demonstrating proficiency. Meta-analyses of cognitive ability tests used for personnel selection and college admissions report score gains of one fourth of a standard deviation (*SD*) for examinees who take a different (parallel) test form on their second attempt; gains for examinees who receive the same test on their second attempt are nearly twice the magnitude, approaching nearly one half *SD* (Hausknecht, Halpert, Di Paolo, Moriarty, & Gerrard, 2007; Kulik, Kulik, & Bangert-Drowns, 1984). The few studies of retest effects on credentialing exams report score gains ranging from one half to three fourths of an *SD* whether examinees see the same form or a different form on their second occasion (Geving, Webb, & Davis, 2005; Raymond, Neustel, & Anderson, 2007, 2009). Although the retest effect is large for credentialing exams, studies indicate that there is little additional advantage to seeing the same form twice. Other factors that moderate score gains include examinee ability, with low-ability examinees showing smaller gains, and type of test, with gains on achievement or knowledge tests demonstrating smaller gains than those on test of cognitive ability (Hausknecht et al., 2007; Kulik et al., 1984). It has been suggested that exam length, item delivery sequence (random vs. fixed), and testing time influence score gains, with smaller gains on long, speeded tests for which items are presented in random order (Raymond et al., 2007).

Scores on credentialing exams increase for a variety of reasons. First, examinees may exhibit a legitimate improvement on the construct being measured. Testing itself promotes learning, particularly when corrective feedback is provided (Butler, Karpicke, & Roediger, 2007; Chan, McDermott, & Roediger, 2006) and when the motivation to pass is high. Score increases that can be explained by an increase in proficiency on the construct of interest are desirable and even inspiring.

A second source of score change can be attributed to construct-irrelevant or method variance (Messick, 1989). An examinee might score higher on a second attempt due to a reduction in test anxiety, an increase in general test-taking skill, or an improvement in some other skill secondary to the test (e.g., reading comprehension, self-confidence). The use of novel and complex item formats has been associated with large score gains on cognitive ability tests, and retest effects can be minimized by using item types that are less susceptible to coaching and practice effects (Powers, 1986). Retest effects on achievement and credentialing tests will be smaller to the extent that they consist of standard MCQs rather than the clever item formats often used on cognitive ability tests. Interestingly, a portion of the large score increases observed on oral exams and standardized patients has been attributed to the novel format of these assessments for some examinees (Raymond & Luciw-Dubas, 2010; Roberts et al., 2000; Swygert, Balog, & Jobe, 2010).

A third type of score gain occurs for examinees who memorize items or topics that comprise a specific test form. Because increases of this nature will be restricted to the content of a particular form, test performance is unlikely to generalize to the domain of interest, thereby compromising validity. This type of increase is suggested by the same-form retest effects for cognitive ability tests reported in the meta-analyses (Hausknecht et al., 2007; Kulik et al., 1984). Large-scale credentialing programs normally control item exposure by developing multiple forms of an examination. Similarly, CAT programs employ item-selection algorithms that prevent examinees from being presented items they were given on previous administrations (Wainer, 1990). Although previously cited evidence suggests that it may not be necessary to assign repeat examinees to alternate forms, it is still good practice assuming that resources are available to assemble and properly equate parallel forms.

A fourth source of score increase can be attributed to regression toward the mean. The magnitude of

the regression effect is positively related to the distance of the initial score from the overall mean and inversely related to reliability. Because only the lowest scoring examinees repeat credentialing exams, scores for these examinees, on average, are all but guaranteed to increase because of regression. The regression effect is of particular concern for performance-based assessments for which reliabilities are typically lower. Simulation studies indicate that scores for repeat examinees on such tests can be expected to regress upward by one fourth to one half of an *SD* (Raymond & Luciw-Dubas, 2010). Regression effects, in combination with a practice effect, can be quite large for performance tests, producing very high pass rates for repeat examinees.

Score gains associated with measurement error can be a problem even for very reliable test scores. Retesting policies help ensure that false-negative errors are short-lived by allowing examinees who were unfavorably affected by measurement error on the first attempt to earn scores consistent with their proficiency on subsequent attempts. Credentialing agencies, however, typically do not have mechanisms in place to ensure that false-positive errors be held to a minimum. This is of particular concern for credentialing exams intended to protect the public because the presence of false-positive errors means that unqualified examinees are granted a credential to practice (Millman, 1989). Examples provided by Clauser and Nungester (2001) illustrated that for test scores with a reliability of .90, the magnitude of false-positive error rates approach 45% after three attempts. Researchers have proposed ways to manage false-positive error rates (Clauser & Nungester, 2001; Millman, 1989). The most obvious strategy is to limit the number of opportunities to retest. This will have a minimal impact because most of the increase in classification errors occurs on the second and third attempts, and it is hard to imagine not allowing at least two additional attempts. Another approach is to use the average of scores rather than the highest score. This is appealing because it is consistent with the principle of reducing uncertainty by obtaining additional observations. Two additional approaches require manipulating the cut score—either by raising it for all examinees to account for measurement error or by raising it only for repeat examinees. All of these policies have been shown to reduce false-positives while having negligible impact on false-negative error rates (Clauser & Nungester, 2001). Current retest policies clearly give examinees the benefit of the doubt. As noted by Millman (1989), such policies should be reevaluated in light of the social consequences of granting a credential to unqualified individuals.

Test Security

Credentialing examinations typically have very high stakes, particularly those taken at the conclusion of a lengthy or expensive education program. Some examinees and test preparation firms engage in a variety of questionable activities to increase the odds of passing an exam. Credentialing agencies must maintain a constant vigilance for attempts by others to subvert the integrity of a testing program through such acts as memorizing test materials, recording and distributing test materials, stealing test booklets, sneaking information (items or answers) into test centers, and hiring imposters to take an examination, to name a few. Although computer-based testing has alleviated some types of security problems, it has created other challenges. For example, the administration of the same test forms continuously for several weeks or months results in the repeated exposure of large numbers of test items, which increases their likelihood of being reproduced and shared. Meanwhile, other types of technology such as miniature recording devices, cell phones, and the Internet provide the capability to record and quickly distribute test items to large numbers of people. Credentialing agencies devote significant resources to preventing and detecting such test fraud. Prevention measures involve activities like developing test materials to minimize their memorability or reproducibility (Impara & Foster, 2006), educating prospective examinees about legal and ethical considerations, fingerprinting and videotaping examinees at test centers, and limiting the number of retakes allowed. Detection methods include such procedures as statistical analyses of correct and incorrect answers (e.g., to flag copying behavior), comparisons of an examinee's scores on different subsets of items to identify access to one subset or another, analyses of examinee response times to test

items, and a review of records for repeat examinees who have very large score gains. In addition, some credentialing agencies routinely monitor the Internet to determine whether examinees or others (e.g., educators, test coaching firms) are actively soliciting or posting copyrighted test materials.

Instances of stealing test books or otherwise distributing test materials have been reported in numerous occupations and professions (Sahagan, 2007; Smydo, 2003). Although copyright infringement is difficult to prove, credentialing agencies have been successful in seizing the assets of test preparation firms and, in some instances, such firms have been ordered to pay considerable damages. Most notably, in 2006, a federal court ordered a test preparation firm to pay more than $11.9 million to the NCBE (*National Conference of Bar Examiners v. Multistate Legal Studies, Inc.*). Among other things, court documents showed a striking similarity between more than 100 items owned by NCBE that were included in the course materials offered by the defendant; documents further noted that an owner of the test preparation firm had personally taken the actual bar exam more than 20 times. The magnitude of the award was apparently influenced by the testing firm's success: In 2004, it had grossed $16 million.

The negative consequences of efforts to subvert the integrity of a credentialing exam are not trivial. When cheating is successful, the validity of score interpretations is directly compromised in that individuals who are not qualified to practice receive a credential. The ultimate consequence is that public protection may be compromised. In addition, test fraud can be unfair to other examinees. Although credentialing exams, strictly speaking, are not norm referenced—passing depends on exceeding a passing score, not outscoring other examinees—a credentialing agency may feel pressured to raise the performance standard if it perceives that pass rates are creeping up for illegitimate reasons. Such actions could adversely affect borderline passers. Furthermore, the credibility of the profession suffers when the public reads headlines to the effect that doctors, lawyers, dentists, or members of some other profession have been caught cheating on their board certification exams (Doctor's Cheating, 2005; Smydo, 2003). If cheating becomes prevalent rather than an isolated incident, then a more serious outcome could be an overall dilution in the collective proficiency of a profession. Finally, credentialing agencies design their programs to deter cheating. Examinees are inconvenienced because some credentialing agencies administer their test infrequently to minimize the risk of item exposure. When a major breach does occur, the expenses associated with recovery can be considerable (e.g., replacement of test forms, legal fees). These costs are ultimately passed on to examinees or to the public. Although some examinees may consider "getting through their boards" as a rite of passage and view the sharing of test questions as a matter of gamesmanship, when taken too far, such activities are unethical and illegal and can undermine the efficacy of a credentialing program.

Controlling Rater and Task Effects

An examinee's score on a performance assessment will be influenced by the particular sample of tasks he or she responded to, the raters who rated the response, and other factors (e.g., the testing occasion commonly reported as test–retest reliability; methods for quantifying these sources of variability are described in Volume 1, Chapters 3 and 20, this handbook). One obvious approach to managing variation due to task difficulty and rater leniency is to administer the exam using a completely crossed rating design such that all examinees respond to the same tasks and are evaluated by the same raters. This typically is not possible with large-scale credentialing programs. Instead, nested rating designs are used, which means that an examinee's score and pass–fail outcome will be influenced to some extent by the sample of tasks or raters he or she happened to receive. Rater training, systematic test development, and other quality control procedures can and should be used to minimize these sources of variability (Johnson et al., 2009). However, these procedures are only partially effective in controlling measurement error (Braun, 1988; Harik et al., 2009).

The use of statistical models to detect and correct for rater and task effects has become increasingly common over the past 20 years. These methods are capable of managing nested and other incomplete designs as long as there is sufficient overlap among

raters and tasks across cohorts. Methods based on IRT—in particular the Rasch model—enjoy widespread use in educational testing and credentialing (McManus, Thompson, & Mollon, 2006; Myford & Wolfe, 2009). Structural equation models have also been employed to calibrate raters and tasks under severely incomplete designs (Kahraman, De Champlain, & Raymond, 2012). A flexible and simple approach is to formulate a linear model that uses ordinary least squares (OLS) regression to estimate examinee, task, and rater parameters (Braun, 1988). This is not multiple regression in the typical sense but rather is more like the use of regression for analysis of variance for which a vector of observed ratings is regressed onto a design matrix consisting of dummy or effect coding to indicate the rater–task–examinee interactions that correspond to each observed score. Using OLS to adjust essay ratings has been shown to substantially increase reliability of essay ratings (Braun, 1988), clinical skills exam for physician licensure (Harik et al., 2009), and oral exams (Raymond & Viswesvaran, 1993). In the Harik et al. (2009) study for example, the improvement in reliability was comparable to what would be achieved by about a 45% increase in testing time.

Although statistical models can correct for systematic variation associated with task difficulty and rater leniency (i.e., the main effects), it typically is not possible to adjust for the error associated with task specificity and rater inconsistency (i.e., interaction effects). About the only way to reduce error associated with such interactions is to administer a larger sample of tasks or by increasing the number of raters. Generalizability theory can help optimize such decisions by informing the manner in which tasks and raters are assigned to examinees to maximize reliability. Consider an oral exam consisting of a series of 30-minute sessions, with each session focusing on a particular class of medical cases (e.g., cardiovascular conditions). Given that a major resource constraint for oral exams is the availability of examiners, the certification board wants to make effective use of an examiner's time. The board considers three strategies, each of which requires the same number of examiners and total amount of examiner time. The three options are as follows:

- Option A: an examinee sees an examiner who administers all six cases over 3 hours
- Option B: an examinee sees six different examiners (one for each case) over 3 hours
- Option C: an examinee sees two examiners for each of three cases over 1.5 hours

Note that doubling up on examiners requires a reduction in the number of cases seen by each examinee because of constraints on examiner availability. Figure 19.2 illustrates the reliability (dependability) associated with these three strategies. Generalizability theory was used to produce Figure 19.2; the necessary data were taken from Wass et al. (2003, Table 3). The lowest of the three lines (Option A) shows the minimal gains in dependability that occur by varying the number of cases while retaining the same examiner for each case. The top line (Option C) shows the improvement in dependability by having two raters for each case, and the middle line (Option B) illustrates the advantage realized by having a single, but different, examiner for each case. It is evident that having one rater per case over multiple cases is the most efficient use of examiner resources. Administering three cases with three pairs of examiners requires a total of six raters and produces a reliability of .73, while assigning a single but different examiner across six cases results in a dependability index of .80. Reliability could be increased even further through the use of one of the statistical models described earlier. These results highlight the point that although rater error is not ignorable, it is the sampling of tasks that typically limits the reliability of performance assessments, and the most efficient way to improve reliability is to add more tasks, not more raters. Although these data are from a single study, similar findings are reported elsewhere (Baker et al., 1993; Turnbull et al., 1996; Swanson et al., 1995; see also Volume 1, Chapter 20, this handbook).

CONCLUSION

Although we have no empirical evidence to support this assertion, it certainly appears as if the number of credentialed occupations and professions has increased dramatically in the 30 years since

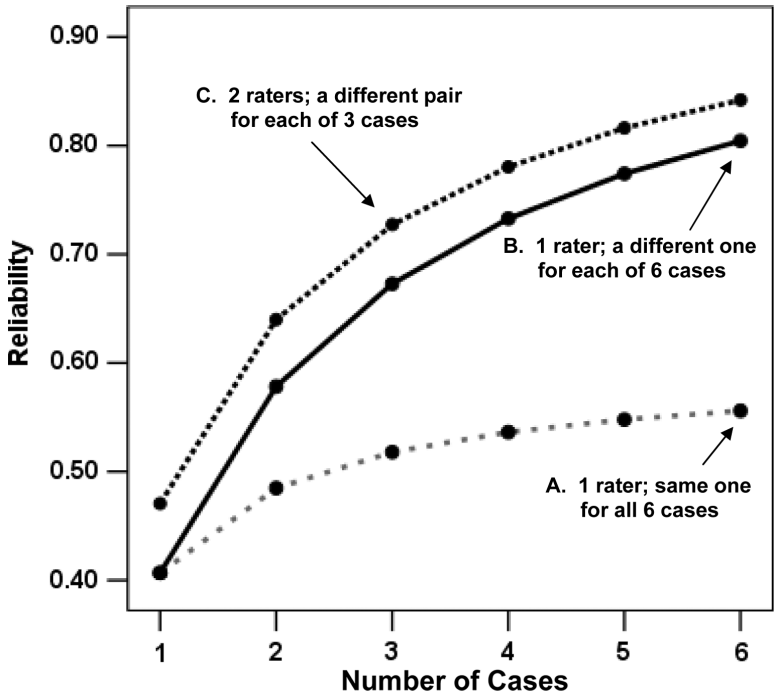


FIGURE 19.2. Score dependability (phi) for an oral examination as a function of number of cases and number of raters. Data adapted from Table 3 of Wass et al. (2003).

Shimberg (1981, 1982) and Kane (1982) first identified credentialing as a unique and important area of testing. In the 21st century, several national and international organizations exist to influence the manner in which credentialing exams are developed, administered, and scored (e.g., American Board of Medical Specialties, American Board of Nursing Specialties, and the National Commission for Certifying Agencies as well as APA, AERA, and NCME). Although there is still considerable variability in the quality of credentialing exams, it is safe to say that the psychometric properties of such exams have improved over the past few decades with advances in standard setting, performance assessment, test administration, and equating. The purpose of this chapter has been to summarize the manner in which credentialing programs presently determine what to test and how to test it. In addition, the chapter touched on some of the practical and psychometric challenges that are particularly germane to the field of credentialing. The number and types of credentialing exams will most certainly grow as methods of performance assessment continue to evolve and as new types of credentialing programs emerge (e.g., recertification) to ensure that professionals remain competent beyond their initial licensure or certification.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Registry of Radiologic Technologists. (2009). *Registered radiologist assistant (RRA) examination statistics.* St. Paul, MN: Author.

Association of State and Provincial Psychology Boards. (2008). *Oral examination requirements by jurisdiction and license type.* Retrieved from http://www.asppb.org/HandbookPublic/HandbookReview.aspx

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessments. *American Psychologist*, *48*, 1210–1218. doi:10.1037/0003-066X.48.12.1210

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76*, 522–532. doi:10.1037/0021-9010.76.4.522

Bejar, I. I. (1995). From adaptive testing to automated scoring of architectural simulations. In E. L. Mancall & P. G. Bashook (Eds.), *Assessing clinical*

reasoning: The oral examination and alternative methods (pp. 115–130). Evanston, IL: American Board of Medical Specialties.

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, predictive, and progressive approach to standard setting. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 31–63). Maple Grove, MN: JAM Press.

Bianchi, L., Gallagher, E. J., Korte, R., & Ham, H. P. (2003). Interexaminer agreement on the American Board of Emergency Medicine oral certification examination. *Annals of Emergency Medicine, 41*, 859–864. doi:10.1067/mem.2003.214

Braun, H. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*, 1–18. doi:10.2307/1164948

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: Praeger.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277–289. doi:10.1111/j.1745-3984.1977.tb00045.x

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*, 27–40. doi:10.1080/08957347.2012.635502

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*, 273–281. doi:10.1037/1076-898X.13.4.273

Chan, J. C., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*, 553–571. doi:10.1037/0096-3445.135.4.553

Cizek, G. (2001). *Setting performance standards*. Mahwah, NJ: Erlbaum.

Cizek, G. J., Robinson, K. L., & O'Day, D. M. (1998). Nonfunctioning options: A closer look. *Educational and Psychological Measurement, 58*, 605–611. doi:10.1177/0013164498058004004

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education, 15*, 413–432. doi:10.1207/S15324818AME1504_05

Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 701–731). Westport, CT: American Council on Education/Praeger.

Clauser, B. E., & Nungester, R. (2001). Classification accuracy for test that allow retakes. *Academic Medicine*, *10*(Suppl.), S108–S110. doi:10.1097/00001888-200110001-00036

Cooke, N. J. (1999). Knowledge elicitation. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. F. Dumais, & M. T. H. Chi (Eds.), *Handbook of applied cognition* (pp. 479–509). New York, NY: Wiley.

Day, S. C., Norcini, J. J., Diserens, D., Cebul, R. D., Schwartz, J. S., & Beck, L. H. (1990). The validity of an essay test of clinical judgment. *Academic Medicine*, *65*(Suppl.), S39–S40. doi:10.1097/00001888-199009000-00034

De Champlain, A. F., Swygert, K. A., Swanson, D. B., & Boulet, J. R. (2006). Assessing the underlying structure of the USMLE Step 2 test of clinical skills using confirmatory factor analysis. *Academic Medicine*, *81*(10, Suppl.), S17–S20. doi:10.1097/00001888-200610001-00006

Doctor's cheating forces test changes. (2005, May 29). *Philadelphia Enquirer,* p. B4.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Westport, CT: American Council on Education/Praeger.

DuBois, P. H. (1970). *A history of psychological testing.* Boston, MA: Allyn & Bacon.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance: The description of human tasks*. New York, NY: Academic Press.

Geving, A. M., Webb, S., & Davis, B. (2005). Opportunities for repeat testing: Practice doesn't always make perfect. *Applied HRM Research, 10*, 47–56.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12*, 109–113. doi:10.1111/j.1745-3984.1975.tb01013.x

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education/Praeger.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Erlbaum.

Harik, P., Clauser, B. E., Grabovsky, I., Margolis, M. J., Dillon, G. F., & Boulet, J. R. (2006). Relationships among subcomponents of the USMLE Step 2

clinical skills examination, the Step 1, and the Step 2 clinical knowledge examinations. *Academic Medicine*, *81*(10, Suppl.), S21–S24. doi:10.1097/01.ACM.0000236513.54577.b5

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability framework. *Journal of Educational Measurement, 46*, 43–58. doi:10.1111/j.1745-3984.2009.01068.x

Harvey, R. J. (1991). Job analysis. In M. Dunnette & L. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 71–163). Palo Alto, CA: Consulting Psychologists Press.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., Moriarty, N. T., & Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373–385. doi:10.1037/0021-9010.92.2.373

Hecht, K. A. (1979). Current status and methodological problems of validating professional licensing and certification examinations. In M. A. Budra & J. R. Sanders (Eds.), *Practices and problems in competency-based education*. Washington, DC: National Council on Measurement in Education.

Impara, J. C., & Foster, D. (2006). Item and test development strategies to minimize test fraud. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 91–114). Mahwah, NJ: Erlbaum.

Johnson, R. R., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: Guilford Press.

Kahraman, N., De Champlain, A. F., & Raymond, M. R. (2012). Modeling the psychometric properties of complex performance assessments using confirmatory factor analysis: A multi-stage model for calibrating tasks. *Applied Measurement in Education, 25*, 79–95. doi:10.1080/08957347.2012.635510

Kane, M. T. (1982). The validity of licensure examinations. *American Psychologist, 37*, 911–918. doi:10.1037/0003-066X.37.8.911

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*, *112*, 527–535. doi:10.1037/0033-2909.112.3.527

Kane, M. T. (1997). Model-based practice analysis and test specifications. *Applied Measurement in Education, 10*, 5–18. doi:10.1207/s15324818ame1001_1

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education/Praeger.

Knapp, J., & Knapp, L. (2007). *Knapp certification industry scan*. Princeton, NJ: Knapp.

Kulik, J. A., Kulik, C. C., & Bangert-Drowns, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*, 435–447.

LaDuca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation and the Health Professions, 17*, 178–197. doi:10.1177/01632787 9401700204

Landy, F. J. (1988). Selection procedure development and usage. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government* (Vol. 1, pp. 271–287). New York, NY: Wiley.

Leigh, I. W., Smith, I. L., Bebeau, M. J., Lichtenberg, J. W., Nelso, P. D., Portnoy, S., . . . Kaslow, N. J. (2007). Competency assessment models. *Professional Psychology: Research and Practice*, *38*, 463–473. doi:10.1037/0735-7028.38.5.463

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retesting effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58*, 981–1007. doi:10.1111/j.1744-6570.2005.00713.x

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.

Luecht, R. M. (2006a). Designing tests for pass–fail decisions using item response theory. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 575–596). Mahwah, NJ: Erlbaum.

Luecht, R. M. (2006b, May). *Engineering the test: Principled item design to automated test assembly*. Invited special event presentation at the Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Luecht, R. M., & Clauser, B. E. (2002). Test models for complex CBT. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing* (pp. 67–88). Mahwah, NJ: Erlbaum.

Luecht, R. M., Dallas, A., & Steed, T. (2010, April). *Developing assessment engineering task models: A new way to develop test specifications*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 229–249. doi:10.1111/j.1745-3984.1998.tb00537.x

Lunz, M. E., Castleberry, B. M., James, K., & Stahl, J. (1987). The impact of the quality of laboratory staff on the accuracy of laboratory results. *JAMA, 258*, 361–363. doi:10.1001/jama.1987.03400030077036

McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency (hawk-dove effect) in the MRCP(UK) clinical examination (PACES) using multi-faceted Rasch modelling. *BMC Medical Education, 6*, 42–64. doi:10.1186/1472-6920-6-42

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *32*(2), 13–23.

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher*, *18*(6), 5–9.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439–483. doi:10.1007/BF02294388

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of bias in job analysis. *Journal of Applied Psychology, 82*, 627–655. doi:10.1037/0021-9010.82.5.627

Morgeson, F. P., Klinger, K. D., Mayfield, M. S., Ferrara, P., & Campion, M. A. (2004). Self-presentation processes in job analysis: A field experiment investigating inflation in abilities, tasks and competencies. *Journal of Applied Psychology, 89*, 674–686. doi:10.1037/0021-9010.89.4.674

Muzzin, L. J., & Hart, L. (1985). Oral examinations. In V. Neufeld & G. R. Norman (Eds.), *Assessing clinical competence* (pp. 71–93). New York, NY: Springer.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*, 371–389. doi:10.1111/j.1745-3984.2009.00088.x

National Commission for the Certification of Crane Operators. (2011). *Certification programs: Mobile crane operator, practical exam—exam outline*. Retrieved from http://www.nccco.org/certification/practicalexam.html

National Conference of Bar Examiners. (2010). *The Multistate Essay Examination: 2010 information booklet*. Madison, WI: Author.

National Conference of Bar Examiners v. Multistate Legal Studies, Inc., d/b/a PMBR, 458 F. Supp. 2d 252, 2006 U.S. Dist. LEXIS 59477 (E. D. Pa. 2006)

National Conference of Examiners for Engineering and Surveying. (1999). *PE examination formats: Principles and practice of engineering examinations.* Clemson, SC: Author.

National Council of Architectural Registration Boards. (2009). *ARE guidelines*. Washington, DC: Author.

Norcini, J. J., Swanson, D., Grosso, L., Shea, J., & Webster, G. (1985). Reliability, validity, and efficiency of multiple choice questions and patient management problem item formats in the assessment of physician competence. *Medical Education, 19*, 238–247. doi:10.1111/j.1365-2923.1985.tb01314.x

Oral psychologist exam judges personality—Not competence—Review finds (2004, March/April). *Professional Licensing Report*, p. 12.

Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin*, *100*, 67–77. doi:10.1037/0033-2909.100.1.67

Ramsey, P. G., Carline, J. D., Inui, T. S., Larson, E. B., LoGerfo, J. P., & Wenrich, M. D. (1989). Predictive validity of certification by the American Board of Internal Medicine. *Annals of Internal Medicine, 110*, 719–726.

Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education, 14*, 369–415. doi:10.1207/S15324818AME1404_4

Raymond, M. R., & Luciw-Dubas, U. A. (2010). The second time around: Accounting for retest effects on oral examinations. *Evaluation and the Health Professions, 33*, 386–403.

Raymond, M. R., & Neustel, S. (2006). Determining the content of credentialing examinations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 181–223). Mahwah, NJ: Erlbaum.

Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology, 60*, 367–396. doi:10.1111/j.1744-6570.2007.00077.x

Raymond, M. R., Neustel, S., & Anderson, D. (2009). Same-form retest effects on credentialing examinations. *Educational Measurement: Issues and Practice*, *28*(2), 19–27. doi:10.1111/j.1745-3992.2009.00144.x

Raymond, M. R., & Viswesvaran, C. (1993). Least-squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement, 30*, 253–268. doi:10.1111/j.1745-3984.1993.tb00426.x

Richards, J. M., Jr., Taylor, C. W., Price, P. B., & Jacobsen, T. L. (1965). An investigation of the criterion problem for one group of medical specialists. *Journal of Applied Psychology, 49*, 79–90. doi:10.1037/h0021960

Roberts, C., Sarangi, S., Southgate, L., Wakeford, R., & Wass, V. (2000). Oral examinations—equal opportunities, ethnicity, and fairness in the MRCGP. *British Medical Journal, 320*, 370–375. doi:10.1136/bmj.320.7231.370

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. doi:10.1111/j.1745-3992.2005.00006.x

Rosenfeld, M., Shimberg, B., & Thornton, R. F. (1983). *Job analysis of licensed psychologists in the United States and Canada.* Princeton, NJ: Educational Testing Service.

Rowland-Morin, P. A., Burchard, K. W., Garb, J. L., & Coe, N. P. W. (1991). Influence of effective communication by surgery students on their oral examination scores. *Academic Medicine*, *66*, 169–171. doi:10.1097/00001888-199103000-00011

Sahagan, L. (2007, November 4). 2 Dentistry scandals rock UCLA. *Los Angeles Times.*

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist, 36*, 1138–1146. doi:10.1037/0003-066X.36.10.1138

Shimberg, B. (1982). *Occupational licensing: A public perspective.* Princeton, NJ: Educational Testing Service.

Smydo, J. (2003, August 3). Health fields fight cheating on tests. *Pittsburgh Post-Gazette*, pp. A1, A4.

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher*, *24*(5), 5–11, 35.

Swygert, K. A., Balog, K. P., & Jobe, A. (2010). The impact of repeat information on examinee performance for a large-scale standardized-patient examination. *Academic Medicine, 85*, 1506–1510. doi:10.1097/ACM.0b013e3181eadb25

Tamblyn, R., Abrahamowicz, M., Dauphinee, W. D., Hanley, J., Norcini, J. J., Girard, N., . . . Brailovsky, C. (2002). Association between licensure examination scores and practice in primary care. *JAMA*, *288*, 3019–3026. doi:10.1001/jama.288.23.3019

Tannenbaum, R. J., Robustelli, S. L., & Baron, P. A. (2008). Evidence-centered design: A lens through which the process of job analysis may be focused to guide the development of knowledge-based test content specifications. *CLEAR Exam Review*, *19*(2), 26–33.

Tannenbaum, R. J., & Wesley, S. (1993). Agreement between committee-based and field-based job analyses: A study in the context of licensure testing. *Journal of Applied Psychology, 78*, 975–980. doi:10.1037/0021-9010.78.6.975

Tekian, A., & Yudkowsky, R. (2009). Oral examinations. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 269–285). New York, NY: Routledge.

Turnbull, J., Danoff, D., & Norman, G. (1996). Content specificity and oral certification examinations. *Medical Education, 30*, 56–59. doi:10.1111/j.1365-2923.1996.tb00718.x

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology, 1*, 386–391. doi:10.1016/0022-2496(64)90010-0

Uniform Guidelines on Employee Selection Procedures. (1978). *Federal Register*, *43*(166), 38296–38309.

U.S. Department of Health, Education, and Welfare, Public Health Services (1977). *Credentialing health manpower* (DHEW Publication No. [05] 77–50057). Washington, DC: Author.

Wainer, H. (1990). *Computer adaptive testing: A primer.* Mahwah, NJ: Erlbaum.

Wang, N., Schnipke, D., & Witt, E. A. (2005). Use of knowledge, skill, and ability statements in developing licensure and certification examinations. *CLEAR Exam Review*, *19*(2), 26–33.

Wass, V., Wakeford, R., Neighbour, R., & Van der Vleuten, C. (2003). Achieving acceptable reliability in oral examinations: An analysis of the royal college of general practitioners membership examination's oral component. *Medical Education, 37*, 126–131. doi:10.1046/j.1365-2923.2003.01417.x

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). "Mental model" comparison of automated and human scoring. *Journal of Educational Measurement, 36*, 158–184. doi:10.1111/j.1745-3984.1999.tb00552.x

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum.

Zara, A. R. (1999). Using computer adaptive testing to evaluate nurse competence: Some history and a forward look. *Advances in Health Sciences Education, 4*, 39–48. doi:10.1023/A:1009866321381

CHAPTER 20

EVALUATING TEACHING AND TEACHERS

Drew H. Gitomer and Courtney A. Bell

Almost everything related to the assessment and evaluation of teaching in the United States is undergoing restructuring. Purposes and uses, data sources, analytic methods, assessment contexts, and policy are all being developed, refined, and reconsidered within a cauldron of research, development, and policy activity. For example, the District of Columbia made headlines when it announced the firing of 241 teachers based, in part, on poor performance results from their new evaluation system, IMPACT (Turque, 2010). The Bill and Melinda Gates Foundation has funded the Measures of Effective Teaching (MET) study, a $45 million study designed to test the ways in which a range of measures including scores on observation protocols, student engagement data, and value-added test scores might be combined into a single teaching evaluation metric (Bill and Melinda Gates Foundation, 2011a). The Foundation is also spending $290 million in four communities in intensive partnerships to reform how teachers are recruited, developed, rewarded, and retained (Bill and Melinda Gates Foundation, 2011b). In addition to pressure from districts and private funders, unions have also pressed for revised standards of teacher evaluation (e.g., American Federation of Teachers [AFT], 2010). Perhaps the most consequential contemporary effort is the federally funded Race to the Top Fund that encourages states to implement teacher evaluation systems based on multiple measures with a significant component based on students' academic growth to achieve funding (U.S. Department of Education, 2010). These and other recent research and policy developments are changing the way the assessment of teaching is understood. The goal of this chapter is to provide an overview and structure to facilitate readers' understanding of the emerging landscape and attendant assessment issues.

As well described in a number of recent reports, current evaluation processes suffer from a number of problems (Toch & Rothman, 2008; Weisberg, Sexton, Mulhern, & Keeling, 2009). For example, the New Teachers Project surveyed evaluation practices in several districts large and small and found that teachers were almost all rated highly. In systems that used binary ratings (i.e., satisfactory or unsatisfactory), almost 99% of teachers were rated satisfactory. To complicate matters, the same administrators who gave all teachers high marks also recognized that staff members varied greatly in performance and that some were actually poor teachers. In addition to an inability to sort teachers, current processes generally do not give teachers useful information to improve their practice, and policymakers do not believe the credibility of the evaluation process (Weisberg et al., 2009).

Measures of teaching should be seen from a validity perspective, and thus, it is critical to begin with the purpose and use of the assessment. As Messick (1989) argued, validity is not an inherent

The authors thank Andrew Croft, Laura Goe, Heather Hill, Daniel McCaffrey, and Joan Snowden for their careful review of an earlier version of this chapter. A special thank-you to Andrew Croft and Evelyn Fisch for their assistance in preparing the chapter. We also appreciate the review of Daniel Eignor, who passed away in 2012. Each of the authors contributed equally to the preparation of this chapter.

DOI: 10.1037/14049-020
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

property of an instrument, but rather it is an evaluation of the inferences and actions made in light of a set of intended purposes. Given the extraordinary and unprecedented focus on evaluating teacher quality, this chapter is focused on measures being used to make inferences about the quality of practicing teachers, and to a lesser degree, the inferences made about prospective teachers who are undergoing professional preparation. These measures are examined through the perspective of modern validity frameworks used to consider the quality of assessments more generally. Building on M. T. Kane's (2006) thinking, the strength of the validity evidence is considered while paying careful attention to the purposes of various teaching evaluation instruments.

In considering the validity of inferences made about teacher quality, the focus is on three issues that may be at the forefront of discussions about teacher evaluation for the foreseeable future. The first issue concerns the validity argument for particular instruments. Guided by M. T. Kane (2006), Messick (1989), and the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), the respective validity arguments for a range of measures being used to evaluate teachers are summarized briefly.

The second issue concerns an often underresearched aspect of any validity argument—causal attribution of scores to particular teachers. Observing teaching or assessing student learning provides a set of observables that produce a score or set of scores. Policymakers and many researchers are seeking ways to establish a causal relationship that attributes these scores to an individual teacher. Yet the nature of teaching and the context in which it occurs raise questions about the extent to which causal attribution can be made. To date, issues of causal attribution have not yet been adequately dealt with across instruments and processes used to measure teaching.

The final issue concerns the consideration of multiple measures in an evaluation. Although the most recent *Standards for Educational and Psychological Testing* (AERA et al., 1999) discussed the use of composite measures in the context of psychological assessments within clinical contexts, current validity research does not address how scores from multiple measures might be combined or considered jointly in the evaluation of teachers. The validity argument for inferences based on multiple measures introduces an additional layer of complexity because support is needed for the composite inference and not simply inferences based on individual measures. As almost all of the current teacher evaluation schemes are contemplating some use of multiple measures, more specific guidance is needed.

TEACHER OR TEACHING QUALITY?

The current policy press is to develop measures that allow for inferences about *teacher* effectiveness. Using particular measures, the goal is to be able to make some type of claim about the qualities of a teacher. Yet, to varying degrees, the measures we examine do not tell us only about the teacher. A broad range of contextual factors also contributes to the evidence of the *teaching* quality, which is more directly observable.

To illustrate why context affects the validity of what inferences can be made from the observation of a performance, consider a scenario from medicine. Assume that under the same conditions, two surgeons would operate using the same processes and their respective patients would have the same outcomes. But, as described in the following example, such simplifying assumptions that conditions are invariant often do not hold.

> Imagine Two Surgeons
>
> We would like to evaluate them on the quality of their surgical skills using multiple measures. We will use the size of the scar, the rate of infection, the quality of pain management, and patient satisfaction as our measures of the quality of their surgical skills. One is in Miami, Florida; the other is in Moshe, Tanzania. Both must remove a benign tumor from a 53-year-old man's abdomen. The surgeon in Miami has a #10 blade steel scalpel that is designed for cutting muscle and

> skin. The surgeon in Moshe has a well-sharpened utility knife that is used for a range of surgical purposes. The excision in Miami will occur in a sterile operating room with no external windows, fans and filters to circulate and clean the air, an anesthesiologist, and a surgical nurse. The excision in Moshe will occur in a clean operating room washed with well water and bleach, windows opened a crack to allow the breeze to circulate the stiflingly hot air, no fans or filters, and a nurse borrowed from the pediatrics unit because she was the only available help.

It is possible that neither patient will get an infection and both will be satisfied with the care they receive. But it is also possible, perhaps likely, that the patient in Miami will have a smaller scar than the Moshe patient, due to the knife used, and that the Miami patient will have better pain management than the Moshe patient because of access to an anesthesiologist. So even in one surgery, one would expect the Miami surgeon to carry out a more effective surgery than the Moshe surgeon. And over a number of years, as these surgeons do 100 similar surgeries, it becomes increasingly likely that the Moshe surgeon will have poorer surgical outcomes than the Miami surgeon.

But has the quality of each surgeon's respective skills really been judged? The quality of medical care the two patients have received has been evaluated. Are surgical skill and medical care the same thing? Perhaps all that has really been learned is that if someone has a tumor, he or she would like it removed in Miami, not in Moshe. The point is that even in medicine, with its more objective outcomes of scar size and infection rate, it is not always so obvious to attribute surgical outcomes to the surgeon alone. There are many factors beyond the surgeon's control that can contribute to her success. Of course, the best conditions in the world will not, over time, make an incompetent surgeon appear to be expert.

Now imagine walking into a classroom and observing a lesson in order to make judgments about a teacher. How much of what is seen is under the sole control of the teacher, and how much might be attributable to contextual factors that influence what the teacher does and how well students learn? For example, although one can judge the quality of the content being taught, that content is frequently influenced by district-imposed curricula and texts. Social interactions that occur among students are certainly a function of the teacher's establishment of a classroom climate, but students also bring a set of interpersonal dynamics into the classroom. Some teachers may design homework assignments or assessments, but others may be compelled to use assessments and assignments developed by the school district. How do parental actions differentially support the intentions of teachers? The point is that it may be impossible to disentangle the individual teacher from all of the classroom interactions and outside variables that influence student outcomes (Braun, 2005a). Outside the classroom, there are additional contextual effects (e.g., interactions within schools and the larger community) that are difficult to isolate (e.g., Pacheco, 2008). At a minimum, if we are to ascribe causal attribution for student learning to teachers, we must attempt to understand these complexities and use analytic processes and methods that can educate stakeholders about the quality and limitations of those causal attributions.

The Purposes of Evaluating Teaching

For a range of reasons, there has been a push for improved teacher evaluation models. The push is strong, in part, because it comes from different constituencies with varying purposes for evaluating teaching. These purposes include educational accountability, strategic management of human capital, professional development of teachers, and the evaluation of instructional policies. The confluence of underlying constituencies and a wide range of purposes have led to intense research and development activity around teacher effectiveness measures.

The first and perhaps most broadly agreed on purpose for teaching evaluation is public accountability. The time period during which this chapter is being written is an era of a pervasive emphasis on educational accountability. Concerns about persistent achievement gaps between Black and White, poor and rich, and English language speakers and

English language learners, coupled with concerns about U.S. academic performance relative to other countries (Gonzales et al., 2008; Programme for International Student Assessment [PISA], 2006), have led policymakers to implement unprecedented policies that focus on achievement and other measurable outcomes. Nowhere is this press for a public accounting on measurable outcomes stronger than in the K–12 accountability policy of the No Child Left Behind revision of the Elementary and Secondary Education Act in 2002 (No Child Left Behind Act, 2002).

Supported by a growing body of research that identifies teachers as the major school-related determinant of student success (Nye, Konstantopoulos, & Hedges, 2004; Raudenbush, Martinez, & Spybrook, 2007), perhaps it was only a matter of time before the public accounting of student performance gave way to a public accounting of teacher performance. The purpose of teaching evaluation in this way of thinking is to document publicly measurable outcomes that drive decision making and ensure the public's financial investment in teachers is maximized. It is important to recognize that out-of-school factors continue to be most predictive of student outcomes; but for the variance that can be accounted for by schools, teachers are the greatest source of variation in student test score gains (Nye et al., 2004; Raudenbush, 2004). Estimates of the size of teachers' contribution vary with the underlying analytic model employed (Kyriakides & Creemers, 2008; Luyten, 2003; Rowan, Correnti, & Miller, 2002).

Earlier efforts to account publicly for teaching quality have not been particularly useful or insightful. Characteristics valued in existing compensation systems, such as postbaccalaureate educational course-taking, credit and degree attainment, and years on the job have modest associations with student achievement (e.g., Clotfelter, Ladd, & Vigdor, 2005; Harris & Sass, 2006; T. J. Kane, Rockoff, & Staiger, 2006).[1] In addition, widely used surface markers of professional preparation, such as certification status and coursework, only weakly predict actual teaching quality (Goe, 2007; Wayne & Youngs, 2003).

Stakeholders have grown increasingly frustrated with the lack of an apparent relationship between student achievement and measures used to evaluate teachers (e.g., Weisberg et al., 2009). This has led to a perspective with a far more empirical view of what defines effective teaching. Largely emanating from individuals who are not representative of the traditional educational research and measurement communities, another goal of teaching evaluation has become prominent—the strategic management of human capital (Odden & Kelley, 2002). This view rests on basic economic approaches to managing the supply of teachers by incentives and disincentives for individuals with specific characteristics. The logic suggests that if the supply of "effective" teachers can be increased by replacing "ineffective" teachers, overall achievement would increase and the achievement gap would decrease (Gordon, Kane, & Staiger, 2006). In this view, the evaluation of teaching is the foundation for managing people via retention, firing, placement, and compensation policies (Heneman, Milanowski, Kimball, & Odden, 2006).

A parsimonious definition of teaching quality guides the measurement approach of human capital management. This is characterized in the following remark by Hanushek (2002): "I use a simple definition of teacher quality: good teachers are ones who get large gains in student achievement for their classes; bad teachers are just the opposite" (p. 3). Hanushek adopted this definition because it is empirically based; straightforward; and in his and others' views, tractable. Most of all, such a definition avoids defining quality by the execution of particular teaching processes or the possession of specific teacher characteristics, factors that have had modest, if any, relationships to valued outcomes (e.g., Cochran-Smith & Zeichner, 2005).

Although recent approaches to the strategic management of human capital have raised the stakes substantially for how teacher evaluations are used, most policies broaden teacher evaluation to include other factors besides student achievement growth.

[1]The relationship of student achievement growth and teacher experience does increase for the first several years of teaching but levels off after only a few years (e.g., Nye et al., 2004).

Nevertheless, student achievement growth estimates typically are a dominant factor in making determinations of effectiveness.

In addition to strategic management of human capital, teacher evaluation has been viewed as a means for improving individual and organizational capacity. There have been longstanding concerns that the professional development of teachers, beginning even in preservice, is disconnected from the particular needs of individual teachers and inconsistent with understandings of how teachers learn (Borko, 2004) and the support they need to teach well (Johnson et al., 2001; Johnson, Kardos, Kauffman, Liu, & Donaldson, 2004; Kardos & Johnson, 2007). There is also increasing research that documents how organizational variables—the alignment of curriculum, the presence of professional learning communities and effective leadership, and the quality of reform implementation—are related to the nature and quality of teaching (Honig, Copland, Rainey, Lorton, & Newton, 2010). With capacity building as a goal, teaching evaluation can be a tool that can diagnose the practices most in need of improvement. The goal of teaching evaluation from this perspective is to improve what teachers and schools know and are able to do around instruction. The measures used toward this end vary dramatically from low-inference checklists of desired behaviors to high-inference holistic rubrics of underlying teaching quality values to school-level measures of teaching contexts (Hirsch & Sioberg, 2010; Kennedy, 2010).

Finally, researchers and evaluators use teaching evaluation to assess whether and how various education policies are working. Deriving from both measurement and evaluation perspectives, teaching evaluation has been used to investigate the degree to which school and curricular reforms and their implementation influence instruction (e.g., Rowan, Camburn, & Correnti, 2004; Rowan & Correnti, 2009; Rowan, Jacob, & Correnti, 2009), the impacts of professional development (Desimone, Porter, Garet, Suk Yoon, & Birman, 2002; Malmberg, Hagger, Burn, Mutton, & Colls, 2010), and how particular policies such as academic tracking (Oakes, 1987) and school and class size (Molnar et al., 1999) relate to teacher practice. Often, the types of measures used for this purpose are logs or other surveys that ask teachers to report on the frequency of important activities or practices. By evaluating teaching, researchers and evaluators can assess the degree to which policies intended to shape teaching and learning are working as intended.

In this chapter, the classes of measures that can be used to support the evaluation of teaching for one or more purposes are described: educational accountability, strategic management of human capital, professional development of teachers, and the evaluation of instructional policies. The next section looks briefly at the history of assessing teaching quality and considers the ways in which these multiple purposes have played out in recent history.

A Selective History of Assessing Teaching Quality

Only a few years ago, S. Wilson (2008) characterized the U.S. national system of assessing teacher quality as "undertheorized, conceptually incoherent, technically unsophisticated, and uneven" (p. 24). Although Wilson focused on the *system* of assessments used to characterize teacher quality, the same characterization can be leveled at the constituent measures and practices that make up what she referred to as a "carnival of assessment" (p. 14). Three dominant assessment purposes at the "carnival" are described, each of which renders judgments about preservice, in-service, and master teaching, respectively. Across the three purposes, there are both strengths and weaknesses that lay the foundation for understanding current research and development activities.

By far, the most common purpose of formal assessment in teaching occurs for beginning licensure, in which the state ensures that candidates have sufficient knowledge, typically of content and basic skills, so that the state can warrant that the individual will "do no harm."[2] These tests have almost always been, and continue to be, standardized assessments that require teacher candidates to meet a particular state-established passing standard to be

[2]That is, in the legal context of licensure, failure to demonstrate sufficient knowledge or skill on an assessment would present some probability of causing *harm* in the workplace (M. T. Kane, 1982).

awarded a license (S. M. Wilson & Youngs, 2005). Tests have differed in terms of the proportion of multiple-choice versus constructed-response questions, whether they are paper-and-pencil or computer-delivered, whether they are linear or adaptive, and the methodology by which passing standards are set. State licensure requirements vary in terms of the number and types of tests required and are not guided by a coherent theory of teaching and learning. Tests are designed most often by testing companies in collaboration with states. Although this system results in adequate levels of standardization within a state, the tests are criticized as being disconnected from teaching practice and based on incomplete views of teaching (e.g., Klein & Stecher, 1991).

Such tests are designed to support inferences about the knowledge of prospective teachers. They publicly account for what teachers know prior to entering a classroom. They deliberately have not been designed to encourage inferences about the quality of teaching, although the implicit assumption on the part of many is that higher scores are associated with higher levels of teaching proficiency, however defined. When researchers have investigated this assumption, the evidence of any relationship has been weak, at best (Buddin & Zamarro, 2009; Goldhaber & Hansen, 2010; National Research Council, 2008).

A number of states more recently adopted the view that, to attain a full license, there ought to be some direct evidence of teaching. Treating the initial license as provisional, they adopted measures that involved more direct evidence of teaching. Almost always grounded in professional standards for teachers, assessments included live classroom observations, interviews, and teacher-developed portfolios that contain artifacts of classroom practice such as planning documents, assignments, and videotapes (California Commission on Teacher Credentialing, 2008; Connecticut State Department of Education, Bureau of Program and Teacher Evaluation, 2001; Educational Testing Service [ETS], 2001). These assessments are intended to provide both public accountability and formative information about how to improve individuals' capacity while continuing to adhere to the "do no harm" principle. Pass rates, particularly given multiple opportunities to complete the assessment as is characteristic of these systems, are very high (more than 95%; e.g., Connecticut State Department of Education, Bureau of Program and Teacher Evaluation, 2001; Ohio Department of Education, 2006).

Taken together, the licensure testing processes serve the function of preventing a relatively small proportion of individuals from becoming teachers, but they do not support inferences about the quality of teachers or teaching beyond minimal levels of competence. Furthermore, because these instruments are disconnected from practice either by not being able to sort teaching or not being close enough to practice to provide information about what a teacher is and is not able to do beyond minimal levels, this group of assessment practices provides modest accountability and capacity-building information.

In addition to supporting judgments about individual teacher candidates, beginning teacher assessment is also influenced by teacher education program accreditation. Almost all states use some combination of teacher testing and program accreditation to regulate and hold programs accountable for the quality of teachers entering classrooms (S. M. Wilson & Youngs, 2005). Accreditation has been governed by state and regional accrediting agencies as well as by two national organizations: National Council for Accreditation of Teacher Education (NCATE) and Teacher Education Accreditation Council (TEAC).[3] Accreditation requirements vary but generally include a site visit and a paper review of program offerings, program coherence, and the alignment of program standards with national organizations' subject matter teaching standards. In some accreditation processes, programs must provide evidence that graduates can teach competently and have acquired relevant subject matter knowledge and teaching experiences. That evidence can come from whatever assessments the program uses, and there are few, if any, common assessments. These processes require much of teacher education programs (e.g., Barnette &

[3]As of October 2010, NCATE and TEAC announced the merger of the two organizations into a unified body, The Council for the Accreditation of Educator Preparation.

Gorham, 1999; Kornfeld, Grady, Marker, & Ruddell, 2007; Samaras et al., 1999) and may produce changes in program structure and processes (Bell & Youngs, 2011); however, there is no research that documents the effects of accreditation on preservice teacher learning or K–12 pupil learning.

A second dominant assessment purpose occurs once teachers are hired and teaching in classrooms. States and districts typically set policies concerning the frequency of evaluation and its general processes. In states with collective bargaining, evaluation is often negotiated by the administration and the union. Despite the variety of agencies that have responsibility for the content of annual evaluations, evaluations are remarkably similar (Weisberg et al., 2009). They are administered by a range of stakeholders—coaches, principals, central office staff, and peers—and use a wide range of instruments each with its own idiosyncratic view of quality teaching.[4] Although evaluations apply to all teachers, the systematic and consistent application of evaluative judgments are rare (e.g., Howard & Gullickson, 2010).

Whereas traditional assessment practices for preservice teachers have had standards but are disconnected from teaching practice, in-service assessment practices have been connected to practice but lack rigorous standards. This has led in-service teaching evaluation to be viewed as a bankrupt and uninformative enterprise (Toch & Rothman, 2008; Weisberg et al., 2009). Evaluations are often viewed as bureaucratic functions that provide little or no useful information to teachers, administrators, institutions, or the public.

Howard and Gullickson (2010) have made the case that teacher evaluation efforts should meet the Personnel Evaluation Standards (Gullickson, 2008) that include the following: *propriety standards,* addressing legal and ethical issues; *utility standards,* addressing how evaluation reports will be used and by whom; *feasibility standards,* addressing the practicality and feasibility of evaluation systems; and *accuracy standards,* addressing the validity and credibility of the evaluative inferences. These standards can be equally applied to particular measures within an evaluative system.

It is fair to say that there is a substantial chasm between the values expressed in these standards and the state of teacher evaluation practice for preservice and in-service teachers. It is rare to find an evaluation system in which there is any information collected as to the validity or reliability of judgments. Often, principals and other administrators are reluctant to give anything but acceptable ratings because of the ensuing responsibilities to continue to monitor and support individuals determined to be in need of improvement. It is extremely rare that teachers—tenured or not—are removed from their jobs simply because of poor instructional performance. Routinely, the propriety and accuracy of the evaluation is challenged at great cost to the school system (Pullin, 2010).

Current policies are attempting to transform this historic state of affairs by largely defining teaching effectiveness as the extent to which student test scores improve on the basis of year-to-year comparisons. The methods that are used necessarily force a distribution of individual teachers and are explicitly tied to student outcomes. The details of these methods and the challenges they present are discussed in subsequent sections of this chapter.

One other major teacher evaluation purpose that was first implemented in the mid-1990s is the National Board for Professional Teaching Standards (NBPTS) certification system. Growing out of the report *A Nation Prepared: Teachers for the 21st Century* (Carnegie Forum on Education and the Economy, 1986), a system of assessments was designed to recognize and support highly accomplished teachers. All NBPTS-certified teachers are expected to demonstrate accomplished teaching that aligns with five core propositions about what teachers should know and be able to do as well as subject- and age range–specific standards that detail the characteristics of highly accomplished teachers (NBPTS, 2010).

The architecture of the NBPTS system has been described by Pearlman (2008) and was used to guide

[4]Annual teaching evaluations are generally idiosyncratic within and across districts; however, there are examples of more coherent district-level practices in places like Cincinnati and Toledo. Increasingly, as a part of the Teacher Incentive Fund grants, districts are experimenting with pilot programs that have higher technical quality.

the development of 25 separate certificates, each addressing a unique combination of subject area and age ranges of students. For all certificates, teachers participate in a year-long assessment process that contains two major components. The first requires teachers to develop a portfolio that is designed to provide a window into practice. Portfolio entries require teachers to write about particular aspects of their practice as well as include artifacts that provide evidence of this practice. Artifacts can include videos and samples of student work. In all cases, teachers are able to choose the lesson(s) they want to showcase, given the general constraints of the portfolio entry. Examples of a portfolio entry include videos of the teacher leading a whole-class discussion or teaching an important concept.

The second major component of NBPTS certification is the assessment center. Candidates go to a testing center and, under standardized testing conditions, respond to six constructed-response prompts about important content and content pedagogy questions within their certificate area. To achieve certification, candidates need to attain a total score across all assessment tasks that exceeds a designated passing standard. On balance, research suggests that the NBPTS system is able to identify teachers who are better able to support student achievement—as measured by standardized test scores—than are teachers who attempt certification but do not pass the assessment, but the differences are quite modest (National Research Council, 2008).

The states in which teachers have been most likely to participate in the NBPTS system are those that have provided monetary rewards or salary supplements for certification. This has led to NBPTS being very active in a relatively small number of states, with only limited participation in other states. In contrast to assessment policies that shape preservice and in-service teaching, NBPTS takes a nuanced and professional view of teaching via a standardized assessment system that is tied to teaching practice. However, it is voluntary, touches relatively few teachers in most states, and is expensive.

Although this discussion does not cover all teacher evaluation practices, it does provide a synopsis of the most common formal assessment and evaluation systems for teachers. Taken together, the evidence suggests that even the most common assessment practices have had a modest impact on the structures and capacity of the system to improve educational performance. Looking across the practices, there is no common view of quality teaching, and sound measurement principles are missing from at least some core practices of in-service evaluation. These findings, along with political reluctance to make evaluative judgments (e.g., Weisberg et al., 2009), have led many researchers and policymakers to conclude that the measures that make up the field's most common assessments will be unable to satisfy the ambitious purposes of accountability, human resource management, and instructional improvement that are driving current policy demands around evaluation.

Thus, the chapter next reviews measures the field is developing and implementing to support purposes ranging from accountability to capacity building. The primary features of different classes of measures, the nature of inferences they potentially can support, and current validation approaches and challenges are described.

Conceptualizing Measures of Teaching Quality

Teaching quality is defined in many different ways and operationalized by the particular sets of measures used to characterize quality. Every measure brings with it, either explicitly or implicitly, a particular perspective as to which aspects and qualities of teaching should receive attention and how evidence ought to be valued. For example, there have been heated political arguments about whether dispositions toward teaching ought to be assessed as part of teacher education (Hines, 2007; Wasley, 2006). Although there is general agreement that the impact on students ought to be a critical evaluative consideration, the indicators of impact are not agreed upon. Some are satisfied with a focus on subject areas that are tested in schools. Others want both to broaden the academic focus and emphasize outcomes that are associated with mature participation in a democratic society (Koretz, 2008; Ravitch, 2010).

Although reasonable people disagree about what distinguishes high-quality teaching, it is important

to identify clearly the constructs that comprise teaching quality and how those constructs may be understood relative to the measures used in evaluation systems. Figure 20.1 describes a model we have developed that presumes that teaching quality is interactional and constructive. Within specific teaching and learning contexts, teachers and students construct a set of interactions that is defined as teaching quality. Six broad constructs make up the domain of teaching quality. These are teachers' knowledge, practices, and beliefs, and students' knowledge, practices, and beliefs. The domain of teaching quality and by extension the constructs themselves are intertwined with critical contextual features, such as the curriculum, school leadership, district policies, and so on. Therefore, by definition, specific instruments measure both context and construct. As can be seen in the figure, instruments may detect multiple constructs or a single teaching quality construct. For example, observation protocols allow the observer to gather evidence on both teacher and student practices, whereas assessments of content knowledge for teaching only measure teachers' knowledge. Finally, the figure suggests that any one measure (e.g., a knowledge test or value-added models [VAM]) does not capture the whole domain of teaching quality.

In many fields, it is reasonable to expect that particular classes of measures are associated with particular stages of educational or professional development. For teaching, that has been partially true, particularly with content knowledge measures being used as a requirement to attain a teaching license. By and large, however, the measures reviewed here are being considered for use throughout the professional span during which teachers are assessed. At the time of the writing of this chapter, how the measures actually are used to meet particular assessment purposes remains to be seen. Nevertheless, because of the lack of any inherent relationship between category of measure and particular use, the remainder of this chapter is organized by construct focus rather than assessment purpose.

MEASURES OF TEACHING QUALITY

In this section, an overview of measures that have been developed to support inferences about constructs associated with teaching quality is presented.

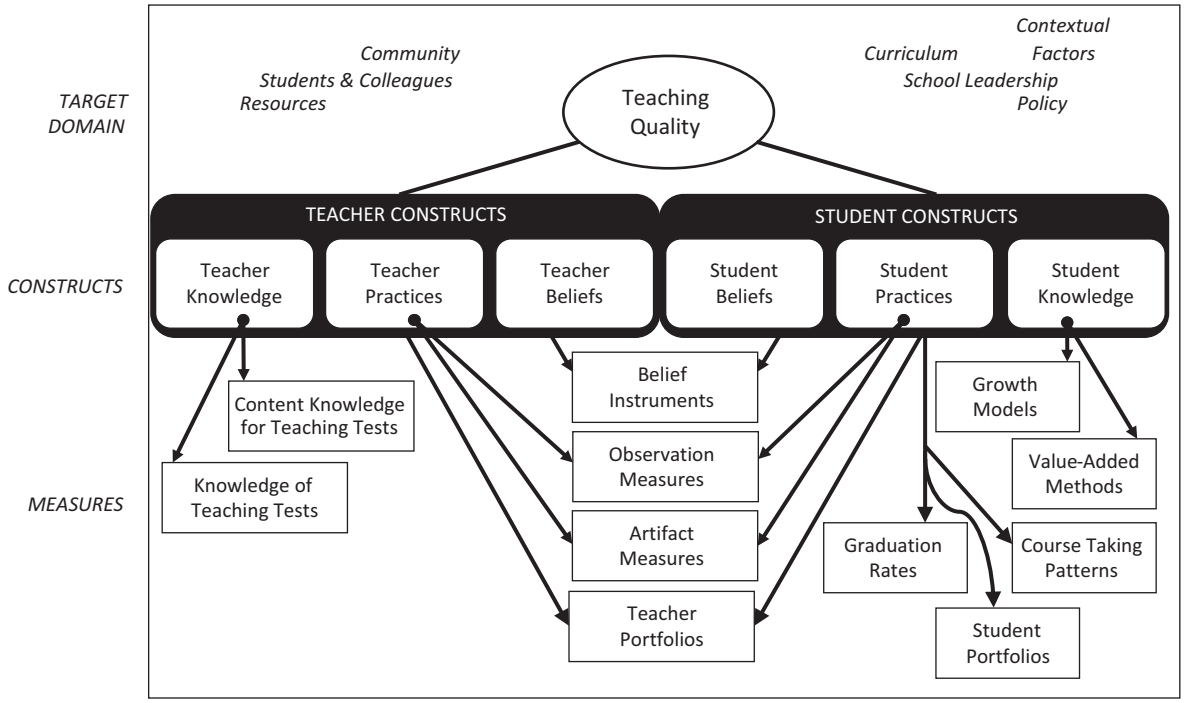


FIGURE 20.1. The contextual factors, constructs, and measures associated with teaching quality.

For each set of measures, its core design characteristics, the range of uses to which it has been put, and the status of evidence to support a validity argument for the evaluation of teaching are reviewed.

Teacher Knowledge

Knowledge of content. Knowledge of content has been a mainstay of the teacher licensure process since the 1930s, with the advent of the National Teacher Examinations (Pilley, 1941). With the requirement for highly qualified teachers in the reauthorization of the Elementary and Secondary Education Act (No Child Left Behind Act, 2002), all states now require teachers to demonstrate some level of content knowledge about the subjects for which they are licensed to teach.

These assessments typically consist of multiple-choice questions that sample content derived from extant disciplinary and teaching standards and then confirmed through surveys of practitioners and teacher educators. Individual states set passing scores for candidates based on standard-setting processes (Livingston & Zieky, 1982) that are used to define a minimum-level "do no harm" threshold. Some states also require tests that assess knowledge of pedagogy and content-specific pedagogy. Although some of these tests may include constructed-response formats, the basic approach to design and validation support is similar for both content and pedagogical tests.[5]

The validity argument for these kinds of tests has long been a source of debate. M. T. Kane (1982) discussed two possible interpretations: one concerned with the ability of the licensure test to predict future professional performance and the other to evaluate current competence on a set of skills and knowledge that was deemed necessary but not sufficient for professional practice. M. T. Kane (1982) argued that the latter interpretation was appropriate for licensure tests as any single instrument would be insufficient to capture the set of complex and coordinated skills, understandings, and experiences necessary for professional competence.

In endorsing the much more limited competence interpretation, M. T. Kane (1982) argued that establishing content validity was the critical task for a licensure test validity argument. Evidence is expected to demonstrate the adequacy of content needed for minimal job performance, both in terms of content representation and expectations for meeting the passing standard. Processes that include job analysis and standard-setting studies typically are used to provide such evidence. The adequacy of scores is typically supported through standard psychometric analyses that include test form equating, reliability, scaling, differential item functioning (DIF), and group performance studies. Other scholars have agreed that it is both inappropriate and infeasible to expect a broader validity argument (e.g., Jaeger, 1999; Popham, 1992).

Even under this relatively constrained set of requirements, the status of validity evidence in practice is uneven. In its 2001 report, the National Research Council reviewed the validity evidence of the two primary organizations that design, develop, and administer these assessments (Mitchell, Robinson, Plake, & Knowles, 2001). ETS[6] was viewed as having evidence to support the content validity argument, although some assessments were using studies that were dated. Information about National Evaluation Systems (NES)[7] tests was typically unavailable, and so, the study panel concluded that for a very substantial amount of teacher licensure testing, the available evidence did not satisfy even the most basic requirements of available information articulated in the *Standards for Educational and Psychological Testing* (AERA et al., 1999).

M. T. Kane's (1982) position that content validation is by itself sufficient to establish the validity of licensure assessments has been argued against

[5]Although all states require demonstrations of content knowledge, some also require candidates to pass assessments of basic knowledge of reading, writing, and mathematics. We do not include these tests in our analysis because these instruments test knowledge and skills that are equally germane for any college student, not just teacher candidates.

[6]The authors of this chapter were both employees of ETS as this chapter was written. The statements included here are a description of the conclusions of the National Research Council study report (Mitchell et al., 2001), *Testing Teacher Candidates: The Role of Licensure Tests in Improving Teacher Quality*. We believe our statements are a fair representation of the study findings.

[7]National Evaluation Systems was acquired by Pearson Education in 2006 and is now known as Evaluation Systems of Pearson.

strongly by other experts (e.g., Haertel, 1991; Haney, Madaus, & Kreitzer, 1987; Moss & Schutz, 1999; Pullin, 1999). A number of researchers and policymakers, including the authors of the National Research Council study (Mitchell et al., 2001), have argued that these assessments ought to be evaluated using the predictive criterion interpretation, including a demonstration of a relationship between scores on the licensure test and other measures associated with teaching. To that end, researchers have conducted studies relating scores on teacher licensure assessments to student gains in achievement by studying practicing teachers who varied on their licensure test scores, including those who would not have met the passing standard in one state even if they scored sufficiently high to teach in another. There is some evidence of a quite modest relationship between licensure test scores and student outcomes (e.g., Goldhaber & Hansen, 2010). Gitomer and Qi (2010), however, observed that the licensure tests were successful in identifying individuals who performed so substantially below the passing standard that such individuals would not have ever become practicing teachers in any locale and, thus, would not have been part of the distribution studied by Goldhaber and Hansen. Because these individuals do not attain a license to teach, any studies examining the relationships between test scores and other outcomes are attenuated.

In part because content knowledge tests have been used so widely, there is a large body of evidence demonstrating disparate impact for minority candidates. Scores and passing rates are significantly lower for African American candidates than White candidates (Gitomer, 2007; Gitomer, Latham, & Ziomek 1999; Gitomer & Qi, 2010), which raises questions about the validity of the assessments and whether bias is associated with them. Although test developers attempt to ensure fairness through their test development and analysis processes (e.g., DIF analyses), it is imperative that research continue not only to examine issues of bias but also to pursue strategies to mitigate unequal educational opportunities that many minority candidates have experienced (e.g., Mitchell et al., 2001; National Research Council, 2008).

Knowledge of content for teaching. Teaching involves much more than simple mastery of content knowledge. Shulman (1986) argued persuasively that teachers also needed to master a body of knowledge he identified as pedagogical content knowledge (PCK). Shulman argued that PCK involves pedagogical strategies and representations that make content understandable to others and also involves teachers grasping what makes particular content challenging for students to understand, what kinds of conceptions and misconceptions students might have, and how different students might interact with the content in different ways.

Building on Shulman's (1986) ideas, Hill, Ball, and colleagues focused on mathematics and developed theory and assessments of what they called mathematical knowledge for teaching (MKT; Ball, Thames, & Phelps, 2008). MKT attempts to specify the knowledge of mathematical content that is used in practice, differentiating the advanced subject knowledge that one might learn as a student majoring in a college discipline from the particular forms of knowledge that teachers need to help their students learn concepts in K–12 education. Content knowledge for teaching (CKT) is the more general term applied to this type of knowledge as it used across different subject matter domains (e.g., science, social studies). CKT incorporates what Shulman called PCK and further specifies both the content knowledge and the PCK that teachers need to know in particular subject areas. The argument that accompanies CKT suggests that teachers must know mathematics differently than someone who uses math in her daily life but is not charged with teaching children math. For example, a common task of teaching requires that teachers assess the degree to which certain problems allow students to practice a particular math objective. In Figure 20.2, the teacher must be able to recognize whether a proportion can be used to solve the word problem. Although adults may use proportions in their professional or personal lives, teachers must be able to look at problems and determine whether that problem can be solved in a specific way that meets a learning objective.

Ball et al. (2008) highlighted six forms of CKT that fall into two categories—content knowledge

Mr. Sucevic is working with his students on understanding the use of proportional relationships in solving problems. He wants to select some problems from a mathematics workbook with which his students can practice. For each of the following problems, indicate whether or not it would be answered by setting up and solving a proportional relationship.

	Would Be Answered by Setting Up and Solving a Proportional Relationship	Would Not Be Answered by Setting Up and Solving a Proportional Relationship
A) Cynthia is making cupcakes from a recipe that requires 4 eggs and 3 cups of milk. If she has only 2 eggs to make the cupcakes, how many cups of milk must she use?		
B) John and Robert are each reading their books at the same rate. When John is on page 20, Robert is on page 15. What page will John be on when Robert is on page 60?		
C) Julie and Karen are riding their bikes at the same rate. Julie rides 12 miles in 30 minutes. How many miles would Karen ride in 35 minutes?		
D) Rashida puts some money into an account that earns the same rate each month. She does not remove or add any money to the account. After 6 months, the balance in the account is $1,093.44. What is the balance in the account after 12 months?		
E) A square with area 16 square units can be inscribed in a circle with area 8π square units. How many square units are in the area of a square inscribed in a circle that has area 24π square units?		

FIGURE 20.2. A sample question from MET Mathematics 6–8 (Gitomer & Phelps, 2012).

and PCK. Each of the main categories has three subcategories. Content knowledge is composed of common content knowledge, specialized content knowledge, and horizon content knowledge. Specialized content knowledge is knowledge that enables work with students around content, but not knowledge that other professionals using the same content (e.g., mathematics) in their jobs might find important. For example, a teacher needs to not only carry out a mathematical operation (e.g., dividing fractions) but also to understand why the operation works so that different student solutions can be understood as being mathematically reasonable or not. Specialized content knowledge contrasts with common content knowledge, which is knowledge held by individuals who use that content in their work and personal lives. Horizon content knowledge involves an understanding of how different content is interrelated across curricular topics both within and across school years.

PCK, the second organizing category of knowledge, is composed of knowledge of content and students, knowledge of content and teaching, and knowledge of content and curriculum. Knowledge of content and students combines content knowledge and knowledge of how students interact with and learn the subject. It includes, for example, knowledge of what aspects of a subject students are likely to find difficult, errors students might make, and difficulties students might encounter in understanding a subject. Knowledge of content and teaching includes knowledge of the best examples to use, how to link subject-specific tasks, and ways of responding to students' ideas and confusion that will develop their understanding of the subject. Finally, knowledge of content and curriculum focuses on knowledge of how to sequence and organize a subject and of the material programs that can be used to support students' developing understanding of the subject.

Hill, Schilling, and Ball (2004) described the developmental processes for constructing items of these types and also provided information about the psychometric quality and structure of assessment

forms built with these items. Schilling and Hill (2007) have laid out a validity argument for these kinds of assessments and have conducted a research program to marshal evidence to evaluate the argument. To date, these assessments have been used in the context of research, particularly in the context of examining the impact of professional development and curricular interventions. They have not been used as part of licensure or other high-stakes testing programs. Thus, the validity argument pertains to use as a research tool.

In one study, Hill, Dean, and Goffney (2007) conducted cognitive interviews of problem solutions by teachers, nonteachers, and mathematicians. Although they found that mathematical knowledge itself was critically important to solving the problems, they observed important differences that were not simply attributable to content knowledge. Mathematicians, for example, sometimes had difficulty interpreting nonstandard solutions, the kinds of solutions that students often generate. Although mathematicians could reason their way through problems, it was teachers who could call on their prior experiences with students to reason through other problems. Krauss, Baumert, and Blum (2008) developed another measure of PCK and also found strong but not uniform relationships with content knowledge—in some cases, teachers brought unique understandings that allowed them to solve problems more effectively than others who had far stronger mathematical content knowledge. Other studies have found modest relationships between CKT measures and student achievement gains (Hill, Rowan, & Ball, 2005) and relationships with judgments of classroom instruction through observation (Hill et al., 2008). The lack of studies that address questions of impact on subgroups of teachers (e.g., special education teachers, teachers of color, or teachers of English language learners) likely is due to the purposes and scope of the existing research studies.

The studies to date have typically relied on relatively small samples. Studies currently being conducted will yield data based on far larger samples and broader sets of measures of teacher quality (e.g., Bill and Melinda Gates Foundation, 2011a). To date, there has been only limited work in other content domains (e.g., Phelps, 2009; Phelps & Schilling, 2004), but assessments are being developed and tested in English language arts. Most validity work has been done on MKT, not the more general CKT. Given the use of MKT as a research tool, there is a relatively strong validity argument. However, the validity argument for MKT for other uses is modest (e.g., teacher preparation program evaluation) but growing. The validity argument for assessments of CKT, used in both research and for personnel decisions, is nascent but also growing.

Teacher Practices

Observations. Scholarship on observation protocols goes back to the turn of the 20th century (Kennedy, 2010). The actual practice of individuals with authority using some type of protocol to make evaluative decisions about a teacher likely dates back even further. Kennedy (2010) suggested that for more than half of the 20th century the protocols in use have been general, poorly defined, idiosyncratic, heavily subjective, and often focused on teachers' personal characteristics rather than on teaching.

Given the view of teaching as one involving socially and culturally situated interactions between teachers and students to support the construction of knowledge, instruments that are unable to detect these types of interactions are not reviewed. This means that the instruments from the productive history of process–product research in the 1970s and 1980s that used observation protocols to assess teaching quality are not included (for a review of this research, see Brophy & Good, 1986). Instead, the focus is on the relatively new and small number of instruments and associated research that has been developed and used over roughly the past 25 to 30 years. These instruments are designed to measure whole-class instruction (e.g., not tutoring situations) and adopt the view that teaching and learning occur through interactions that support the construction of knowledge.

The observation protocols currently in use generally adhere to the following description: The protocol begins with an observer developing a record of evidence from the classroom for some defined segment of time, typically without making any evaluative judgments. At the end of the segment, observers use a set of scoring criteria or rubric that typically

includes a set of Likert scales to make both low- and high-inference judgments based on the record of evidence. Those judgments result in numerical scores. Although some of the protocols have been used to evaluate thousands of teachers (e.g., Charlotte Danielson's Framework for Teaching has been the most widely used), the protocols have rarely been used for summative consequential decisions, although this is changing rapidly. Despite the fact that many districts are considering or have already begun using these observation protocols for consequential decisions, there is still much not known about the strength of the validity argument for these protocols as a group as well as the strength of the validity argument for individual protocols. Although there are exceptions, the instruments have been used in both live and video-based observation settings. Bell et al. (2012) have recently used an argument approach to evaluate the validity of one observation protocol.

Protocols tend to fall into two broad categories—protocols for use across subject areas and those intended for use in specific subject areas (S. K. Baker, Gersten, Haager, & Dingle, 2006; Danielson, 1996; Grossman et al., 2010; Hill et al., 2008; Horizon Research, 2000; Pianta, La Paro, & Hamre, 2007; Taylor, Pearson, Peterson, & Rodriguez, 2003). There are subject-specific protocols in mathematics, science, and English language arts, but none are evident for social studies classrooms (e.g., social studies, history, government, geography, etc.). There are more protocols for use at the elementary grades than the secondary ones. Many of the subject-specific protocols have been studied in K–3, K–5, or K–8 classrooms, so it is unclear whether or how the protocols might function differently in high school classrooms. These protocols reflect a particular perspective on teaching quality—some privilege a belief in constructivist perspectives on teaching, and others are more agnostic to the particular teaching methods used.

Observation protocols are generally developed and vetted within a community of practice that has a corresponding set of teaching standards (Danielson & McGreal, 2000; Gersten, Baker, Haager, & Graves, 2005; La Paro, Pianta, & Stuhlman, 2004; Piburn & Sawada, 2000). Because much of the research on these protocols has happened in the context of university-based research projects, the raters themselves are often graduate students or faculty members. With this rater group, trainers are able to teach raters to see teaching and learning through the lens of their respective protocol at acceptable levels of interrater agreement (e.g., Hill et al., 2008). Initial qualification of raters typically requires agreement with master codes at some prespecified level (e.g., 80% exact match on a 4-point scale).

Among both researchers and practitioners, the best methods and standards for judging rater agreement on holistic observation protocols are evolving. The most simple and common way of judging agreement is to calculate the proportion of scores on which raters agree. For many protocols, agreement requires an exact match in scores (e.g., Danielson & McGreal, 2000). But for others with larger scales, raters are deemed to agree if their scores do not differ by more than 1 score point (e.g., Pianta et al., 2007). Such models do not take into account the overall variation in scores assigned—raters may appear to agree by virtue of not using more than a very narrow range of the scale. More sophisticated analyses make use of rater agreement metrics that take into account the distribution of scores, including Cohen's kappa,[8] intraclass correlations, and variance component decomposition.

Emerging models attempt to understand a range of factors that might affect rater quality and agreement. For example, in addition to rater main effects, Raudenbush, Martinez, Bloom, Zhu, and Lin (2010) consider how rater judgments can interact with the classrooms, days, and lesson segments observed. To the extent that these variance components (or facets, if *g*-study approaches are used; see Volume 1, Chapter 3, this handbook) can be estimated, it may be possible to develop observation scores that adjust for such rater effects. When using these models, preliminary findings suggest there are substantial training challenges in obtaining high levels of agreement,

[8]It is important to note that kappa can be sensitive to skewed or uneven distributions and, therefore, may be of limited value depending on the particular score distributions on a given instrument (e.g., Byrt, Bishop, & Carlin, 1993).

particularly with higher inference instruments (e.g., Gitomer & Bell, 2012; McCaffrey, 2011). As observation systems are included in evaluation systems, systems will need to ensure not only that raters are certified but also that effective monitoring and continuing calibration processes are in place. In general, there is little or no information provided about whether or how raters are calibrated over time (Bell, Little, & Croft, 2009).

A research literature is now beginning to amass around these observation protocols. Research is being conducted examining the extent to which empirical results support the underlying structure of the instruments (e.g., La Paro et al., 2004) and changes in practice as a result of teacher education (Malmberg et al., 2010) and professional development (Pianta, Mashburn, Downer, Hamre, & Justice, 2008; Raver et al., 2008). A number of studies are now being reported that examine the relationship of observation scores to student achievement gains (Bell et al., 2012; Bill and Melinda Gates Foundation, 2011b; Grossman et al., 2010; Hamre et al., in press; Hill, Umland, & Kapitula, 2011; Milanowski, 2004). Thus, over the next 5 to 10 years, a very strong body of research is likely to emerge that will provide information about the validity and potential of classroom observation tools.

It is important to understand that these protocols are designed to evaluate the quality of classroom practice. As described in Figure 20.1, factors such as curriculum, school policy, and environment, as well as the characteristics of the students in the classroom, are being detected by these observation protocols. Thus, causal claims about the teacher require another inferential step and are not transparent. Furthermore, given the high-stakes uses to which these instruments are being applied, the state of the current validity argument is weak.

Instructional collections and artifacts. A second group of instruments to measure teaching quality has emerged in the past 15 to 20 years. Instructional collections (sometimes referred to as portfolios) and artifacts have a shorter history than observations. Research began in earnest on these types of instruments in the early to mid-1990s with peer-reviewed articles and book chapters beginning to appear in the late 1990s. Thus far that work has produced a relatively small number of instruments used and studied by a relatively small number of researchers. In contrast to observation protocols that were largely designed as professional development tools, the design and development of instructional collections and artifact protocols gave more attention to psychometric quality from the outset. Even so, research remains highly uneven—a moderate number of studies with very small numbers of teachers and a handful of studies with large numbers of teachers. Claims about such protocols as a group should therefore be taken as preliminary.

Instructional collections are evidence collection and scoring protocols that typically involve one or more holistic judgments about a range of evidence that often addresses the multiple constructs that comprise the teaching quality construct in Figure 20.1. Instructional collections draw inferences from evidence that can include lesson plans, assignments, assessments, student work samples, videos of classroom interactions, reflective writings, interviews, observations, notes from parents, evidence of community involvement, and awards or recognitions. These protocols identify what types of evidence the teacher is expected to submit within broad guidelines; the teacher is able to choose the specific materials upon which the judgment is based. Often, the teacher provides both an explicit rationale for the selection of evidence in the collection and a reflective analysis to help the reader or evaluator of the collection make sense of the evidence.

Artifact protocols can be thought of as a particular type of instructional collection that is much narrower. The protocols most widely researched are designed to measure the intellectual rigor and quality of the assignments teachers give students as well as the student work that is produced in response to those assignments (e.g., Borko, Stecher, & Kuffner, 2007; Newmann, Bryk, & Nagaoka, 2001). These protocols are designed to be independent of the academic difficulty of a particular course of study. For example, an advanced physics assignment would receive low scores if students were simply asked to provide definitions. The judgments made focus on an important but limited part of the teaching quality domain, focusing almost exclusively on teacher and

student practices, with much less teacher description and analysis called for than with other instructional collections. The protocols circumscribe what types of assignments are assessed, often asking for a mix of four to six typical and challenging assignments that produce written student work. Often, researchers sample assignments across the school year and allow for some teacher choice in which assignment is assessed.

Both artifact and instructional collection instruments have been used for various purposes, ranging from formative feedback for the improvement of teaching practice to licensure and high-stakes advanced certification decisions. For example, the Scoop Notebook is an instructional collection protocol that has been used to improve professional practice (Borko et al., 2007; Borko, Stecher, Alonzo, Moncure, & McClam, 2005). The portfolio protocol for NBPTS certification is used as part of a voluntary high-stakes assessment for advanced certification status (e.g., Cantrell, Fullerton, Kane, & Staiger, 2008; National Research Council, 2008; Szpara & Wylie, 2007). Related protocols have been used as part of licensure (e.g., California Commission on Teacher Credentialing, 2008; Connecticut State Department of Education, Bureau of Program and Teacher Evaluation, 2001), and three artifact protocols documented in the research literature have been used for the improvement of practice, judgments about school quality, and the evaluation of school reform models (Junker et al., 2006; Koh & Luke, 2009; Matsumura & Pascal, 2003; Mitchell et al., 2005; Newmann et al., 2001). These protocols vary in the degree to which they require the teacher to submit evidence that is naturalistic (i.e., already exists as a regular part of teaching practice) or evidence that is created specifically for inclusion in the assessment (e.g., written reflections or interviews).

All of the protocols reviewed have been developed to reflect a community's view of quality teaching. In the high-stakes assessments (e.g., the now-redesigned Connecticut's Beginning Educator Support and Training [BEST] portfolio assessment[9] and the NBPTS observation protocol), stakeholder committees were consulted extensively. As a class of protocols, there has been significant attention to raters and score quality. Although there have been graduate students, principals, and other education professionals trained to rate, raters have overwhelmingly been teachers with experience at the grade level and subject area being assessed (Aschbacher, 1999; Boston & Wolf, 2006; Matsumura et al., 2006; Matsumura, Garnier, Slater, & Boston, 2008; Newmann et al., 2001). Training on both instructional collection and artifact protocols is usually intensive (e.g., 3 to 5 days for artifacts and sometimes longer for instructional collections) and makes use of benchmark and training papers. For almost all protocols, raters are required to pass a certification test before scoring. Although the quality of the training as judged by interrater agreement varies across protocols and studies, the literature suggests it is possible to train raters to acceptable levels of agreement (more than 70%) with significant effort (Borko et al., 2007; Gitomer, 2008b; Ingvarson & Hattie, 2008; Matsumura et al., 2006; M. Wilson, Hallam, Pecheone, & Moss, 2006).

As with observations, score accuracy is often a challenge to the validity of interpretations of evidence for instructional collections. Accuracy problems, most often in the form of rater drift and bias, have been addressed by putting in place procedures for bias training (e.g., Ingvarson & Hattie, 2008) and retraining raters, rescoring, and, in some cases, modeling rater severity using Rasch models (Gitomer, 2008b; Kellor, 2002; National Research Council, 2008; Shkolnik et al., 2007; Wenzel, Nagaoka, Morris, Billings, & Fendt, 2002). Because there is such a wide range of practices to account for rater agreement across the instruments and purposes of those instruments, it is difficult to generalize about the quality of scores except to say it is uneven.

Instructional collections and artifact protocols examine evidence that is often produced as a regular part of teaching and learning. Perhaps in part because of this closeness to practice, instructional collections have high levels of face validity, and for at least some protocols, teachers report that

[9]BEST has been redesigned and, as of the 2009–2010 school year, is now known as the Teacher Education and Mentoring (TEAM) Program. This paper considers BEST as it existed before the redesign.

preparing an instructional collection improves their practice (e.g., Moss et al., 2004; Tucker, Stronge, Gareis, & Beers, 2003). Across protocols, however, teachers often feel they are burdensome.

Evidence is modest and mixed on the relationship to teaching practice and student achievement, depending on the instrument under investigation (e.g., National Research Council, 2008; M. Wilson et al., 2006). Instruments that focus on evaluating the products of classroom interactions rather than the teacher's commentary on those products in collections seem to have stronger evidence for a relationship to student learning (e.g., Cantrell et al., 2008; Matsumura et al., 2006). Consistent with this trend, there is a somewhat stronger, more moderate relationship between scores on artifact protocols and student achievement (Matsumura & Pascal, 2003; Matsumura et al., 2006; Mitchell et al., 2005; Newmann et al., 2001). This relationship may be due to the fact that artifact protocols are, by definition, more narrowly connected to teaching practice. If these instruments are to become more widely used in teacher evaluation, there will need to be a stronger understanding of teacher choice in selecting assignments and teacher-provided description and reflection. There will also have to be a stronger understanding of the role of grade-level, school, and district curricular decisions that could prove thorny when attributing scores to individual teachers.

Validity challenges to the measurement of teacher practices. Across these different measures of teacher practice, valid inferences about teaching quality will depend, in large part, on the ability to address the following issues. First, claims about teacher effectiveness must take into account contextual factors that individuals do not control. For example, as teachers are required to focus on test preparation activities, an increasingly common practice in recent years (Stecher, Vernez, & Steinberg, 2010), qualities of instruction valued by particular protocols may become less visible. Teachers who work within certain curricula may be judged to be more effective, not necessarily because of their own abilities, but because they are working with a curriculum that supports practices valued by particular measurement instruments (e.g., Cohen, 2010). Causal claims based on any single instrument may be inappropriate and can be better justified by considering evidence from multiple measures.

Second, issues of bias and fairness need to be examined and addressed. As with other assessment measures, there must be vigilance to ensure that measures do not, for construct-irrelevant reasons, privilege teachers with particular backgrounds. Aside from the NBPTS and Connecticut's previous BEST instructional collection research, there is very little research to suggest the field understands the bias and fairness implications of specific protocols. This is understandable given the more formative uses of many of the instruments; however, as stakes are attached, this will not be an acceptable state of affairs for either legal or ethical reasons.

Finally, implementation of these protocols is critical to the validity of the instrument for specific uses. Even if there is validity evidence for a particular measure, such evidence is dependent on implementing the protocols in particular ways, for example, with well-trained and calibrated raters. Using a protocol that has yielded valid inferences in one context with a specific set of processes in place does not guarantee that inferences made in a similar context with different implementation processes will yield those valid inferences. States and districts will have to monitor implementation specifics closely, given the budgetary and human capital constraints under which they will operate.

Teacher Beliefs

This category represents a mix of various kinds of measures that have been used to target different constructs about teaching. They include measures that range from personal characteristics and teacher beliefs to abilities to make judgments on others' teaching, typically through some type of survey or questionnaire methodology. Collectively, this body of work has tried to identify proxy measures of beliefs, attitudes, and understandings that could predict who would become a good teacher and that could provide guidance for individuals and systems as to whether individuals were suited to the profession of teaching, generally, and to particular teaching specialties and job placements, more specifically.

Fifty years ago, Getzels and Jackson (1963) reviewed the extant literature linking personality characteristics to teaching quality. Finding relationships somewhat elusive, they highlighted three substantial obstacles that remain relevant in the 21st century. First, they raised the problem of defining personality. Although personality theory has certainly evolved substantially over the past half-century, the identification of personality characteristics that are theoretically and empirically important to teaching is still underspecified. Second, they argued that instrumentation and theory to measure personality was relatively weak. The reliance on correlations of measures without strong theories that link personality constructs to practice continues to persist (e.g., Fang, 1996). The third fundamental challenge is the limitation of the criterion measures—what are the measures of teacher quality that personality measures are referenced against? Typical criterion measures that Getzels and Jackson reviewed included principal ratings, teacher self-reports, and experience. As can be seen throughout this review, although great effort has been and is being made in defining quality of teaching, the issues are hardly resolved.

Reviewing a large body of research, their conclusions were humbling:

> The regrettable fact is that many of the studies have not produced significant results. Many others have produced only pedestrian findings. For example, it is said after the usual inventory tabulation that good teachers are friendly, cheerful, sympathetic, and morally virtuous rather than cruel, depressed, unsympathetic, and morally depraved. But when this has been said, not very much that is especially useful has been revealed. . . . What is needed is not research leading to the self-evident but to the discovery of specific and distinctive features of teacher personality and of the effective teacher. (Getzels & Jackson, 1963, p. 574)

In the ensuing years, efforts have been undertaken to make progress beyond this earlier state of affairs. A large body of work has focused on *teacher efficacy*—that is, the extent to which an individual believes that teachers in general can determine student outcomes. This work highlights the continuing challenges in clarifying the personality constructs of interest. Ashton and Webb (1986), Gibson and Dembo (1984), and Woolfolk and Hoy (1990) all made the distinction between beliefs about what teachers in general can do to affect student outcomes (teacher efficacy) and what he or she as an individual could do to affect student outcomes (personal efficacy). Guskey and Passaro (1994) rejected this distinction as an artifact of instrument design and instead argued that two factors of efficacy—internal and external locus of control—reflect the extent to which teachers view themselves as having the ability to influence student learning. This work builds on the finding of Armor et al. (1976), who did find a modest relationship between student achievement gains and a composite measure of teacher beliefs based on the following statements:

1. When it comes right down to it, a teacher really can't do much because most of a student's motivation and performance depends on his or her home environment.
2. If I try really hard, I can get through to even the most difficult or unmotivated student. (Armor et al., 1976, p. 73)

Students who showed the greatest gains had teachers who disagreed with the first statement and agreed with the second.

The field continues to be characterized by, at best, modest correlations between measures of personality, dispositions, and beliefs and academic outcome measures. This, however, has not stopped the search for such measures. Metzger and Wu (2008) reviewed the available evidence for a widely used commercially available product, Gallup's Teacher Perceiver Interview. They attributed the modest findings to possibilities that teachers' responses in these self-report instruments may not be accurate reflections of their operating belief systems and that the manifestation of characteristics may be far more context bound than general instruments acknowledge. They concluded, as others have, that the constructs being examined are "slippery" (Metzger & Wu, 2008, p. 934) and multifaceted, making it very difficult to detect relationships. The validity argument for this group of measures is weak.

Student Beliefs and Student Practices

Both teachers and students contribute to teaching quality. The measures used to assess teaching quality through the assessment of student beliefs and practices may be considered. As Figure 20.1 indicates, some of the instruments being used to assess teacher beliefs and practices also assess student beliefs and practices. For example, on the holistic observation protocol called the Classroom Assessment Scoring System (CLASS), developed by Pianta, Hamre, Haynes, Mintz, and La Paro (2007), raters are trained to observe both teacher practices and student practices. Secondary classrooms that, for example, receive high scores on the quality of feedback dimension of CLASS would have students engaging in back-and-forth exchanges with the teacher, demonstrating persistence, and explaining their thinking in addition to all of the teacher's actions specified in that dimension. This focus on both teacher and student practices is common across the observation protocols reviewed in this section.

Many instruments are designed to measure student beliefs and practices on a wide range of topics from intelligence to self-efficacy to critical thinking (e.g., Dweck, 2002; Stein, Haynes, Redding, Ennis, & Cecil, 2007; Usher & Pajares, 2009). A summary of this research is outside the scope of this chapter, but only one identified belief instrument is being used to evaluate teachers. On the basis of a decade of work by Ferguson and his colleagues in the Tripod Project (Ferguson, 2007), the MET project is using the Tripod assessment to determine the degree to which students' perceptions on seven topics are predictive of other aspects of teaching quality (Bill and Melinda Gates Foundation, 2011b). Preliminary evidence suggests the assessment is internally reliable (coefficient alpha > .80) when administered in such a way that there are no stakes for students and teachers (i.e., a research setting). Results on how the instrument functions in situations in which there are consequences for teachers have not yet been published.

Student Knowledge

Value-added models. Over recent years, there has been great enthusiasm for incorporating measures of student achievement into estimates of how well teachers are performing. This approach has led policymakers and researchers to advocate for the use of value-added measures to evaluate individual teachers. Value-added measures use complex analytic methods applied to longitudinal student achievement data to estimate teacher effects that are separate from other factors shaping student learning. Comprehensive, nontechnical treatments of value-added approaches are presented by Braun (2005b) and the National Research Council and National Academy of Education (2010).

The attraction of valued-added methods to many is that they are objective measures that avoid the complexities associated with human judgment. They are also relatively low cost once data systems are in place, and they do not require the human capital and ongoing attention required by many of the previously described measures. Finally, policymakers are attracted to the idea of applying a uniform metric to all teachers, provided test scores are available. Although these models are promising, they have important methodological and political limitations that represent challenges to the validity of inferences based on VAM (Braun, 2005b; Clotfelter et al., 2005; Gitomer, 2008a; Kupermintz, 2003; Ladd, 2007; Lockwood et al., 2007; National Research Council and National Academy of Education, 2010; Raudenbush, 2004; Reardon & Raudenbush, 2009).

These challenges are actually not unique to VAM. However, because VAM has been so widely endorsed in policy circles and because it is viewed as having an objective credibility that other measures do not, it is particularly important to highlight these challenges with respect to VAM.

These challenges are summarized into two broad and related categories. A first validity challenge concerns the nature of the construct. One distinguishes between teacher and teaching effectiveness because a variety of factors may influence the association of scores with an individual teacher. For example, school resources, particularly those targeted at instruction (e.g., Cohen, Raudenbush, & Ball, 2003); specific curricula (e.g., Cohen, 2010; Tyack & Cuban, 1995); and district policies that provide financial, technical, and professional support to achieve instructional goals (e.g., Ladd, 2007), all

can influence what gets taught and how it gets taught, potentially influencing the student test scores that are used to produce VAM estimates and inferences about the teacher. There are other interpretive challenges as well: Other adults (both parents and teachers) may contribute to student test results, and the limits of student tests may inappropriately constrain the inference to the teacher (for a broad discussion of construct-relevant concerns, see E. L. Baker et al., 2010).

A second set of issues concerns the internal validity of VAM. One aspect of internal validity requires that VAM estimates are attributable to the experience of being in the classroom and not attributable to preexisting differences between students across different classrooms. Furthermore, internal validity requires that VAM estimates are not attributable to other potential modeling problems. Substantial treatment of these methodological issues associated with VAM is provided elsewhere (Harris & McCaffrey, 2010; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; National Research Council and National Academy of Education, 2010; Reardon & Raudenbush, 2009). Key challenges include the fact that students are not randomly assigned to teachers within and across schools. This makes it difficult to interpret whether VAM effects are attributable to teachers or the entering characteristics of students (e.g., Clotfelter et al., 2005; Rothstein, 2009). Model assumptions that attempt to adjust for this sorting have been shown to be problematic (National Research Council and National Academy of Education, 2010; Reardon & Raudenbush, 2009). Finally, choices about the content of the test (e.g., Lockwood et al., 2007), the scaling (e.g., Ballou, 2008; Briggs, 2008; Martineau, 2006), and the fundamental measurement error inherent in achievement tests and especially growth scores can "undermine the trustworthiness of the results of value-added methods" (Linn, 2008, p. 13).

Bringing together these two sets of validity concerns suggests that estimates of a particular teacher's effectiveness may vary substantially as a function of the policies and practices in place for a given teacher, the assignment of students to teachers, and the particular tests and measurement models used to calculate VAM. Substantial research into VAM continues to attempt to address these validity challenges and to understand the most appropriate use of VAM within evaluation systems. Researchers and policymakers vary in their confidence that these issues will be to the improvement of educational practice (for two distinct perspectives, see E. L. Baker et al., 2010; Glazerman et al., 2010).

Student learning objectives. Evaluation policies must include all teachers. If student achievement is to be a core component of these evaluation systems, policymakers must address the fact that there are no annual achievement test data appropriate to evaluate roughly 50% to 70% of teachers, either because of the subjects or grade levels that they teach. One of the solutions proposed has been the development of measures using student learning objectives (SLOs; Community Training and Assistance Center, 2008). In these models, teachers articulate a small set of objectives and appropriate assessments to demonstrate that students are learning important concepts and skills in their classrooms. SLOs are reviewed by school administrators for acceptability. Teachers are evaluated by considering how well the SLOs are achieved on the basis of assessment results. Because of the limited applicability of VAM, SLOs are being considered for use in numerous state teaching evaluation systems (e.g., Rhode Island, Maryland, and New York). Many of these models include the development of common SLOs for use across a state.

The integrity of the process rests on the quality of the objectives and the rigor with which they are produced and reviewed inside the educational system. To date, there is a very limited set of data to judge the validity of these efforts. Available studies have found, first, that developing high-quality objectives that identify important learning goals is challenging. The Community Training and Assistance Center (2004) reported that for the first 3 years of a 4-year study, a majority of teachers produced SLOs that needed improvement. Teachers failed to identify important and coherent learning goals and had low expectations for students. Studies do report, however, that teachers with stronger learning goals tend to have students who demonstrate better achievement (Community Training and Assistance Center, 2004; Lussier & Forgione, 2010). There are

also indications that across systems, SLOs can lead teachers to have stronger buy-in to the evaluation system than has been demonstrated with other evaluation approaches (Brodsky, DeCesare, & Kramer-Wine, 2010).

COMBINING MULTIPLE MEASURES

Policy discussions are now facing the challenge of integrating information from the various measures considered thus far as well as measures that are specific to particular assessment purposes. The *Standards for Educational and Psychological Testing* (AERA et al., 1999) provide guidance on the use of multiple measures in decisions about employment and credentialing:

> *Standard 14.13*—When decision makers integrate information from multiple tests or integrate test and nontest information, the role played by each test in the decision process should be clearly explicated, and the use of each test or test composite should be supported by validity evidence;
>
> *Standard 14.16*—Rules and procedures used to combine scores on multiple assessments to determine the overall outcome of a credentialing test should be reported to test takers, preferably before the test is administered.

Current policies and practices are only beginning to be developed. For example, the U.S. Department of Education's (2010) Race to the Top competition asks states to

> design and implement rigorous, transparent, and fair evaluation systems for teachers and principals that (a) differentiate effectiveness using multiple rating categories that take into account data on student growth (as defined in this notice) as a significant factor, and (b) are designed and developed with teacher and principal involvement. (p. 34)

How these multiple ratings are accounted for, however, is left unstated. As states and districts grapple with these issues, a number of fundamental measurement questions will need to be considered to address the *Standards for Educational and Psychological Testing* (AERA et al., 1999).

Compensatory or Conjunctive Decisions

One question concerns the nature of the decision embedded in the system. In a conjunctive system, individuals must satisfy a standard for each constituent measure, whereas in a compensatory system, individuals can do well on some measures and less well on others as long as a total score reaches some criterion. In a conjunctive model, the reliability of each individual measure ought to be sufficiently high such that decisions based on each individual measure are defensible. A compensatory model, such as that used by NBPTS, does not carry the same burden, but it does lead to situations in which someone can satisfy an overall requirement and perform quite poorly on constituent parts. One compromise that is sometimes taken is to adopt a compensatory model, yet set some minimum scores for particular measures.

Determining and Using Weighting Schemes

Some proposed systems (e.g., Bill and Melinda Gates Foundation, 2011a) are trying to establish a single metric of teacher effectiveness that is based on a composite of measures. Efforts like these attempt to determine the weighting of particular measures based on statistical models that will maximize the variance accounted for by particular measures.

At least two complexities will need to be kept in mind by whatever weighting scheme is used. First, if two measures have the same "true" relationship with a criterion variable, the one that is scored more reliably will be more predictive of the criterion and thus will be assigned a greater weight. Because of the reliability of scoring, some measures, or dimensions of measures, may be viewed as more predictive of the outcome than they actually are when compared with other less reliable measures.

A second source of potential complexity is that measures that have greater variability across individuals are likely to have a stronger impact on a total evaluation score and that the effective weighting will be far larger than the assigned weights would indicate. Imagine a system that derived a composite

teaching quality score based on value-added and principal evaluation scores and also imagine that each was assigned a weight of 50%. Now imagine that the principal did not differentiate teachers much, if at all. In this case, even though each measure was assigned a weight of 50%, the value-added measure actually contributes almost all the variance to the total score. Thus, it is important not to just assign an intended weight but also to understand the effective weight given the characteristics of the scores when implemented (e.g., range, variance, measurement error, etc.).

The exercise of judgment. Systems can range from those in which a single metric is derived from multiple measures via a mathematical model to ones in which decision makers are required to exercise a summative judgment that takes into account multiple measures. Systems that avoid judgment often do so because of a lack of trust in the judgment process. If judgment is valued, as it is in many high-performing education systems, then it will be imperative to ensure that judgments are executed in ways that are credible and transparent.

Rare yet important teaching characteristics. Finally, there may be characteristics that do not contribute to variance on valued outcomes that should contribute to composite measures. For example, we may believe that teachers should not make factual errors in content or be verbally abusive to students. These might be rare events and do little to help distinguish among teachers; however, robust evaluation systems might want to include them to make standards of professional conduct clear. Weighting schemes that rely solely on quantitative measurable outcomes run the risk of ignoring these important characteristics.

CONCLUSION

An ambitious policy agenda that includes teacher evaluation as one of its cornerstones places an unprecedented obligation on the field of education measurement to design, develop, and validate measures of teaching quality. There is a pressing need for evaluation systems that can support the full range of purposes for which they are being considered—from employment and compensation decisions to professional development. Doing this responsibly obligates the field to uphold the fundamental principles and standards of education measurement in the face of enormous policy pressures. Well-intentioned policies will be successful only if they are supported by sound measurement practice.

Building well-designed measures of effective teaching will require coordinated developments in theory, design, and implementation, along with ongoing monitoring processes. Through ongoing validation efforts, enhancements to each of these critical components should be realized. This process also will require implementation of imperfect systems that can be subject to continued examination and refinement. The discipline needs to continue to examine the fairness and validity of interpretations and develop processes that ensure high-quality and consistent implementation of whichever measures are being employed. Such quality control can range from ensuring quality judgments from individuals rating teacher performance to ensuring that adequate data quality controls are in place for value-added modeling.

It is important that sound measurement practices be developed and deployed for these emerging evaluation systems, but there may be an additional benefit to careful measurement work in this area. Theories of teaching quality continue to be underdeveloped. Sound measures can contribute to both the testing of theory and the evolution of theories about teaching. For example, as educators understand more about how contextual factors influence teaching quality, theories of teaching will evolve. Understanding the relationship between context and teaching quality also may lead to the evolution and improvement of school and district decisions that shape student learning.

For the majority of instruments reviewed in this chapter, their design can be considered first generation. Whether measures of teacher knowledge, instructional collections, or observation methods, there is a great deal to be done in terms of design of protocols, design of assessments and items, training and calibration of raters, aggregation of scores, and psychometric modeling. Even the understanding of expected psychometric performance on each class of

measures is at a preliminary stage. Importantly, most of the work on these measures done to date has been conducted in the context of research studies. There is little empirical understanding of how these measures will work in practice, with all the unintended consequences, incentives, disincentives, and competing priorities that characterize education policy.

There is at least one crucial aspect of the current policy conversation that may prove to be the Achilles' heel of the new systems being developed, should it go unchecked. In general, all of the currently envisioned systems layer additional tasks, costs, and data management burdens on school, district, and state resources. Observations take principals' time. SLOs take teachers', principals', and districts' time. Student questionnaires take students' time. Data systems that track all of these new measures require money and time. And the list goes on. These systems are massive because they are intended to apply to all teachers in every system. Serious consideration has not been given to how institutions can juggle existing resource demands with these new demands. The resource pressures these evaluation systems place on institutions may result in efficiencies, but they may also result in significant pressure to cut measurement corners that could pose threats to the validity of the systems. Such unintended consequences must be monitored carefully.

Although the new evaluation systems will require substantial resources, the justification for moving beyond measures that simply assign a ranking is that these kinds of measures can provide helpful information to stakeholders about both high-quality teaching and the strengths and weaknesses of teachers and school organizations in providing students access to that teaching. Actualizing such a useful system will require commitments by researchers, policymakers, and practitioners alike to proceed in ways that support valid inferences about teaching quality.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Federation of Teachers. (2010). *A continuous improvement model for teacher development and evaluation* (Working paper). Washington, DC: Author.

Armor, D., Conroy-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., & Zellman, G. (1976). *Analysis of the school preferred reading programs in selected Los Angeles minority schools* (Report No. R-2007-LAUSD). Santa Monica, CA: RAND Corporation.

Aschbacher, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform* (CSE Technical Report No. 513). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

Ashton, P. T., & Webb, R. B. (1986). *Making a difference: Teacher efficacy and student achievement* (Monograph). White Plains, NY: Longman.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing Paper No. 278). Washington, DC: Economic Policy Institute.

Baker, S. K., Gersten, R., Haager, D., & Dingle, M. (2006). Teaching practice and the reading growth of first-grade English learners: Validation of an observation instrument. *The Elementary School Journal*, *107*, 199–220. doi:10.1086/510655

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*, 389–407. doi:10.1177/0022487108324554

Ballou, D. (2008, October). *Value-added analysis: Issues in the economics literature*. Paper presented at the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC. Retrieved from http://www7.nationalacademies.org/bota/VAM%20Analysis%20-%20Ballou.pdf

Barnette, J. J., & Gorham, K. (1999). Evaluation of teacher preparation graduates by NCATE accredited institutions: Techniques used and barriers. *Research in the Schools*, *6*(2), 55–62.

Bell, C. A., Gitomer, D. H., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 62–87.

Bell, C. A., Little, O. M., & Croft, A. J. (2009, April). *Measuring teaching practice: A conceptual review*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Bell, C. A., & Youngs, P. (2011). Substance and show: Understanding responses to NCATE accreditation. *Teaching and Teacher Education*, *27*, 298–307. doi:10.1016/j.tate.2010.08.012

Bill and Melinda Gates Foundation. (2011a). *Learning about teaching: Initial findings from the Measures of Effective Teaching project*. Washington, DC: Author. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

Bill and Melinda Gates Foundation. (2011b). *Intensive partnerships for effective teaching*. Washington, DC: Author. Retrieved from http://www.gatesfoundation.org/college-ready-education/Pages/intensive-partnerships-for-effective-teaching.aspx

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*(8), 3–15. doi:10.3102/0013189X033008003

Borko, H., Stecher, B., & Kuffner, K. (2007). *Using artifacts to characterize reform-oriented instruction: The Scoop Notebook and rating guide* (CSE Technical Report No. 707). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

Borko, H., Stecher, B. M., Alonzo, A. C., Moncure, S., & McClam, S. (2005). Artifact packages for characterizing classroom practice: A pilot study. *Educational Assessment*, *10*, 73–104. doi:10.1207/s15326977ea1002_1

Boston, M., & Wolf, M. K. (2006). *Assessing academic rigor in mathematics instruction: The development of the Instructional Quality Assessment Toolkit* (CSE Technical Report No. 672). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

Braun, H. I. (2005a). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Braun, H. I. (2005b). Value-added modeling: What does due diligence require? In R. Lissitz (Ed.), *Value-added models in education: Theory and applications* (pp. 19–39). Maple Grove, MN: JAM Press.

Briggs, D. (2008, November). *The goals and uses of value-added models*. Paper presented at the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC. Retrieved from http://www7.nationalacademies.org/bota/VAM%20Goals%20and%20Uses%20paper%20-%20Briggs.pdf

Brodsky, A., DeCesare, D., & Kramer-Wine, J. (2010). Design and implementation considerations for alternative teacher compensation systems. *Theory Into Practice, 49*, 213–222. doi:10.1080/00405841.2010.487757

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 328–375). New York, NY: Macmillan.

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics*, *66*, 103–115. doi:10.1016/j.jue.2009.05.001

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*, 423–429. doi:10.1016/0895-4356(93)90018-V

California Commission on Teacher Credentialing. (2008). *California teaching performance assessment*. Sacramento, CA: Author.

Cantrell, S., Fullerton, J., Kane, T. J., & Staiger, D. O. (2008). *National board certification and teacher effectiveness: Evidence from a random assignment experiment* (NBER Working Paper No. 14608). Cambridge, MA: National Bureau of Economic Research.

Carnegie Forum on Education and the Economy. (1986). *A nation prepared: Teachers for the 21st century*. New York, NY: Carnegie.

Clotfelter, C., Ladd, H., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, *24*, 377–392. doi:10.1016/j.econedurev.2004.06.008

Cochran-Smith, M., & Zeichner, K. M. (Eds.). (2005). *Studying teacher education: The report of the AERA panel on research and teacher education*. Mahwah, NJ: Erlbaum.

Cohen, D. (2010). Teacher quality: An American educational dilemma. In M. M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality: A handbook* (pp. 375–402). San Francisco, CA: Jossey-Bass.

Cohen, D., Raudenbush, S., & Ball, D. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*, 119–142. doi:10.3102/01623737025002119

Community Training and Assistance Center. (2004). *Catalyst for change: Pay for performance in Denver*. Boston, MA: Author.

Community Training and Assistance Center. (2008). *Tying earning to learning: The link between teacher compensation and student learning objectives*. Boston, MA: Author.

Connecticut State Department of Education, Bureau of Program and Teacher Evaluation. (2001). *A guide to the BEST program for beginning teachers*. Hartford, CT: Author.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.

Desimone, L., Porter, A. C., Garet, M., Suk Yoon, K., & Birman, B. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, *24*, 81–112. doi:10.3102/01623737024002081

Dweck, C. S. (2002). The development of ability conceptions. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 57–88). San Diego, CA: Academic Press. doi:10.1016/B978-012750053-9/50005-X

Educational Testing Service. (2001). *PRAXIS III: Classroom performance assessments orientation guide*. Princeton, NJ: Author.

Fang, Z. (1996). A review of research on teacher beliefs and practices. *Educational Research*, *38*, 47–65. doi:10.1080/0013188960380104

Ferguson, R. F. (2007). *Toward excellence with equity: An emerging vision for closing the achievement gap*. Boston, MA: Harvard Education Press.

Gersten, R., Baker, S. K., Haager, D., & Graves, A. W. (2005). Exploring the role of teacher quality in predicting reading outcomes for first-grade English learners. *Remedial and Special Education*, *26*, 197–206. doi:10.1177/07419325050260040201

Getzels, J. W., & Jackson, P. W. (1963). The teacher's personality and characteristics. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 506–582). Chicago, IL: Rand McNally.

Gibson, S., & Dembo, M. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology, 76*, 569–582. doi:10.1037/0022-0663.76.4.569

Gitomer, D. H. (2007). *Teacher quality in a changing policy landscape: Improvements in the teacher pool* (ETS Policy Information Report No. PIC-TQ). Princeton, NJ: Educational Testing Service.

Gitomer, D. H. (2008a). Crisp measurement and messy context: A clash of assumptions and metaphors—Synthesis of Section III. In D. H. Gitomer (Ed.), *Measurement issues and the assessment for teacher quality* (pp. 223–233). Thousand Oaks, CA: Sage.

Gitomer, D. H. (2008b). Reliability and NBPTS assessments. In L. Ingvarson & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards* (pp. 231–253). Greenwich, CT: JAI Press. doi:10.1016/S1474-7863(07)11009-7

Gitomer, D. H., & Bell, C. A. (2012, August). *The instructional challenge in improving instruction: Lessons from a classroom observation protocol*. Paper presented at the European Association for Research on Learning and Instruction Sig 18 Conference, Zurich, Switzerland.

Gitomer, D. H., Latham, A. S., & Ziomek, R. (1999). *The academic quality of prospective teachers: The impact of admissions and licensure testing* (ETS Teaching and Learning Report Series No. ETS RR-03-25). Princeton, NJ: Educational Testing Service.

Gitomer, D. H., & Phelps, G. C. (2012). *Measures of effective teaching: Assessment of content knowledge for teaching*. Unpublished manuscript.

Gitomer, D. H., & Qi, Y. (2010). *Score trends for Praxis II*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: Brown Center on Education Policy at Brookings.

Goe, L. (2007). *The link between teacher quality and student outcomes*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D., & Hansen, M. (2010). Race, gender, and teacher testing: How informative a tool is teacher licensure testing? *American Educational Research Journal, 47*, 218–251. doi:10.3102/0002831209348970

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context* (NCES 2009-001 Revised). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://nces.ed.gov/pubs2009/2009001.pdf

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (Hamilton Project Discussion Paper). Washington, DC: The Brookings Institution.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010, May). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores* (NBER Working Paper No. 16015). Cambridge, MA: National Bureau of Economic Research.

Gullickson, A. R. (2008). *The personnel evaluation standards: How to assess systems for evaluating educators*. Thousand Oaks, CA: Corwin Press.

Guskey, T. R., & Passaro, P. D. (1994). Teacher efficacy: A study of construct dimensions. *American Educational Research Journal*, *31*, 627–643.

Haertel, E. H. (1991). New forms of teacher assessment. *Review of Research in Education, 17*, 3–29.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S., . . . Hakigami, A. (in press). Teaching through interactions: Testing a developmental framework of teacher effectiveness

in over 4,000 classrooms. *The Elementary School Journal*.

Haney, W. M., Madaus, G., & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of education. *Review of Research in Education, 14*, 169–238.

Hanushek, E. A. (2002). Teacher quality. In L. T. Izumi & W. M. Evers (Eds.), *Teacher quality* (pp. 1–13). Stanford, CA: Hoover Institution Press.

Harris, D., & McCaffrey, D. (2010). Value-added: Assessing teachers' contributions to student achievement. In M. M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality: A handbook* (pp. 251–282). San Francisco, CA: Jossey-Bass.

Harris, D. N., & Sass, T. R. (2006). *Value-added models and the measurement of teacher quality*. Unpublished manuscript.

Heneman, H. G., Milanowski, A., Kimball, S. M., & Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay* (CPRE Policy Briefs No. RB-45). Philadelphia: Consortium for Policy Research in Education, University of Pennsylvania.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430–511. doi:10.1080/07370000802177235

Hill, H. C., Dean, C., & Goffney, I. M. (2007). Assessing elemental and structural validity: Data from teachers, non-teachers, and mathematicians. *Measurement: Interdisciplinary Research and Perspectives*, *5*(2–3), 81–92. doi:10.1080/15366360701486999

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*, 371–406. doi:10.3102/00028312042002371

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal, 105*, 11–30. doi:10.1086/428763

Hill, H. C., Umland, K. L., & Kapitula, L. R. (2011). A validity argument approach to evaluating value-added scores. *American Educational Research Journal*, *48*, 794–831. doi:10.3102/0002831210387916

Hines, L. M. (2007). Return of the thought police? The history of teacher attitude adjustment. *Education Next*, *7*(2), 58–65. Retrieved from http://educationnext.org/files/ednext_20072_58.pdf

Hirsch, E., & Sioberg, A. (2010). *Using teacher working conditions survey data in the North Carolina educator evaluation process*. Santa Cruz, CA: New Teacher Center. Retrieved from http://ncteachingconditions.org/sites/default/files/attachments/NC10_brief_TeacherEvalGuide.pdf

Honig, M. I., Copland, M. A., Rainey, L., Lorton, J. A., & Newton, M. (2010, April). *School district central office transformation for teaching and learning improvement: A report to the Wallace Foundation*. Seattle, WA: The Center for the Study of Teaching and Policy.

Horizon Research. (2000). *Inside classroom observation and analytic protocol*. Chapel Hill, NC: Author.

Howard, B. B., & Gullickson, A. R. (2010). Setting standards in teacher evaluation. In M. M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality: A handbook* (pp. 337–354). San Francisco, CA: Jossey-Bass.

Ingvarson, L., & Hattie, J. (Eds.). (2008). *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards*. Greenwich, CT: JAI Press.

Jaeger, R. M. (1999). *Some psychometric criteria for judging the quality of teacher certification tests*. Paper commissioned by the Committee on Assessment and Teacher Quality. Greensboro: University of North Carolina.

Johnson, S. M., Birkeland, S. E., Kardos, S. K., Kauffman, D., Liu, E., & Peske, H. G. (2001, September/October). Retaining the next generation of teachers: The importance of school-based support. *Harvard Education Letter*. Retrieved from http://www.umd.umich.edu/casl/natsci/faculty/zitzewitz/curie/TeacherPrep/99.pdf

Johnson, S. M., Kardos, S. K., Kauffman, D., Liu, E., & Donaldson, M. L. (2004). The support gap: New teachers' early experiences in high-income and low-income schools. *Education Policy Analysis Archives*, *12*(61). Retrieved from http://epaa.asu.edu/ojs/article/viewFile/216/342

Junker, B., Weisberg, Y., Matsumura, L. C., Crosson, A., Wolf, M. K., Levison, A., & Resnick, L. (2006). *Overview of the Instructional Quality Assessment* (CSE Technical Report No. 671). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125–160.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). New York, NY: Praeger.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City*. New York, NY: National Bureau of Economic Research.

Kardos, S. K., & Johnson, S. M. (2007). On their own and presumed expert: New teachers' experiences with their colleagues. *Teachers College Record*, *109*, 2083–2106.

Kellor, E. M. (2002). *Performance-based licensure in Connecticut* (CPRE-UW Working Paper Series TC-02–10). Madison, WI: Consortium for Policy Research in Education.

Kennedy, M. M. (2010). Approaches to annual performance assessment. In M. M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality: A handbook* (pp. 225–250). San Francisco, CA: Jossey-Bass.

Klein, S. P., & Stecher, B. (1991). Developing a prototype licensing examination for secondary school teachers. *Journal of Personnel Evaluation in Education, 5*, 169–190. doi:10.1007/BF00117336

Koh, K., & Luke, A. (2009). Authentic and conventional assessment in Singapore schools: An empirical study of teacher assignments and student work. *Assessment in Education: Principles, Policy, and Practice*, *16*, 291–318.

Koretz, D. (2008). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Kornfeld, J., Grady, K., Marker, P. M., & Ruddell, M. R. (2007). Caught in the current: A self-study of state-mandated compliance in a teacher education program. *Teachers College Record*, *109*, 1902–1930.

Krauss, S., Baumert, J., & Blum, W. (2008). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV Constructs. *ZDM—The International Journal on Mathematics Education, 40*, 873–892.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, *25*, 287–298. doi:10.3102/01623737025003287

Kyriakides, L., & Creemers, B. P. M. (2008). A longitudinal study on the stability over time of school and teacher effects on student outcomes. *Oxford Review of Education*, *34*, 521–545. doi:10.1080/03054980701782064

Ladd, H. F. (2007, November). *Holding schools accountable revisited.* Paper presented at APPAM Fall Research Conference, Washington, DC.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal*, *104*, 409–426. doi:10.1086/499760

Linn, R. L. (2008, November 13–14). *Measurement issues associated with value-added models.* Paper presented at the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC. Retrieved from http://www7.nationalacademies.org/bota/VAM_Robert_Linn_Paper.pdf

Livingston, S., & Zieky, M. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*, 47–67. doi:10.1111/j.1745-3984.2007.00026.x

Lussier, D. F., & Forgione, P. D. (2010). Supporting and rewarding accomplished teaching: Insights from Austin, Texas. *Theory Into Practice*, *49*, 233–242. doi:10.1080/00405841.2010.487771

Luyten, H. (2003). The size of school effects compared to teacher effects: An overview of the research literature. *School Effectiveness and School Improvement*, *14*, 31–51. doi:10.1076/sesi.14.1.31.13865

Malmberg, L. E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology*, *102*, 916–932. doi:10.1037/a0020920

Martineau, J. A. (2006). Distorting value-added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*, 35–62. doi:10.3102/10769986031001035

Matsumura, L. C., Garnier, H., Slater, S. C., & Boston, M. (2008). Toward measuring instructional interactions at-scale. *Educational Assessment, 13*, 267–300. doi:10.1080/10627190802602541

Matsumura, L. C., & Pascal, J. (2003). *Teachers' assignments and student work: Opening a window on classroom practice* (CSE Report No. 602). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., & Resnick, L. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the Instructional Quality Assessment* (CSE Technical Report No. 681). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

McCaffrey, D. F. (2011, April). *Sources of variance and mode effects in measures of teaching in algebra classes.* Paper presented at the annual meeting of the National Council on Measurement, New Orleans, LA.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council of Education and Macmillan.

Metzger, S. A., & Wu, M. J. (2008). Commercial teacher selection instruments: The validity of selecting teachers through beliefs, attitudes, and values. *Review of Educational Research*, *78*, 921–940. doi:10.3102/0034654308323035

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33–53. doi:10.1207/s15327930pje7904_3

Mitchell, K. J., Robinson, D. Z., Plake, B. S., & Knowles, K. T. (Eds.). (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality*. Washington, DC: National Academies Press.

Mitchell, K., Shkolnik, J., Song, M., Uekawa, K., Murphy, R., & Means, B. (2005). *Rigor, relevance, and results: The quality of teacher assignments and student work in new and conventional high schools*. Washington, DC: American Institutes for Research and SRI.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A., & Ehrle, K. (1999). Evaluating the SAGE program: A pilot program in targeted pupil-teacher reduction in Wisconsin. *Educational Evaluation and Policy Analysis*, *21*, 165–177.

Moss, P. A., & Schutz, A. (1999). Risking frankness in educational assessment. *Phi Delta Kappan*, *80*, 680–687.

Moss, P. A., Sutherland, L. M., Haniford, L., Miller, R., Johnson, D., Geist, P. K., . . . Pecheone, R. L. (2004). Interrogating the generalizability of portfolio assessments of beginning teachers: A qualitative study. *Education Policy Analysis Archives*, *12*(32), 1–70.

National Board for Professional Teaching Standards. (2010). *The standards*. Retrieved from http://nbpts.org/the_standards

National Research Council. (2008). *Assessing accomplished teaching: Advanced-level certification programs*. Washington, DC: National Academies Press.

National Research Council and National Academy of Education. (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: National Academies Press.

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Chicago, IL: Consortium on Chicago School Research.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115 Stat 1425 (2002).

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237–257. doi:10.3102/01623737026003237

Oakes, J. (1987). *Tracking in secondary schools: A contextual perspective*. Santa Monica, CA: RAND.

Odden, A., & Kelley, C. (2002). *Paying teachers for what they know and can do: New and smarter compensation strategies to improve student learning*. Thousand Oaks, CA: Corwin Press.

Ohio Department of Education. (2006). *Report on the quality of teacher education in Ohio: 2004–2005*. Columbus, OH: Author.

Pacheco, A. (2008). Mapping the terrain of teacher quality. In D. H. Gitomer (Ed.), *Measurement issues and assessment for teacher quality* (pp. 160–178). Thousand Oaks, CA: Sage.

Pearlman, M. (2008). The design architecture of NBPTS certification assessments. In L. Ingvarson & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards* (pp. 55–91). Greenwich, CT: JAI Press. doi:10.1016/S1474-7863(07)11003-6

Phelps, G. (2009). Just knowing how to read isn't enough! What teachers know about the content of reading. *Educational Assessment, Evaluation, and Accountability*, *21*, 137–154. doi:10.1007/s11092-009-9070-6

Phelps, G., & Schilling, S. (2004). Developing measures of content knowledge for teaching reading. *The Elementary School Journal*, *105*, 31–48. doi:10.1086/428764

Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S. L., & La Paro, K. M. (2007). *Classroom assessment scoring system manual, middle/secondary version*. Charlottesville: University of Virginia.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2007). *Classroom assessment scoring system*. Baltimore, MD: Brookes.

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 23*, 431–451. doi:10.1016/j.ecresq.2008.02.001

Piburn, M., & Sawada, D. (2000). *Reformed Teaching Observation Protocol (RTOP) reference manual*. Tempe: Arizona State University.

Pilley, J. G. (1941). The National Teacher Examination Service. *School Review*, *49*, 177–186. doi:10.1086/440636

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5*, 285–301. doi:10.1207/s15324818ame0504_1

Programme for International Student Assessment. (2006). *PISA 2006 science competencies for tomorrow's world*.

Retrieved from http://www.oei.es/evaluacioneducativa/InformePISA2006-FINALingles.pdf

Pullin, D. (1999). *Criteria for evaluating teacher tests: A legal perspective*. Washington, DC: National Academies Press.

Pullin, D. (2010). Judging teachers: The law of teacher dismissals. In M. M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality: A handbook* (pp. 297–333). San Francisco, CA: Jossey-Bass.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*, 121–129. doi:10.3102/10769986029001121

Raudenbush, S. W., Martinez, A., Bloom, H., Zhu, P., & Lin, F. (2010). *Studying the reliability of group-level measures with implications for statistical power: A six-step paradigm* (Working paper). Chicago, IL: University of Chicago.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*, 5–29. doi:10.3102/0162373707299460

Raver, C. C., Jones, S. M., Li-Grining, C. P., Metzger, M., Smallwood, K., & Sardin, L. (2008). Improving preschool classroom processes: Preliminary findings from a randomized trial implemented in Head Start settings. *Early Childhood Research Quarterly*, *23*, 10–26. doi:10.1016/j.ecresq.2007.09.001

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York, NY: Basic Books.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Educational Finance and Policy 4*, 492–519. doi:10.1162/edfp.2009.4.4.492

Rothstein, J. (2009). Student sorting and bias in value added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*, 537–571. doi:10.1162/edfp.2009.4.4.537

Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum in large-scale surveys: A study of literacy teaching in third-grade classrooms. *The Elementary School Journal, 105*, 75–101. doi:10.1086/428803

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from A Study of Instructional Improvement. *Educational Researcher, 38*(2), 120–131. doi:10.3102/0013189X09332375

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study of elementary schools. *Teachers College Record*, *104*, 1525–1567. doi:10.1111/1467-9620.00212

Rowan, B., Jacob, R., & Correnti, R. (2009). Using instructional logs to identify quality in educational settings. *New Directions for Youth Development, 121*, 13–31. doi:10.1002/yd.294

Samaras, A. P., Francis, S. L., Holt, Y. D., Jones, T. W., Martin, D. S., Thompson, J. L., & Tom, A. R. (1999). Lived experiences and reflections of joint NCATE-state reviews. *Teacher Educator*, *35*, 68–83. doi:10.1080/08878739909555218

Schilling, S. G., & Hill, H. C. (2007). Assessing measures of mathematical knowledge for teaching: A validity argument approach. *Measurement: Interdisciplinary Research and Perspectives*, *5*(2–3), 70–80. doi:10.1080/15366360701486965

Shkolnik, J., Song, M., Mitchell, K., Uekawa, K., Knudson, J., & Murphy, R. (2007). *Changes in rigor, relevance, and student learning in redesigned high schools*. Washington, DC: American Institutes for Research and SRI.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Stecher, B. M., Vernez, G., & Steinberg, P. (2010). *Reauthorizing No Child Left Behind: Facts and recommendations*. Santa Monica, CA: RAND Corporation.

Stein, B., Haynes, A., Redding, M., Ennis, T., & Cecil, M. (2007). Assessing critical thinking in STEM and beyond. In M. Iskander (Ed.), *Innovations in E-learning, instruction technology, assessment, and engineering education* (pp. 79–82). Dordrecht, the Netherlands: Springer. doi:10.1007/978-1-4020-6262-9_14

Szpara, M. Y., & Wylie, E. C. (2007). Writing differences in teacher performance assessments: An investigation of African American language and edited American English. *Applied Linguistics*, *29*, 244–266. doi:10.1093/applin/amm003

Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal*, *104*, 3–28. doi:10.1086/499740

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Washington, DC: Education Sector.

Tucker, P. D., Stronge, J. H., Gareis, C. R., & Beers, C. S. (2003). The efficacy of portfolios for teacher evaluation and professional development: Do they make a difference? *Educational Administration Quarterly*, *39*, 572–602. doi:10.1177/0013161X03257304

Turque, B. (2010, July 24). Rhee dismisses 241 D.C. teachers; union vows to contest firings. *Washington Post*. Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2010/07/23/AR2010072303093.html

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform.* Cambridge, MA: Harvard University Press.

U.S. Department of Education. (2010). *Race to the Top program: Executive summary.* Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology*, *34*, 89–101. doi:10.1016/j.cedpsych.2008.09.002

Wasley, P. (2006, June 16). Accreditor of education schools drops controversial "social justice" language. *Chronicle of Higher Education*, p. A13.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, *73*, 89–122. doi:10.3102/00346543073001089

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New York, NY: The New Teacher Project.

Wenzel, S., Nagaoka, J. K., Morris, L., Billings, S., & Fendt, C. (2002). *Documentation of the 1996–2002 Chicago Annenberg Research Project Strand on Authentic Intellectual Demand exhibited in assignments and student work: A technical process manual.* Chicago, IL: Consortium on Chicago School Research.

Wilson, M., Hallam, P. J., Pecheone, R., & Moss, P. (2006, April). *Using student achievement test scores as evidence of external validity for indicators of teacher quality: Connecticut's Beginning Educator Support and Training Program.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wilson, S. (2008). Measuring teacher quality for professional entry. In D. H. Gitomer (Ed.), *Measurement issues and assessment for teaching quality* (pp. 8–29). Thousand Oaks, CA: Sage.

Wilson, S. M., & Youngs, P. (2005). Research on accountability processes in teacher education. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA panel on research and teacher education* (pp. 591–643). Mahwah, NJ: Erlbaum.

Woolfolk, A. E., & Hoy, W. K. (1990). Prospective teachers' sense of efficacy and beliefs about control. *Journal of Educational Psychology, 82*, 81–91. doi:10.1037/0022-0663.82.1.81

CHAPTER 21

PREPARING EXAMINEES FOR TEST TAKING

Ruth A. Childs and Pei-Ying Lin

In a recent study, third-grade teachers in Ontario were asked how they would prepare their students for that province's large-scale assessment, administered each May (Childs & Fung, 2009; Childs & Umezawa, 2009). Among other questions, they were asked to imagine attending a test preparation workshop at which the presenter suggested they should teach their students strategies for answering multiple-choice questions, such as eliminating wrong choices before picking an answer. A teacher captured the ambivalence many teachers feel about test preparation:

> I would follow this suggestion [to teach strategies for answering multiple-choice items]. Practicing test taking skills would not take the place of teaching critical thinking skills or areas of the curriculum, but I don't see the purpose to not provide my students every opportunity for success, regardless of my political view. Test taking strategies are skills that they will require for high school and especially post-secondary studies. (from a web questionnaire completed by a Grade 3 teacher, February 22, 2007)

Where states or provinces administer large-scale assessments to elementary and secondary students (for an overview of such assessments, see Chapter 16, this volume), teachers may be expected—or even required—to prepare their students for these tests. Students themselves, and sometimes their parents, may seek help preparing for tests, especially entrance examinations for undergraduate or graduate university programs.

WHAT IS TEST PREPARATION?

Test preparation, defined broadly, "refers to activities, beyond normal classroom instruction or study, specifically undertaken to (a) review content likely to be covered on a test and (b) practice skills necessary to demonstrate that knowledge in the format of the test" (Crocker, 2006, p. 16). Teachers or students preparing for a test may choose to focus only on content or only on developing test-taking skills, but more commonly, they choose to address both by, for example, reviewing tests from previous years.

How do teachers and students decide how to prepare for a test? They consider the effectiveness of the activities, that is, which activities they believe will increase test scores. They may also consider the ethics of test preparation activities, especially which activities might be viewed as cheating. The time and materials available and teachers' and students' beliefs about the tests may also influence their choices. This chapter begins by discussing the ethics of test preparation.

ETHICS OF TEST PREPARATION

In a 1989 article about the pressures on schools to increase test scores, Mehrens and Kaminski listed seven practices for preparing students to take multiple-choice tests. "General instruction on objectives not determined by looking at the objectives

DOI: 10.1037/14049-021
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

measured on standardized tests" was, they wrote, always ethical, and "teaching test taking skills" was almost always ethical (p. 16). At the other extreme were two practices that were never ethical: "practice (instruction) on a published parallel form of the same test" and "practice (instruction) on the same test" (p. 16). There were also three practices that they suggested might be ethical, depending on the intended use of the test's results:

> Instruction on objectives generated by a commercial organization where the objectives may have been determined by looking at objectives measured by a variety of standardized tests . . . instruction based on objectives (skills, subskills) that specifically match those on the standardized test to be administered . . . [and] instruction on specifically matched objectives (skills, subskills) where the practice (instruction) follows the same format as the test questions. (p. 16)

Although others have questioned Mehrens and Kaminski's designations (for early challenges, see S. A. Cohen & Hyman, 1991; Haladyna, Nolen, & Haas, 1991; Kilian, 1992) or suggested additions to the list of ethical practices—for example, Haladyna et al. (1991) recommended "increasing student motivation to perform on the test through appeals to parents, students, and teachers" (p. 4)—Mehrens and Kaminski's list remains the reference against which other lists are compared.

To someone reading the test preparation literature for the first time, its emphasis on ethics might be surprising. What does it mean to call a test preparation practice ethical or unethical? Green, Johnson, Kim, and Pope (2007), in their study of teachers' perceptions of ethics in both classroom and large-scale assessment, defined ethical behavior as "acting based on one's judgment of an obligation—a duty by virtue of a relationship with a person, persons, or social institution" (p. 1000) and suggested two principles that together determine which test preparation and administration practices are ethical. One principle is "avoid score pollution"—that is, avoid practices that increase or decrease "test performance *without connection to the construct* represented by the test, producing construct-irrelevant test score variance" (Haladyna et al., 1991, p. 4; for a detailed discussion of construct-irrelevant variance as a threat to the validity of score inferences, see Haladyna & Downing, 2004; for a broader discussion of test validity, see Volume 1, Chapter 4, this handbook). Mehrens made a similar point in a 1991 symposium on test preparation, asserting that "the most general . . . principle is that a teacher should not engage in any type of instruction that attenuates the ability to infer from the test score to the domain of knowledge/skill/or ability of interest" (p. 4) and Popham (1991) argued that to be educationally defensible, "no test preparation practice should increase students' test scores without simultaneously increasing student mastery of the content domain tested" (p. 13). Indeed, the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; for an overview of the *Standards*, see Volume 1, Chapter 2, this handbook) specify that "the integrity of test results should be maintained by eliminating practices designed to raise test scores without improving performance on the construct or domain measured by the test" (AERA et al., 1999, p. 168).

The complexity of determining what precisely is—or should be—measured by a test is illustrated by Koretz, McCaffrey, and Hamilton (2001; also Koretz & Hamilton, 2006) in their framework for evaluating the validity of score gains. Test performance, according to the framework, depends on students' mastery of the curriculum content (what Koretz et al., 2006, have called substantive elements of performance) and of test-taking skills (nonsubstantive elements of performance). As Koretz et al. pointed out, tests often inadvertently over- or underemphasize some elements of each type and may even include substantive elements that are not in the test specifications. Judging whether a particular test preparation practice is likely to pollute test scores, therefore, requires understanding not only which elements a test actually measures but also which elements users of the test scores believe the test measures.

It is easy to see how advance practice with the same test form, for example, might lead to inflated

scores and therefore to a state or province's testing office or a university's admissions office overestimating a student's knowledge and skill—after all, the users of the test scores almost certainly believe that the test results represent a wider knowledge than knowing the answers to the precise items on test. Even focusing classroom instruction on what is expected to be on the test can compromise the interpretation of the test results if the results are taken to represent mastery of a larger domain of knowledge and skill than the students actually learned (Mehrens, 1991). Koretz et al. (2001) have pointed out that such narrowing of instruction may represent "reallocation"—that is, shifting of instructional time and resources from other curriculum areas to areas covered by the test—as well as "alignment"—that is, changing what is taught to better match the test, with the assumption that the test reflects the most important parts of a state or province's curriculum. These and other effects of testing on classroom practice are sometimes referred to as *washback*, or less frequently, *backwash*. Some researchers (e.g., Popham, Keller, Moulding, Pellegrino, & Sandifer, 2005) have argued that alignment to a test can be positive if the test is carefully designed. As argued elsewhere (Childs, Emenogu, Falenchuk, Herbert, & Xu, 2005), however, if tests are to be accepted as de facto curriculum documents, it will be critically important that curriculum experts and not test developers alone control their content.

Less obviously, neglecting to do any test preparation can cause test results to underestimate what students actually know. In their 1998 guidelines for large-scale performance-based assessments, Mehrens, Popham, and Ryan included the exhortation to "make certain that the student is not surprised, and hence confused, by the performance assessments' format" (p. 20). A student who is unfamiliar with the format of test items or with the timing and instructions of the test administration may find it difficult to demonstrate what she or he knows. More controversial is what is sometimes called *test wiseness*, which can include, in addition to time management and familiarity with item formats, strategies for deducing the correct response to a multiple-choice item from clues in the response options (Millman, Bishop, & Ebel, 1965). Scruggs, White, and Bennion (1986) argued that this last strategy "enable[s] test takers to score higher than would be expected based on their knowledge of the content being tested" (p. 70) and so should not be considered acceptable. Others have countered that such test-wiseness strategies are effective only on poorly developed test items (indeed, Dolly & Williams, 1986, found that a 1-hour lecture on how to guess on multiple-choice items improved undergraduate students' performance on poorly constructed items but not on well-constructed items) and so are best controlled by developing better tests (Powers, 1986, also makes this point).

Beyond the accuracy of each student's score is another consideration: the comparability of scores. If students in one school have the chance to practice answering multiple-choice items and students in another school do not, comparisons between the schools will be affected. Haladyna et al. (1991) made this point eloquently, writing that "even ethical practices are polluting if they are unevenly administered" (pp. 4–5).

Thus far, this chapter has avoided using a term that appears often in media coverage of testing: *cheating*. Haladyna and Downing (2004) defined cheating as "any deception committed to misrepresent a group's or a student's level of achievement" (p. 25) and gave as examples advance practice on the actual test items, "reading answers to students during the test or helping students select correct answers; giving hints to correct answers in the classroom during the test or changing wrong answers to rights answers after the test is given," and "intentionally excluding [low-achieving] students" (p. 25). It is striking that only one of the examples (practice on test items) happens before the test. The other examples happen during the test administration or after, and so they cannot properly be called test preparation. Indeed, the ambiguity and variability, described by Koretz et al. (2001), in what test users believe a test score represents makes it hard to conclude that any but the most extreme test preparation practices (e.g., teaching the students the answers to the actual items on the test) are deliberate deception.

Recall that Green et al. (2007) proposed two principles for determining whether a test preparation or test administration activity was ethical: "avoid

score pollution" was only one of them; the other was "do no harm." Green et al. listed several possible harms from assessments:

> There is the potential educational harm done as the result of assessments that fail to accurately measure the knowledge or skills that they claim to measure. There is also the potential emotional harm done to students in the form of anxiety or other stress that high-stakes assessments often bring about. There is also the potential for harm of the teacher-student relationship. Teacher-student trust can be damaged by assessments that the student perceives as unfair or unfounded. (p. 1009)

As Green et al. (2007) noted, teachers may not be able to avoid all possible harms, and they may not be able to simultaneously avoid harm and score pollution. These conflicting obligations mean that individual teachers may choose to act differently. As Rex and Nelson (2004), in their narrative study of two teachers' test preparation practices, observed,

> Teachers' choices about how and what to teach in preparation for a test emerge not from following, disobeying, or transcending rules. Rather, teachers act practically in the moment, over time, and in different but related contexts based upon what they are able to discern as honorable and necessary amidst conflict and ambiguity. (p. 1320)

We are not suggesting that all choices are equal—that, for example, a teacher should be excused for letting her students practice with the actual test items because she believes it will decrease their test anxiety. Test developers and state and provincial testing offices, however, should understand that exhorting teachers to be ethical is unlikely to lead all teachers to the same set of practices (Kilian, 1992, writing from a school district perspective, made a similar point). Test developers and testing offices would do better to acknowledge and seek to minimize the potential harms teachers see in the testing.

In summary, in preparing their students to take tests, teachers may consider both how test preparation activities will affect the meaning of the test scores and how the experiences of preparing for and taking the test will affect their students academically and emotionally. These considerations may also affect students' and parents' decisions about test preparation. Ideally, test preparation practices should jeopardize neither the interpretation of the test scores nor the well-being of the students.

Referring to Mehrens and Kaminski's (1989) list of test preparation practices, "general instruction on objectives not determined by looking at the objectives measured on standardized tests"—that is, teaching the curriculum content—is always ethical. Furthermore, echoing Mehrens, Popham, and Ryan's (1998) call to "make certain that the student is not surprised, and hence confused, by the performance assessments' format" (p. 20), this can be accomplished, for any type of test, by teaching test-taking skills, such as familiarizing students with the terminology used in test administration instructions and teaching students how to manage their time during a test. Some other test-taking skills, however, such as how to use clues within a multiple-choice item to guess the correct response, are likely to introduce construct-irrelevant variance. This guessing strategy is not necessary to help students understand the format of the test (some may argue that fairness requires that, if any students are taught these skills, all must be; this is a version of "everyone else is doing it"—an ethically dubious argument). In contrast, practicing on previous versions of a test, when those versions have been released for such use, or on sample tests, can be an efficient way to familiarize students with the format and administration formalities of the test. Mehrens et al.'s focus on reducing construct-irrelevant variance by preventing students from being confused by what they called "the logistics of the assessments" (p. 20) is useful in deciding which test-taking skills should be taught.

Although Mehrens and Kaminski (1989) asserted that "instruction based on objectives (skills, subskills) that specifically match those on the standardized test to be administered" (p. 16) is unethical, others (e.g., Popham et al., 2005) have argued that

narrowing of the curriculum to match a test can be appropriate if the test measures the most important parts of the curriculum. Finally, as Mehrens and Kamiski have noted, instruction or practice on the actual test items on which the students will be evaluated is always unethical.

EFFECTIVENESS OF TEST PREPARATION

Which test preparation practices work? Unfortunately, the research evidence is not conclusive—as Sturman (2003) observed, "preparation seems to arise from an intuitive belief that it can make a difference, rather than from evidence that it actually does" (p. 263). Some practices may even increase students' worry about tests: Kwok (2004), in a study of commercial after-school programs (sometimes called "cram schools") in Hong Kong, Macao, Seoul, Taipei, and Tokyo, found that the programs "reinforced open examination pressure and encouraged students to value the importance of open examinations to their life/career" (p. 71).

One of the simplest forms of preparation is practice on previous tests or on sample tests designed to be parallel in content and format to the real test. Kulik, Kulik, and Bangert (1984), in a meta-analysis of studies with students ranging in age from kindergarten to the end of university, found an average effect of almost a quarter of a standard deviation for a single practice on a parallel test. A few of the studies in their meta-analysis included additional practice and those resulted in further increases in scores.

Research on the effectiveness of teaching test-taking skills—sometimes referred to as test coaching—is more complicated. (Some researchers, such as Bangert-Drowns, Kulik, and Kulik, 1983, have reserved the term coaching for programs that teach test-taking skills, distinguishing such programs from those that offer test-taking practice or tutoring in the broad domain of content to be tested; in practice, coaching programs teach test-taking skills, but they may also include practice and tutoring.) As A. D. Cohen (2006) noted in his review of research on language testing, it is difficult to study what strategies students actually use when answering questions because asking students to think aloud may change strategy use and asking students to describe their strategy use in retrospect requires both that they were aware of using strategies at the time and that they remember afterward. Not surprisingly, most studies have taught students a variety of test-taking strategies and then have looked for changes in performance, or for differences in test performance between students who received the training and those who did not.

In a meta-analysis of 24 studies investigating the effects of teaching test-taking skills to elementary students, Scruggs et al. (1986) found an average effect of about a tenth of a standard deviation as the result of training in test-taking skills. More interesting, however, he found that students in the first to third grades benefited only when given at least 4 hours of training, but fourth to sixth graders benefited significantly from shorter training (although they too benefited more from longer training). In a meta-analysis of 10 studies including students of all ages, Powers (1986) found that teaching test-taking skills was more effective for multiple-choice items that had complex instructions and that had to be answered rapidly than for those with simpler instructions and less time pressure. Bunting and Mooney (2001) found that 11-year-old British students who received a 3-hour session providing hints and demonstrating techniques for solving the short-answer items on the high-stakes 11-plus exams performed significantly better than students who did not receive the session.

Much of the research on preparation for university or postgraduate entrance examinations, such as, in the United States, the SAT, American College Test (ACT), Medical College Admission Test (MCAT), and Law School Admission Test (LSAT), has been criticized for methodological flaws. As Becker observed in her 1990 review of test preparation programs for the SAT, research on formal programs is complicated by the wide variability in the programs' length, whether the programs teach content, such as vocabulary or mathematics, in addition to test-taking skills, and by the self-selection of the students who attend the programs. In addition, studies conducted by the test developer have tended to find smaller effects of test preparation. Nevertheless, based on a careful meta-analysis of 25 published studies, Becker concluded that test

preparation programs produced an average increase of 0.09 of a standard deviation for the verbal section of the SAT and 0.16 for the mathematics section and found that explicitly teaching test-taking skills was positively related to the size of the increase. In a review of research on preparation for the newer version of the SAT, Briggs (2009) concluded that test preparation offered by commercial testing companies or private tutoring for the SAT produced a small positive effect (0.10 to 0.20 of a standard deviation for mathematics and 0.05 to 0.10 for critical reading). He found insufficient research to conclude whether test preparation has an effect on ACT scores.

Methodological problems, including reliance on students' self-report, were also emphasized by McGaghie, Downing, and Kubilius (2004) in their review of 10 studies of commercial test preparation programs for high-stakes medical examinations, including the MCAT. McGaghie et al. concluded that "there is no trustworthy evidence that shows medical commercial test preparation courses and services have a measurable educational impact or are cost effective" (p. 210). Stricker and Wilder (2002) also relied on students' self-reports in their study of the Pre-Professional Skills Tests (PPST), required for admission to teacher education programs in some U.S. states. They found that those who reported preparing for the PPST scored lower than those who did not prepare; they suggested that "the most likely explanation for this seemingly counter-intuitive relationship may simply be that test takers who prepare most do so because they are the most deficient in the abilities assessed by the PPST" (Stricker & Wilder, 2002, p. 273).

A related research literature has investigated the effects of test anxiety (described by Sarason, 1984, as causing test-irrelevant thinking, worry, tension, and bodily reactions) on test performance and, relevant to this chapter, the effects of test-taking preparation on test anxiety. For example, Hembree (1988), concluded, on the basis of 73 studies of the performance on intelligence, aptitude, and achievement tests of students ranging from first grade to university, that students with high test anxiety scored on average almost half a standard deviation below those with low test anxiety. Hembree further found from six studies in which students were trained to increase their test wiseness that the training significantly reduced students' test anxiety, but it did not significantly improve their test performance. More recently, Beidel, Turner, and Taylor-Ferreira (1999) reported some success in teaching test wiseness to students in fourth to seventh grades, with the students reporting decreased test anxiety and earning higher classroom test scores in most subjects, but not in math.

In summary, the research evidence suggests that both practice taking similar tests and instruction in how to take tests improve students' test performance, although the improvement is typically small. It may be that some of the improvement is due to the fact that test preparation decreases test anxiety and this in turn increases performance; however, the research evidence is inconclusive.

WHAT AFFECTS TEST PREPARATION?

Ethics and effectiveness are important, but they are not the only influences on test preparation. Several recent studies in the United States have asked teachers how they prepare their students for large-scale assessments—and why. For example, Lai and Waltman (2008) surveyed a sample of schools in Iowa and found that, although teachers in most schools reported engaging in some test preparation, the specific practices varied across schools and were related to grade level. For example, teachers in elementary schools were more likely than those in middle or high schools to use practice tests or structure their classroom tests to resemble the large-scale assessment. In interviews in the same study, teachers justified their practices by referring to professional ethics, integrity, fairness, score meaning, the likelihood of a practice raising scores, and the importance of learning.

In a study across many states, Pedulla et al. (2003) investigated whether the consequences of the test results—and so the pressure on teachers to produce good results—affected test preparation practices. Not surprisingly, the teachers in states with testing programs that had important consequences for both schools and individual students were more likely to report teaching test-taking skills

(85%), although even in the states with the lowest stakes (i.e., the states with moderate stakes for schools and low stakes for students), 67% of the teachers taught these skills. Pedulla et al. also found that elementary school teachers spent more time on test preparation activities than secondary teachers.

Firestone, Monfils, and Schorr (2004) observed two distinct types of responses by New Jersey teachers to a newly introduced large-scale mathematics assessment: "Some . . . teachers responded to the new test by intensifying conventional didactic instruction and adopting short-term, decontextualized test preparation strategies, while others explored more inquiry-oriented approaches and built test preparation into their regular teaching" (p. 68). Firestone et al. found that teachers who felt more pressure about the test did more test preparation in the month before the test and were more likely to use didactic instructional approaches. Those who felt supported by their principals were more likely to prepare their students for the test throughout the year and to use inquiry-oriented instructional approaches. Being knowledgeable about the curriculum standards predicted inquiry-oriented instruction and test preparation both in the month before the test and throughout the year. Other studies of classroom practices after the introduction of large-scale assessments have also suggested that the effects may depend on teachers' knowledge about what will be on the test (e.g., Noble & Smith, 1994; Stecher & Chun, 2001).

In a study of how teachers in England prepared sixth-grade students for examinations in science, Sturman (2003) found that teachers in schools that had previous low achievement on the exams started test preparation earlier than those in middle-achieving schools and used more materials and approaches to test preparation than those in high-achieving schools. Overall, 95% of the teachers surveyed reported preparing their students for the tests and, of these,

> all exhorted their pupils to read questions carefully, to follow written instructions and to answer in sufficient detail. Almost all encouraged pupils to derive "clues" about each test item from the number of mark boxes (to find out how many marks were available) and the size of the response space (to guide the level of detail required). It was also almost universal to give pupils either advice about or opportunities to practise the skills of timing and pace, ticking the required number of boxes in multiple-choice items, using scientific vocabulary, and reading and/or interpreting information from tables or graphs. (Sturman, 2003, p. 265)

Although they were preparing their students, 50% of teachers reported that "preparation activities replaced some of their normal science activities" (p. 264), and 10% reported that test preparation replaced all other science activities.

As described earlier in this chapter, Green et al. (2007) and others have investigated teachers' beliefs about the ethics of test preparation practices, finding that teachers vary in their beliefs about what practices are ethical. This chapter has discussed some of the reasons for the variation. Additional valuable insights into teachers' perspectives on large-scale assessments and the possible implications of those perspectives for their actions have been provided by Smith (1991). On the basis of her research with teachers in Arizona, she noted that

> to chastise teachers for unethical behavior or for "polluting" the inference from the achievement test to the underlying construct of achievement is to miss a critical point: The teachers already view the indicator as polluted. . . . With an interpretive context unavailable to other groups, teachers noted the inadequacies of the mandated achievement test: its poor fit with what they teach, the influence on its scores of pupils' socioeconomic status and ethnic group, the influence of pupils' emotional state and intentional effort on test results, its many sources of error, its poor relationship with other indicators of achievement, and its limited scope. (p. 538)

The influences on students' test preparation decisions are also complex. Stricker and Wilder (2002), in their study of students taking the PPST, asked detailed questions about time and money considerations and about the students' own attitudes and those of their friends. They concluded that

> test takers who reported that they prepared the least also reported that they were confident that they would do well on the test and were not test anxious, thought that the test was easy and preparing for it was unimportant, thought that their peers shared these beliefs, and thought that their peers did little preparation, too. Ignorance of test preparation resources and experience in taking the PPST also seem to play a role, but a lesser one, and issues of time or money appear to have little or no involvement. (p. 272)

Teachers' choices of test preparation activities are affected not only by ethical considerations but also by how well the school's students have performed in the past and how important the test results are to the school, by teachers' knowledge and beliefs about the test, and by other factors, such as the age of the students. Students' choices may be affected by their confidence about their knowledge of the content and familiarity with the test format, although research on students' choices is limited.

CONCLUSION

Although there is no agreed-on "best" way to prepare students for large-scale assessments, the research suggests that familiarizing students with the test through practice taking previous or sample tests provides consistent benefits; formal instruction on test-taking skills can also lead to improvement in students' scores. Whether these test preparation practices are ethical for a specific test depends, as Green et al. (2007) suggested, on how they affect the meaning of the test scores and how they affect students academically and emotionally. Beyond simply teaching the curriculum, test preparation activities that familiarize students with "the logistics of the assessments" (Mehrens et al., 1998, p. 20) seem most likely to be both ethical and effective across a wide range of situations. This may include, for example, teaching students how to manage their time during a test or familiarizing them with the format of the test questions or the terminology used in the instructions. Practicing on released versions of a test or on sample tests can help students cope with the logistics of the actual test.

Test preparation poses continuing challenges, in part because it is but one of many activities that vie for teachers' and students' time and attention, but also because, as Smith (1991), Rex and Nelson (2004), and others have noted, the perspectives of those who develop tests may be quite different from the perspectives of those tasked with preparing for the tests. Preparing students, both effectively and ethically, to take tests presents challenges for all involved in testing. The challenge for teachers and students is to carefully consider the ethics of their situation and the time and resources available to them. The challenge for test developers is to create tests that are resistant to test-wiseness strategies and that minimize the potential harm to students. Test developers can assist teachers and students by providing wide access to information about the test and to appropriate practice materials. The challenge for researchers in this field is to design and conduct better research on how students answer test questions—and, by extension, how best to prepare them to demonstrate their knowledge and skills on tests.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research, 53*, 571–585.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60*, 373–417.

Beidel, D. C., Turner, S. M., & Taylor-Ferreira, J. C. (1999). Teaching study skills and test-taking strategies to elementary school students. *Behavior Modification, 23*, 630–646. doi:10.1177/0145445599234007

Briggs, D. C. (2009). *Preparation for college admission exams* (2009 NACAC Discussion Paper). Arlington, VA: National Association for College Admission Counseling

Bunting, B. P., & Mooney, E. (2001). The effects of practice and coaching on test results for educational selection at eleven years of age. *Educational Psychology, 21*, 243–253. doi:10.1080/01443410120065450

Childs, R. A., Emenogu, B. C., Falenchuk, O., Herbert, M., & Xu, Y. (2005). Beyond viability: Are instructionally supportive accountability tests an advisable assessment option? (Commentary on "Instructionally Supportive Accountability Tests in Science: A Viable Assessment Option?"). *Measurement: Interdisciplinary Research and Perspectives, 3*, 191–194.

Childs, R. A., & Fung, L. (2009). "The first year they cried": How teachers address test stress. *Canadian Journal of Educational Administration and Policy, 64*, 1–14.

Childs, R. A., & Umezawa, L. (2009). When the teacher is the test proctor. *Canadian Journal of Education, 32*, 618–651.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly, 3*, 307–331.

Cohen, S. A., & Hyman, J. S. (1991). Can fantasies become facts? *Educational Measurement: Issues and Practice*, *10*(1), 20–23. doi:10.1111/j.1745-3992.1991.tb00174.x

Crocker, L. (2006). Preparing examinees for test taking: Guidelines for test developers and test users. In S. M. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 115–128). Mahwah, NJ: Erlbaum.

Dolly, J. P., & Williams, K. S. (1986). Using test-taking strategies to maximize multiple-choice test scores. *Educational and Psychological Measurement, 46*, 619–625. doi:10.1177/0013164486463014

Firestone, W. A., Monfils, L., & Schorr, R. Y. (2004). Test preparation in New Jersey: Inquiry-oriented and didactic responses. *Assessment in Education: Principles, Policy, and Practice, 11*, 67–88.

Green, S. K., Johnson, R. L., Kim, D.-H., & Pope, N. S. (2007). Ethics in classroom assessment practices: Issues and attitudes. *Teaching and Teacher Education, 23*, 999–1011. doi:10.1016/j.tate.2006.04.042

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. doi:10.1111/j.1745-3992.2004.tb00149.x

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*, 2–7.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*, 47–77.

Kilian, L. J. (1992). A school district perspective on appropriate test-preparation practices: A reaction to Popham's proposals. *Educational Measurement: Issues and Practice*, *11*(4), 13–15. doi:10.1111/j.1745-3992.1992.tb00256.x

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: Praeger.

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Technical Report No. 551). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/Reports/TR551.pdf

Kulik, J. A., Kulik, C.-L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*, 435–447.

Kwok, P. (2004). Examination-oriented knowledge and value transformation in East Asian cram schools. *Asia Pacific Education Review, 5*, 64–75. doi:10.1007/BF03026280

Lai, E. R., & Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement: Issues and Practice*, *27*(2), 28–45. doi:10.1111/j.1745-3992.2008.00120.x

McGaghie, W. C., Downing, S. M., & Kubilius, R. (2004). What is the impact of commercial test preparation courses on medical examination performance? *Teaching and Learning in Medicine, 16*, 202–211. doi:10.1207/s15328015tlm1602_14

Mehrens, W. A. (1991, April). *Defensible/indefensible instructional preparation for high stakes achievement tests: An exploratory trialogue.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. Retrieved from ERIC database (ED334202).

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent. *Educational Measurement: Issues and Practice*, *8*(1), 14–22. doi:10.1111/j.1745-3992.1989.tb00304.x

Mehrens, W. A., Popham, W. J., & Ryan, J. M. (1998). How to prepare students for performance assessments. *Educational Measurement: Issues and Practice*, *17*(1), 18–22. doi:10.1111/j.1745-3992.1998.tb00617.x

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and*

Psychological Measurement, 25, 707–726. doi:10.1177/001316446502500304

Noble, A. J., & Smith, M. L. (1994). Old and new beliefs about measurement-driven reform: Build it and they will come. *Educational Policy, 8*, 111–136. doi:10.1177/0895904894008002002

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching.* Boston, MA: National Board on Educational Testing and Public Policy.

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice*, *10*(4), 12–15. doi:10.1111/j.1745-3992.1991.tb00211.x

Popham, W. J., Keller, T., Moulding, B., Pellegrino, J., & Sandifer, P. (2005). Instructionally supportive accountability tests in science: A viable assessment option? *Measurement: Interdisciplinary Research and Perspectives, 3*, 121–179. doi:10.1207/s15366359mea0303_1

Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin*, *100*, 67–77. doi:10.1037/0033-2909.100.1.67

Rex, L. A., & Nelson, M. C. (2004). How teachers' professional identities position high-stakes test preparation in their classrooms. *Teachers College Record, 106*, 1288–1331. doi:10.1111/j.1467-9620.2004.00380.x

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46*, 929–938. doi:10.1037/0022-3514.46.4.929

Scruggs, T. E., White, K. R., & Bennion, K. (1986). Teaching test-taking skills to elementary-grade students: A meta-analysis. *Elementary School Journal, 87*, 69–82. doi:10.1086/461480

Smith, J. L. (1991). Meanings of test preparation. *American Educational Research Journal, 28*, 521–542.

Stecher, B., & Chun, T. (2001). *School and classroom practices during two years of education reform in Washington State* (CSE Tech. Rep. No. 550). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/Reports/TR550.pdf

Stricker, L. J., & Wilder, G. Z. (2002). Why don't test takers prepare for the Pre-Professional Skills Tests? *Educational Assessment, 8*, 259–277. doi:10.1207/S15326977EA0803_03

Sturman, L. (2003). Teaching to the test: Science or intuition? *Educational Research, 45*, 261–273. doi:10.1080/0013188032000137256

CHAPTER 22

STANDARD SETTING

Richard J. Tannenbaum and Irvin R. Katz

Imagine a fourth grader being categorized as "proficient" on a state mathematics test. What does this label say about what the student likely knows and can do in fourth-grade mathematics? Who defined the set of knowledge and skills that describe proficiency? What score range corresponds to proficient performance and how was the minimally acceptable proficient score determined? Or, consider the case of a secondary school English teacher candidate who must pass a state licensure test to demonstrate an acceptable level of subject matter knowledge to enter the profession. What subject matter specific to secondary school English should this teacher candidate know? What test score does the preservice teacher need to earn to be considered ready to enter the profession? In both examples, use of the test involves a defined criterion or expectation, a *performance standard*, and the decision of whether the test taker has met or exceeded that performance standard. Testing professionals and policymakers use the activities and methodologies of *standard setting* to answer these questions.

Standard setting is the process of operationally defining a performance standard, such as proficient, and setting that standard to a point on a test score scale. The standard-setting process recommends a minimum test score—called the *cut score*—that must be achieved to meet the performance standard. Of course, depending on the intended use of a test, more than one performance standard might be defined and, therefore, more than one cut score recommended.

But standard setting consists of more than recommending cut scores. Standard setting occurs in the context of the overall validity argument for a test (Kane, 2006; see also Volume 1, Chapter 4, this handbook) because performance standards, and their corresponding range on the test score scale, define the intended interpretation of test results (Wiliam, 1996). Validation of this interpretation involves providing evidence for at least two underlying assumptions. First, that the performance standard reasonably represents the type and degree of knowledge and skills that should be met to reach a desired level of performance in the domain of interest. Second, that the cut score appropriately reflects the amount of the tested knowledge and skills needed to reach the performance standard. If either claim cannot be supported, it brings into doubt the use of the test score to conclude that a test taker has sufficient levels of relevant knowledge and skills.

Standard setting also occurs in the context of policy formation. The need to establish a performance standard reflects the need for a policy (decision rule). For example, a state department of education that incorporates a test into its teacher licensure process has recognized the need for a standardized measure of knowledge and skills to inform its licensure decision. The conduct of a standard-setting study provides a reasoned mechanism for developing a uniform standard of performance, both descriptively (the knowledge and skills that define the standard) and quantitatively (the recommended cut score that signifies that a test taker has met the standard).

This chapter uses the qualifier *recommended* when referring to the cut score. This usage is deliberate because in most instances, a panel of experts

DOI: 10.1037/14049-022
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

recommends a cut score to policymakers. The policymakers, not the panel, are responsible and accountable for setting the operational cut score. In our example, the policymakers would be the state department of education. The state department of education considers the recommendation from its panel of experts and decides on the cut score that will best meet its needs, which may be the recommended score or a score somewhat higher or lower. Nonetheless, by applying the same performance standard and cut score to all licensure applicants, the state department of education implements its policy deliberately and consistently.

This chapter is written for the researcher, graduate student, or novice standard-setting practitioner who might know about testing and measurement, and major psychometric issues, but may not have had formal exposure to standard setting. This chapter, therefore, is intended to provide a broad perspective of this area of measurement; it is not a how-to manual for setting a standard. There are far too many standard-setting methods and procedures for this chapter to cover. Kaftandjieva (2010) identified 60 methods, and there are numerous variations and derivations. Readers who want a more detailed accounting of specific methods are encouraged to refer to such texts as Cizek (2012); Cizek and Bunch (2007); and Zieky, Perie, and Livingston (2008).

The first section of this chapter provides an overview of standard setting, discusses the elements common to most standard-setting methods, and describes a few of the more frequently applied methods. The second section focuses on the policy-based context within which standard setting occurs, including (a) the roles and responsibilities of those persons involved in standard setting, (b) how policy making and validity concerns affect the design of standard-setting studies and the use of study outcomes, and (c) the sources of evidence that support the validity of the standard-setting process and its primary outcomes, performance standards and cut scores. The chapter closes with a discussion of evolving issues and areas for future research.

STANDARD SETTING OVERVIEW

Standard setting refers to a variety of systematic, judgment-based processes that identify a minimum test score that separates one level of *performance* (e.g., understanding, competence, expertise, or accomplishment) from another (Tannenbaum, 2011). Cizek (1993) defined standard setting as the "proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more conceivable states or degrees of performance" (p. 100). Those assignments flow from expert judgment rather than statistical calculations. Both definitions state that standard setting is foremost a process of informed judgment; setting a standard is not the equivalent of applying a statistical test to estimate a population parameter. Cizek elaborated on this point,

> Indeed, the definition [of standard setting] specifically eschews the notion of an empirically true cutting score that separates real, unique states on a continuous underlying trait (such as minimal competence) . . . standard setting rests simply on the ability of panelists to rationally derive, consistently apply, and explicitly describe procedures by which inherently judgmental decisions must be made. (Cizek, 1993, p. 100)

Zieky (2001) added: "There is general agreement now that cutscores are constructed, not found. That is, there is no 'true' cutscore that researchers could find if only they had unlimited funding and time, and could run a theoretically perfect study" (p. 45). Although the judgment-based or subjective nature of standard setting was once a source of concern (Glass, 1978), the field of measurement now recognizes standard setting as a process of reasoned judgment with a critical role in supporting test score validity (Bejar, Braun, & Tannenbaum, 2007; Kane, 2001, 2006; American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999; see also Volume 1, Chapter 4, this handbook).

Not all uses of test scores require that a standard be set. If a test score will be used to make decisions about the next step in an instructional sequence, a standard is not needed. If a test score will be used to locate the position of the test taker relative to other

test takers (e.g., percentile ranking), a standard is not needed. In general, a standard is needed when there is a defined criterion or absolute level of performance and the purpose of the test is to determine whether the test taker has met or exceeded the expectation.

Common Elements of Standard-Setting Studies

Each standard-setting study implements a particular standard-setting methodology, such as Angoff, Bookmark, Contrasting Groups, or Body of Work. These methods differ in the procedures used to define cut scores. However, implementation of these methods—the activities comprising a standard-setting study—contain similar elements regardless of methodology (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Zieky et al., 2008). For example, every method includes the judgment of experts, the definition of performance-level descriptions, and the training of experts in how to make standard-setting judgments. Standard-setting studies that include these common elements are more likely to provide the necessary procedural evidence for the validity of the study (Kane, 2001).

Selecting Panelists

Necessary expertise. Panelist judgment is at the heart of standard setting, so the makeup of that panel in terms of the represented expertise must be carefully considered when planning a standard-setting study (Hambleton & Pitoniak, 2006; Raymond & Reid, 2001). Panelists should have expertise in the content measured by the test and experience with the test-taking population. For example, in K–12 student achievement testing (see Chapter 16, this volume), panelists might be teachers of the subject area and grade level covered by the test. Sometimes panels include other stakeholders who bring a different perspective but still have the needed content expertise, such as curriculum specialists, school administrators, or even certain members of the general public.

The prevailing wisdom on the makeup of the panel has shifted somewhat over the past decade, especially concerning the role of stakeholders who are not educators or not involved in the preparation of educators. Kane (2001) stated that "given the wide-ranging impact of the policy decisions involved in setting standards for high-stakes tests, it is important to have broad representation from groups with an interest in the stringency of the standard" (p. 65). The breadth of representation on a standard-setting panel was noted by Hambleton and Pitoniak (2006) who reported that it is common for 30% of a National Assessment of Educational Progress standard-setting panel to include noneducators, such as members of the business community, military, and parent groups.

Although broad representation helps ensure that cut scores reflect all relevant stakeholder perspectives, the judgments made by panelists require some command of the tested content and familiarity with the typical performance of test takers. It is not always the case that a stakeholder, with a legitimate interest in the outcome of the standard-setting study, has the requisite content understanding or sufficient contact with the test-taking population to serve on a standard-setting panel. Plake (2008) recognized the value of involving a range of stakeholders to inform final cut score decisions of the authorized policy-making agency (e.g., a state department of education) but questioned the practice of including nonexperts on the panel.

Plake's (2008) viewpoint reflects a deliberate distinction between the *policy* aspect of setting a standard and the *procedural* aspect of recommending a cut score. The authorized agency (policymakers) establishes a policy that reflects the purpose and use of the cut scores, and ultimately decides on the operational cut scores, while taking into account the recommendation of the standard-setting panel. Nonexperts, who may be affected in some way by the application of the cut score, may offer valuable insights for policymakers to consider as they decide on the final cut scores. But, as Plake (2008) has argued, these nonexperts might not have enough relevant background knowledge to make the kinds of judgments required for standard setting. She cautioned,

> The consequences of involving these unqualified people in the standard-setting panel are severe: At best, their

> invalid responses can be removed from the data set before computing the cutscore(s); at worst, these unqualified panelists can delay, distract, and even derail the whole process. (p. 5)

Of course, Plake's concern is with expertise, not a specific job title. A noneducator who has the relevant expertise to make informed judgments would likely contribute as well as someone with an educator title.

Brandon (2004) concluded from his review of the influence of expertise on Angoff-based standard-setting judgments that expertise enhances judgments but not all panelists need to have a high level of expertise. According to Brandon, "it is appropriate to select as judges people who have at least general knowledge of the subject matter addressed in the examination" (p. 66). Stated another way, the level of content expertise required should be at least commensurate with the level represented on the test. For example, the level of subject matter knowledge needed to recommend a cut score for advanced certification in a specialty area likely exceeds that required to recommend a cut score for initial licensure in the general field to which that specialty belongs (for a discussion of licensure and certification testing, see Chapter 19, this volume).

Number of panelists. The number of panelists is an issue of both practical and measurement concern. On the one hand, study costs and recruitment challenges increase with each panelist needed. On the other hand, a greater number of panelists lends credibility to a standard-setting study, increasing the likelihood that policymakers, and other stakeholders, will accept the panel's recommendations. In addition, larger panels lead to more statistically stable results: the greater the number of panelists' judgments that inform a cut score, the smaller the standard error associated with that cut score. Smaller standard errors provide confidence that another similarly composed panel of experts, following the same standard-setting procedures, would reach a similar cut score.

Although different standard-setting methods might require different numbers of panelists to achieve credible and stable recommendations, Raymond and Reid's (2001) review of the literature suggested that 10 to 15 panelists generally result in an acceptably low standard error. Other authors have agreed: Zieky et al. (2008) suggested 12 to 18 panelists, whereas Hambleton and Pitoniak (2006) recommended slightly higher numbers (15 to 30) for K–12 student achievement tests to reach the needed panelist diversity.

Familiarizing Panelists With the Test

In most standard-setting methods, the focus of panelists' judgments is on the relationship between the tested content and the knowledge and skills that define the performance level of interest. Panelists must have an accurate understanding of what the test measures to make informed judgments. Although panelists might be chosen for their content expertise and familiarity with the test-taking population, some may be unfamiliar with the test being considered. Cizek and Bunch (2007) stated: "From the perspective of amassing validity evidence and documenting the qualifications (broadly speaking) of participants to engage in the standard-setting task, it would seem inappropriate for them *not* to have experience with the assessment" (p. 52).

To become familiar with the test, panelists typically take the test and then self-score (Cizek & Bunch, 2007). Importantly, this familiarization should be done without panelists having access to the answer keys. Having the answer key may lead panelists to view the items as easier than if the key were not available (Hambleton, 2001). The keys should be provided only after the panelists have completed the test. If a test includes extended constructed-response items, panelists may simply jot down how they would respond, rather than writing complete responses; they would then compare their notes to the scoring rubric.

In addition, panelists as a group should have an opportunity to discuss the tested content (Morgan & Michaelides, 2005), with the goal of developing a shared understanding of what the test measures and what it does not measure. Making these two observations explicit helps to ground the panel in the realities of the existing test. Nevertheless, because panelists are content experts with a genuine interest in the outcomes of the standard setting, they sometimes become distracted by what they would *prefer*

the test to measure rather than what it *actually* measures. This bias might affect their consideration of the test items, which may lead to less accurate judgments (accurate in the sense that the judgment might not reflect the panelist's intended cut score; Reckase, 2006). One way to reduce this potential bias is for the panelists to discuss their reactions to the tested content. Getting issues "on the table" eases the task of turning the discussion back to the test as it exists.

Defining Performance Levels

Arguably, the most critical component in a standard-setting study is the construction of clear and reasonable performance-level descriptions (PLDs). A PLD delineates the knowledge and skills expected of any test taker who is believed to be performing at that level (Cizek & Bunch, 2007). A PLD states what the cut score is intended to mean—that is, a test taker who meets or exceeds the cut score has the defined set of knowledge and skills. A PLD is the performance standard (Kane, 2001) and the objective of the standard-setting task is to identify the test score (cut score) that best aligns with that performance standard. PLDs, therefore, play a central role in setting a standard (Egan, Ferrara, Schneider, & Barton, 2009; Egan, Schneider, & Ferrara, 2012; Kane, 2002; Perie, 2008).

Kane (2001) included the reasonableness of PLDs in his validity argument as applied to standard setting. There are two related aspects of reasonableness: (a) that the knowledge and skills defining a performance level accurately reflect the content domain of interest, an issue of content relevance and representation, and (b) that the knowledge and skills are pitched at an appropriate level of demand or complexity.

A parallel to these reasonableness criteria may be found in the alignment framework proposed by Webb (2007), in which part of the evaluation of the alignment between a test and a set of content standards is determined by both (a) the extent to which the same content categories are present in the test and the content standards and (b) the match between the cognitive demands represented by the content standards and the test. For example, alignment would be weak if a set of reading content standards emphasizes comprehension and fluency, but the corresponding reading test emphasizes vocabulary and grammar; alignment would similarly be weak if both emphasize comprehension, but the standards value comprehension of extended, complex passages, whereas the test values comprehension of short and simple passages. Stronger alignment requires convergence of both content and cognitive demand.

PLDs should similarly be aligned with the content measured by the test, such that the expectations they embody of what a test taker should know and be able to do are connected to what the test holds the test taker accountable for knowing and demonstrating. A mismatch between the tested content and the PLDs undermines the meaningfulness of the cut scores and, therefore, the validity of the resultant classification decisions.

The alignment between the PLDs and the test content can be reinforced by developing the PLDs from a careful review of the test content specifications or blueprint (Perie, 2008). Perie cautioned that although test items may be used as supplemental sources to clarify PLDs, PLDs should not be item specific. Different forms of tests will contain different items; yoking PLDs to specific items would therefore reduce the applicability of the PLDs to other test forms. It can be helpful for panelists to support their delineation of expected knowledge and skills with behavioral indicators: observable behaviors or actions that would signal to the panelists that a test taker has the expected knowledge and skills. The example indicators are not test items, but critical incidents that help to make more tangible the defined set of knowledge and skills.

When multiple PLDs are needed for the same test, as is the case in K–12 student achievement testing, care must be taken to ensure that PLDs represent a reasonable progression of expectations (Perie, 2008). What is expected to reach "proficient" on a fourth-grade mathematics test should be more than what is expected on that same test to reach "basic," but the steepness of the increased expectation must also be considered.

Each PLD represents a range of knowledge and skills. One test taker may be at the lowest end of that range and another at the highest, but each test taker is still within that same performance level.

However, standard-setting studies seek to identify the minimally acceptable test score that signals entrance into a performance level (e.g., that a test taker on a licensure test has met the passing standard or a K–12 student has met the expectations for the basic, proficient, or advanced standard). A test taker who demonstrates enough knowledge and skill to enter a performance level is variously referred to as a *borderline test taker*, *target test taker*, or *threshold test taker; just-qualified candidate* is often used in the context of educator licensure testing because that seems to resonate better with the panelists. (This chapter uses the label *just-qualified candidate*.) The standard-setting task focuses specifically on the just-qualified candidate, so that the cut score signals the requirements to enter a performance level.

Clearly written PLDs that are understood similarly by all panelists are a prerequisite for meaningful cut scores. Hambleton (2001) stated, "When these descriptions are unclear, panelists cannot complete their tasks and the resulting performance standards could be questioned" (p. 97). There has been surprisingly little research about PLDs, however. Mills and Jaeger (1998), Perie (2008), and Egan et al. (2012) have all provided advice on how to construct PLDs. Beyond development guidelines, research is needed about how panelists actually use the descriptions to inform their standard-setting judgments. What are the cognitive processes that panelists use as they incorporate these descriptors into their judgments of the test content? PLDs are not likely to be exhaustive delineations of knowledge and skills; therefore, it is likely that panelists will need to make inferences from the descriptions as applied to the test content. What is the inferential process panelists go through? What personal preferences or biases come into play as panelists interpret the description of what a just-qualified candidate knows and is able to do? Such research would inform practical methods for developing, representing, and discussing PLDs that support panelist judgments.

Training and Practice

Panelists must receive enough training on the standard-setting method to understand the overall procedures and confidently carry out the specific standard-setting tasks (Cizek & Bunch, 2007). Sufficient time should be devoted to instruct the panelists in how they need to complete each step of the standard setting—what they must think about, the materials they are to use, and how they are to complete the judgment task and record their judgments (Hambleton & Pitoniak, 2006). After training, panelists should have the opportunity to practice completing the judgment task (Plake, 2008).

Following practice, panelists should discuss the rationales for their judgments, helping to uncover misunderstandings with the judgment task and also to make explicit some of their personal preferences (biases) that may have introduced irrelevant variance into their standard-setting judgments. For example, a panelist may attribute certain competencies to the just-qualified candidate because these competencies are highly valued by the panelist, but they do not appear in the PLD and they cannot be reasonably inferred from the description—an error of commission. An example encountered in the context of initial teacher licensure occurred when a faculty member on the panel justified a judgment based on his belief that a just-qualified teacher *develops* curriculum, even though the performance description stated that a just-qualified teacher designs a sequence of instruction aligned with the *given* content curriculum. During the discussion of the practice judgments, the person(s) facilitating the study should help panelists recognize these personal preferences, while refocusing discussion on the performance level of the just-qualified candidate.

After training and practice, panelists should be asked to complete a survey documenting their understanding of the process, the clarity with which instructions and training were provided, and their readiness to proceed to make operational standard-setting judgments (Cizek & Bunch, 2007; Zieky et al., 2008).

Training and practice are recognized as essential to establishing the validity of the standard-setting process and recommended cut scores (Morgan & Michaelides, 2005). Nonetheless, as Cizek and Bunch (2007) have observed, "In our experience—and as reflected by the comparatively thinner research base on training standard-setting participants—the topic of [standard-setting] training is one of the least well developed" (p. 50).

Standard-Setting Judgments

The nature of the standard-setting judgments panelists make depends on the specific methodology implemented. Several popular methodologies are outlined in the next section (Standard-Setting Methods). The panelists' judgments have been called the "kernel" of the standard-setting process (Brandon, 2004) because they are the data used to calculate the recommended cut scores and because the surrounding activities (e.g., taking the test, defining PLDs, training) have the primary goal of supporting judgment validity: that judgments reflect panelists' content expertise and experience by minimizing various irrelevant (to the standard-setting task) factors.

Feedback and Discussion

After panelists have made their judgments, it is common to provide panelists with feedback about their standard-setting judgments and then to engage them in a discussion (Cizek, Bunch, & Koons, 2004). The use of feedback implies the need for at least two rounds of judgments to allow panelists the opportunity to refine their judgments. In fact, three rounds often are included so that different types of feedback may be staggered over the rounds (Hambleton & Pitoniak, 2006).

Reckase (2001) discussed a continuum of feedback from "process feedback" to "normative feedback." Process feedback includes information about the relative difficulty of test items, allowing panelists to gauge how their judgments align with test takers' performance. Depending on the measurement model being used, these data may include *p* values or item response theory (IRT)–based difficulty estimates. Normative feedback includes information about the consequences, or impact, of setting cut scores. For example, the percentage of test takers who would be classified into each of the performance levels based on the recommended cut scores may be provided to panelists, followed by discussion of the reasonableness of these classifications. Reckase also defined *rater location feedback* as a point between the two ends of the continuum. This feedback focuses on showing panelists how their individual cut score judgments compare with those of other panelists, illustrating the level of agreement or convergence in the judgments of the panelists.

The specific timing or staging of feedback may vary. Reckase (2001) noted that when three rounds are included, process feedback may be introduced at the conclusion of the first round and normative feedback at the conclusion of the second round. (It would seem reasonable to include "rater location feedback" at the end of each round.) This ordering means that the panelists first have an opportunity to refine their understanding of the process with respect to item difficulty, without the potentially compelling influence of the normative feedback (the percentage of test takers classified into each performance level). This sequencing, therefore, places more emphasis on the criterion-referenced aspects of standard setting (panelists' judgments of item difficulty in relation to the defined performance levels) than on the norm-referenced aspects of the process. This sequencing is also consistent with the Peer Review Guidelines (2007) applicable to K–12 student achievement testing, which state that feedback about the percentage of test takers classified should not be a primary driver of standard-setting judgments.

In fact, not all standard-setting implementations use normative information. Hambleton and Pitoniak (2006) noted that policymakers may not support showing panelists the percentage of test takers who would be classified into each of the performance levels. Some policymakers hold that it is their responsibility alone to review this type of data and to make the appropriate decisions. Morgan and Michaelides (2005) have underscored that the use of consequence data "is controversial because it may influence the cut score decision by introducing information with potentially sensitive political ramifications that could unintentionally alter panelists' judgments. For this reason, this step may be omitted" (p. 6). The issue surrounding the inclusion of normative feedback reinforces the more fundamental truth about standard setting: It is inextricably tied to policy considerations and decisions and, as such, is subject to political, social, and economic forces (Cizek & Bunch, 2007).

STANDARD-SETTING METHODOLOGIES

This section outlines four frequently used standard-setting methods: Angoff, Bookmark, Contrasting

Groups, and Body of Work. All the methods are based on expert judgment, but differ in the decisions asked of panelists. The discussion includes two popular variations of the Angoff method, one for multiple-choice items and one for constructed-response items. As noted, interested readers should review other texts such as Cizek and Bunch (2007) and Zieky et al. (2008) for details on the design and implementation of specific standard-setting methods. The discussion for each method focuses on the judgment task facing panelists, as other aspects of standard setting were described in the section Common Elements.

Table 22.1 summarizes the methods discussed in this section. The choice of method tends to flow

TABLE 22.1

Summary of Popular Standard-Setting Methods

Method	Judgment	Comparison	Responses	Cut score calculation
Angoff (MC)	What is the likelihood that a just-qualified candidate would answer the item correctly? Of 100 just-qualified candidates, how many would answer the item correctly?	Level of the knowledge and skills of the just-qualified candidate versus knowledge and skill level required by the item	0–1 0–100	Panelist cut score: sum of all item ratings (weighted as per test scoring) Overall cut score: Mean of panelists' cut scores
Extended Angoff (CR)	What score is a just-qualified candidate likely to obtain on the item? What is the average score that 100 just-qualified candidates would likely obtain on the item?	Level of knowledge and skills of the just-qualified candidate versus knowledge and skill level implied by the item and the score categories of the item rubric	A value on the item rubric scale A value within the range of the rubric scale	Panelist cut score: sum of all item ratings (weighted as per test scoring) Overall cut score: Mean of panelists' cut scores
Bookmark (MC and CR)	Would the just-qualified candidate have a .67 probability of answering an MC item correctly, or a .67 probability of obtaining at least a specified rubric value for a CR item?	Level of the knowledge and skills of the just-qualified candidate versus knowledge and skill level required by the item and, for CR items, the score categories of the rubric	Bookmark placed on the first item that the just-qualified candidate would NOT solve at the .67 probability level	Panelist cut score: Midpoint of the RP67 theta values associated with the bookmarked item and the preceding item Overall cut score: Median of panelists' cut scores
Contrasting groups	Does a test taker have enough knowledge and skills to be classified into a performance level, as defined by the just-qualified candidate description?	Level of the knowledge and skills of the just-qualified candidate versus level of the knowledge and skills of the test taker	Categorization of test taker as (a) below or (b) at the performance level	Overlap of score distributions, logistic regression, or midpoint of the mean of scores in adjacent categories
Body of work	Does this set of test taker responses reflect enough knowledge and skills to be classified into a performance level, as defined by the just-qualified candidate description?	Level of the knowledge and skills of the just-qualified candidate versus level of the knowledge and skills implied by the response set	Categorization of response set as (a) below or (b) at the performance level	Overlap of score distributions, logistic regression, or midpoint of the mean of scores in adjacent categories

Note. MC = multiple choice; CR = constructed response.

from practical, rather than theoretical, concerns. Hambleton and Pitoniak (2006) have identified several factors to consider in this selection decision, including the structure of the test (the types of items that form the test), the number of performance levels to be differentiated, the kind of data (item and test level) available, the test's scoring model, and the available time to conduct the standard-setting study.

Angoff

The Angoff method is the most widely used and accepted method for defining cut scores, especially for licensure and certification tests (Brandon, 2004; Cizek & Bunch, 2007; Plake & Cizek, 2012). It is often referred to as "modified" Angoff because the original discussion of the approach (Angoff, 1971) presented only a general description of how a cut score may be established rather than specific implementation steps. The following discussion begins with the Angoff method for multiple-choice items, then describes the Extended Angoff method for constructed-response items.

In the Angoff method, the fundamental judgment task is for panelists to consider each item on the test and decide the likelihood (or probability) that a just-qualified candidate would answer that item correctly (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006). Panelists must consider both (a) what an item requires a test taker to know or be able to do and (b) the knowledge and skills that define the just-qualified candidate. If an item requires knowledge and skills that exceed those that define the just-qualified candidate, the item would be difficult for the just-qualified candidate, and so this test taker has a low probability of answering the item correctly. Conversely, if an item requires knowledge and skills below those expected of the just-qualified candidate, that item would be easy and this candidate would be more likely to answer the item correctly. Panelists provide their judgments using a scale from 0 to 1; scale increments may be .10 or .05, although research indicates that using .05 increments may produce a less biased cut score, especially if the cut score is likely to be at the tails of the score scale (Reckase, 2006). Each panelist's recommended cut score is the sum—weighted, if applicable—of the item probability judgments. The panel's recommended cut score is the average of the panelists' recommendations.

The same judgment task may be presented in slightly different ways to panelists. For example, panelists might be asked to consider 100 just-qualified candidates and to judge the number who would answer an item correctly. Some panelists may be more comfortable with thinking about the task in this way, rather than as a probability. During the analysis of the judgments, the numbers out of 100 are simply converted into decimals and then the same computations described previously are applied. The panels' recommended cut score is the same in either approach.

The Angoff method cannot be applied directly to constructed-response items, which are scored using multipoint rubrics; instead, one would use the Extended Angoff method (Hambleton & Plake, 1995). In this approach, panelists judge the rubric score that a just-qualified candidate would likely earn on the constructed-response item. A variation asks panelists to estimate the average score that 100 just-qualified candidates would likely earn; this variation is known as the Mean Estimation Method (Loomis & Bourque, 2001). As in the other Angoff variations, the panel's recommended cut score is the average of the sum (weighted, if applicable) of each panelist's judgments.

The Angoff method is relatively simple to prepare for and implement. The method requires only commonly available data, such as item difficulty estimates (*p* values or average item scores) and score distributions. In fact, the method may be implemented with no data, although having data is preferable. The Angoff method becomes less efficient as the number of desired cut scores increases, however, because each item must be judged once for each cut score. If three performance levels need to be differentiated (e.g., basic, proficient, and advanced), two cut scores are required: one that defines the border between basic and proficient, and one that defines the border between proficient and advanced. Each item on the test, therefore, needs to be judged twice. On a short test this may be feasible, but as the number of items increases, panelist fatigue may influence judgments.

Bookmark

The Bookmark method typically applies to K–12 student achievement testing in which multiple performance levels must be differentiated and for which tests often consist both of multiple-choice and constructed-response items (Karantonis & Sireci, 2006; Mitzel, Lewis, Patz, & Green, 2001). This method relies on the use of an ordered item booklet (OIB): a booklet of test items placed in order of their difficulty from easiest (first in the OIB) to hardest (last in the OIB). Item difficulty derives from IRT calibrations (see Volume 1, Chapter 6, this handbook). An IRT model expresses item difficulty in terms of the "response probability" that test takers at different abilities (theta values) will answer a multiple-choice item correctly or earn at least a specified rubric point on a constructed-response item. For the Bookmark method, the most frequently used response probability is .67, often denoted as RP67 (Cizek, Bunch, & Koons, 2004; Karantonis & Sireci, 2006).

Each multiple-choice item appears once in an OIB at its location of difficulty; each constructed-response item appears at a different location for each nonzero rubric value (Cizek & Bunch, 2007). For example, a constructed-response item scored on a 1-to-3 rubric will appear three times in the OIB, once at the difficulty location associated with earning at least 1 point, at the location associated with earning at least 2 points, and at the location associated with earning 3 points.

The judgment task for panelists is to review the items presented in the OIB and to place a bookmark (in practice this is often a "sticky note") on the first item encountered for which they judge a just-qualified candidate has less than a .67 probability of answering correctly or earning at least the indicated rubric value (Reckase, 2006). In essence, the bookmark differentiates content that the test taker needs to know from content the test taker does not need to know to be considered meeting the performance level (Mitzel et al., 2001). The cut score may be determined by computing the median of the RP67 theta values (ability levels) associated with the bookmarked item and the item immediately preceding the bookmarked item (Reckase, 2006). The scaled score associated with this median theta value is the recommended cut score. In practice, panelists often work in small groups, at separate tables, with a median cut score computed for each table. The recommend cut score is the median of the group medians (Zieky et al., 2008).

The Bookmark method accommodates both multiple-choice and constructed-response items, and efficiently accommodates the need to differentiate multiple performance levels (e.g., basic, proficient, advanced). The Bookmark method achieves this efficiency because panelists need to review each item only once no matter the number of performance levels to be differentiated, placing their bookmarks corresponding to progressively higher performance levels. Each higher performance level represents a greater expectation of what the just-qualified candidate at that level can do. For example, the first bookmarked location marks the boundary between basic and proficient; panelists then continue from that location in the OIB to the next location that marks the boundary between proficient and advanced.

Although efficient, the Bookmark method requires that the test was administered to sufficiently large numbers of test takers to provide stable IRT estimates of item difficulty. The method also necessitates the involvement of a statistician or psychometrician with expertise in IRT to perform the needed item calibrations and analyses. Karantonis and Sireci (2006) and Hambleton and Pitoniak (2006) have detailed other technical challenges.

Contrasting Groups

Unlike Angoff and Bookmark, which focus on *test items*, the Contrasting Groups method focuses on *test takers*. The basic judgment task is for experts, who may or may not be assembled as a panel, to assign test takers to performance levels (Brandon, 2002; Zieky et al., 2008). The experts consider (a) the knowledge and skills that define the levels (what is expected of just-qualified candidates) and (b) the knowledge and skills of test takers with whom the experts are highly familiar. Experts complete their classifications without knowledge of a test taker's score to reduce that source of potential bias. The basic concept of classifying test takers based on what is known about their knowledge and skills is likely familiar to experts, which is an appealing aspect of the method (Zieky et al., 2008).

The method places a premium on each expert being highly familiar with the knowledge and skills of each test taker he or she will be classifying. The method assumes that the experts (e.g., teachers) have been working with test takers (e.g., students) regularly, so that they have a well-formed understanding of the test takers' knowledge and skill levels. The method also necessitates that the experts focus only on test-taker knowledge and skills relevant to the test and performance levels and not on test-taker characteristics that may be admirable but outside the scope of the test, such as personality or perceived effort (Brandon, 2002). If the experts are not highly familiar with the test taker's knowledge and skills—and so cannot make reasonable classifications—the validity of any resultant cut score is in serious doubt.

On the basis of the classifications of test takers, a cut score may be determined in at least three ways (see Brandon, 2002). One method considers the distributions of the scores of the test takers classified in each level. The distributions of adjacent levels (e.g., basic and proficient) are plotted to identify the score that best differentiates the two distributions. Another approach uses logistic regression to identify the test score (at a defined probability) that best predicts classification into the higher performance level. A third approach computes the mean (or median) of the test scores in adjacent levels, taking the midpoint of those two values as the recommended cut score.

When planning a Contrasting Groups study, one must consider both the number of test takers who should be classified and the number of experts who will conduct the classifications. Larger numbers leads to greater stability of the cut score, but they may be logistically infeasible. Unfortunately, the research literature provides little direction on determining the appropriate numbers. Zieky et al. (2008) has suggested that at a minimum, 100 test takers are needed for each cut score desired but also stressed that the range of abilities reflected in the sample of test takers classified needs to be representative of the population of test takers. Regarding the number of experts, because the method is grounded in expert classifications, no single expert's classifications should be overrepresented (e.g., no more than 15% to 20% of the classifications for any cut score), as that could potentially bias the resulting cut scores.

Body of Work

The Body of Work method (Kingston, Kahl, Sweeney, & Bay, 2001; Kingston & Tiemann, 2012) is similar to the Contrasting Groups approach in that experts make classifications into performance levels, but the focus is on test-taker products, such as essays or portfolios, and not on the test takers themselves. In this method, a panel of experts reviews the collection of evidence ("the body of work") provided by the test takers (Cizek & Bunch, 2007). Similar to the Contrasting Groups approach, the experts assign the responses (products)—without knowledge of the earned scores—to one or more performance levels. Most often, the first set of classifications identify the likely range of scores within which a cut score may be located; this step is referred to as *range finding*. Next, experts review additional responses from within that range to identify the cut score; this is referred to as *pinpointing* (Kingston & Tiemann, 2012). The method can be time-consuming and logistically challenging if the test-taker products are large or complex, as is often the case with portfolios. Experts, however, should be familiar with the judgment task. For example, many teachers regularly evaluate student products and responses to assignments. Cut scores may be computed in the same ways used for the Contrasting Groups method.

STANDARD SETTING WITHIN CONTEXT

The previous section outlined the activities directly involved in recommending cut scores. As stated at the beginning of this chapter, however, standard setting includes more than just the recommending of cut scores. This section outlines the larger context surrounding a standard-setting study, including the various stakeholders and their responsibilities as well as the activities and decisions that take place before, during, and after a standard-setting study.

Standard-Setting Roles

Policy-making group. The policy-making group is the state department, board, council, agency, association, or organization that sanctions the

need for setting standards, and, in the end, is both responsible for deciding on the cut scores that will be applied and accountable for the consequences of those decisions. The decisions facing policymakers are presented throughout this section.

Standard setting involves other stakeholders, beyond the policymakers, but these groups have a less direct impact on the standard-setting study. For example, the outcomes of a standard-setting study, and the decisions made by the policymakers, affect test takers and people who make decisions based on the results of testing (e.g., parents, teachers, administrators). In some instances, policymakers assemble groups of these stakeholders to provide feedback on the outcomes of a study, which may influence the policymakers' decisions about the cut scores to apply.

Standard-setting consultant. The consultant provides the expertise on the theory and practice of standard setting. The consultant represents the organization or institution that is accountable for the standard-setting process. The consultant helps guide decisions of the policymakers or standard-setting study sponsors that influence the design of the study and, subsequently, the usage of the test scores. The standard-setting consultant works with this group to plan the study and recruit panelists, and keeps validation issues in mind so that there will be sufficient evidence to judge the reasonableness of the study's primary outcomes, the cut scores. The consultant might conduct the study (serving as the facilitator) or might train others to facilitate the study. After the study, the consultant analyses the data and writes the report to help the policy-making group interpret the results and decide on the final cut scores.

Panelists. Policymakers select and approve the panelists who serve on the standard-setting panel. Through the study designed by the standard-setting consultant and implemented by the standard-setting facilitator, the panelists make a recommendation to the policymakers about what cut scores should be applied. As described earlier, panelists should have sufficient expertise in the content of the test, be knowledgeable of the skills and abilities of test takers, and represent the larger population of people having the necessary expertise (Hambleton & Pitoniak, 2006; Plake, 2008; Raymond & Reid, 2001).

Facilitator. The facilitator leads the panelists through the standard-setting process, which has been approved by the policymakers. If someone other than the standard-setting consultant serves in this role, the consultant trains the facilitator on the specific methods to be used in a study as well as any details unique to the study (e.g., data to be provided to panelists, nature of the test and its scoring). Although the facilitator is expected to implement the standard-setting process according to the approved plan, the facilitator may need to modify the process to handle unexpected circumstances (e.g., several panelists not attending or showing up much later than the start time, policymakers deciding to include unplanned information or excluding planned information, technology resources failing). The facilitator, therefore, must be well-versed not only in the planned process but also with standard setting and measurement in general, so that he or she can make reasonable accommodations without jeopardizing the credibility and meaningfulness of the study outcomes.

Although the facilitator plays a key role in standard setting, there is scant literature devoted to facilitator training or preparation. This section outlines some implementation considerations facing facilitators and offers some implementation "tips" that we have found help a standard-setting study run smoothly and support defensible outcomes.

The facilitator reinforces the purpose of setting the standard and trains the panelists in the standard-setting method, providing corrective feedback, as needed. In general, he or she is responsible for engaging the panelists throughout the process, presenting and clarifying information and data, eliciting discussion, and monitoring interactions (Zieky et al., 2008). With regards to the latter, the facilitator must encourage all panelists to engage in the process, managing the flow of discussion so that, for example, a few panelists do not dominate the discussion; refocus panelists, as needed, to frame their judgments by the agreed-on PLDs (performance standards) and not their personal standards; and

remain cognizant of the time, so that the process progresses smoothly and is completed on time. Often at a standard setting others may be present who assist the facilitator, such as a data analyst and subject matter specialist; the facilitator is responsible for managing how these functions fit into the overall flow of the process (Zieky et al., 2008).

The facilitator also needs to maintain a balance between implementation and persuasion. That is, although the facilitator is there to implement a high-quality standard-setting process, he or she must not persuade panelists to alter their judgments to support his or her own policy or the preferred outcomes of the policymakers. Although the facilitator may, for example, communicate the consequences of a proposed cut score, he or she must not communicate that a proposed cut score is too low or too high. The latter is a value statement and, therefore, is a policy-based interpretation, which is the responsibility of the policymakers. Similarly, the facilitator's role is not to defend the test content, answer keys, or scoring rubrics. Often one or more panelists will likely find fault with some aspect of the test. Although the facilitator (or the test specialist present) may try to clarify test content, the facilitator should encourage the panelists to write down their specific concerns so that the test specialist may review them and consider how they may be addressed. This act of recording critiques of the test helps keep the concerns from being raised throughout the study, which might obstruct the standard-setting process.

Panelists should be encouraged to ask questions to get clarification and explanation as needed. However, there may be one or two panelists who are reluctant to ask a question; one strategy to encourage questions is to place the focus on the facilitators of the standard-setting process. The focus is not do *the panelists* understand, but have *the facilitators* provided sufficiently clear instructions and directions. By placing the focus on an evaluation of the facilitators' performance, panelists tend to ask more questions than when simply asked, "Are there any questions?"

The facilitator needs to be perceptive enough to recognize when a panelist seems uncomfortable with a step in the process or with how a discussion is unfolding. While surveys of panelists' reactions are collected, the facilitator must be attuned to signals of misunderstanding or discomfort throughout the process; he or she needs to address the reasons for any discomfort with tact and respect for the panelists. The facilitator needs to be flexible and resourceful, and no matter the challenge, remain outwardly calm. An appropriate amount of humor goes a long way in putting panelists at ease, in supporting a collegial and productive climate, and in helping to make the process an enjoyable experience for the panelists.

Activities Surrounding the Standard-Setting Study

Prestudy issues. Several issues regarding the test and its usage should be considered by the policy-making group before they sponsor a standard-setting study. If the group contains the appropriate blend of expertise, some of these decisions might be made within the group. Often, however, a policy-making group contracts with an external organization (i.e., the standard-setting consultant) or assembles an external technical advisory committee for advice when planning the standard-setting study. Ultimately, however, the policy-making group is accountable for the usage of the test scores, and so the final say in all matters rests with this group.

Should a cut score be set? One of the first tasks that policymakers need to complete is to define how they plan to use the test scores. What specific decision(s) will be made from the test scores? How will that planned use of the scores meet the policymaker's intended objective(s)? These and similar questions help policymakers determine the appropriateness of setting standards. Not all score uses require that there be cut scores. For example, cut scores are unnecessary if a test will be used to compare test takers' scores to each other (e.g., percentile rank) or to identify a fixed number or percentages of test takers (e.g., only the top 30 test takers may enroll in a certain course; Zieky et al., 2008).

A cut score is applicable when a defined set of knowledge, skills, abilities, or other measureable characteristics is needed to be considered ready to meet a specific purpose or expectation (Zieky et al.,

2008). Who, for example, has the needed occupationally relevant knowledge and skills to enter a profession (a licensure purpose)? Which internationally educated students, for whom English is a foreign language, have enough English listening, speaking, writing, and reading skills to be considered ready to study at a North American university, where English is the language of instruction (university admissions purpose)? In general, cut scores are appropriate when a defined criterion or a set of criteria needs to be met.

How many cut scores are needed? Although the policy-making group decides on the number of cut scores (i.e., the number of categories to classify test takers' performance) needed to fulfill the purpose of the test, this decision also should be informed by the measurement properties of the test. The test scale and its population-dependent error, the difficulty distribution of test items, the reliability of the test, and other properties all potentially limit the number of cut scores supportable by a test. For example, Subkoviak (1988) has argued that a test should contain at least six items corresponding to each performance level to be distinguished. Using that limit, Norman and Buckendahl (2008) provided a method whereby panelists judge which test items correspond to which performance levels (related to the perceived difficulty of each item). If some performance levels contain fewer than six items (Norman & Buckendahl, 2008, cited other researchers who have suggested that a greater number of items are needed to reliably discriminate between categories), a test might not support the desired number of cut scores.

In addition to the number of cut scores, the policy-making group should consider whether the cut scores will be based on a compensatory or conjunctive scoring model (O'Neill, Buckendahl, Plake, & Taylor, 2007). In a compensatory model, a total test score is the focus of the decision: Strengths in some test content areas compensate for weaknesses in other areas. In a conjunctive model, test takers need to demonstrate readiness in every subdomain of a test. The conjunctive model may be appropriate when the tested subdomains are distinct enough to apply a separate cut score to each. A conjunctive model, however, is more likely than a compensatory model to result in a greater number of false-negative outcomes (e.g., incorrectly ascribing test takers to a lower performance level; Hambleton & Slater, 1997; O'Neill et al., 2007).

Prestudy activities. Standard-setting consultants, in consultation with policymakers, should complete three prestudy activities: define performance standards, design the study, and consider validity evidence.

Define the performance standards. As described in the section Standard Setting Overview, performance standards (operationally defined by performance level descriptions [PLDs]) are central to the standard-setting process: Cut scores reflect the panelists' judgments of the minimum test scores needed to enter the performance levels (the score expected of the just-qualified candidate). The PLDs reflect the policymakers' expectations about the level of knowledge and skills required for each performance standard. These descriptions must therefore correspond to the policymakers' intended usage(s) of the test scores.

Sometimes, policymakers may construct complete PLDs; other times, they may construct only general descriptions of what the levels mean and then have a group of experts flesh out these general descriptions or rely on the standard-setting panelists to flesh them out. However constructed, the complete PLDs must be approved by the policymakers as corresponding to their expectations of performance.

Design the study. The study design may be done by the standard-setting consultant, but the policymakers review and sign off on the study, sometimes with the input of their technical advisory committee. A key element in the design of the study is the standard-setting method. The judgments that a method requires of panelists influence the study activities and the nature of validity evidence to be gathered. The previous section outlined a few popular standard-setting methods. The choice of method should take into consideration the testing purpose, intended score use, structure of the test (item types) and scoring model, and logistic and financial feasibility.

Consider validity evidence to be collected. In the process of designing the standard-setting study, the consultant (in communication with the policymakers) should (a) begin to document the activities leading to the study (e.g., development

of performance standards) and (b) plan the validity evidence to be collected during the study. The next section outlines the types of validity evidence that might be collected.

Poststudy activities. Standard-setting consultants should complete the following activities to help policymakers decide on final cut scores.

Document validity evidence. The report of the standard-setting study should document the entire process to allow the policymakers to interpret the results and make the determination of the final cut scores. Therefore, the report should contain evidence of the validity of the standard-setting process and information about the validity of the cut scores obtained to serve the intended purpose of the test.

Kane (1994; 2001) described three sources of validity evidence: (a) procedural, evidence that recommended cut scores were constructed in a reasonable way; (b) internal, evidence that the recommended cut scores represent stable judgments by the panelists; and (c) external, evidence that inferences about test takers based on the recommended cut scores correspond with inferences based on other information.

Procedural evidence. Procedural evidence focuses on the quality of the standard-setting design and its implementation. Setting a standard is a judgment-based, policy-making process, albeit grounded in measurement practices, and so documenting the process of collecting the judgments provides relevant and important validity evidence. According to Kane (1994),

> We can have some confidence in standards if they have been set in a reasonable way (e.g., by vote or by consensus), by persons who are knowledgeable about the purpose for which the standards are being set, who understand the process they are using, who are considered unbiased, and so forth. (p. 437)

Procedural evidence should include documentation regarding the purpose of the study and rationale for the chosen standard-setting method; the number, expertise, experiences, diversity, and representativeness of the panelists; the training and practice provided; how the PLDs were defined; the specific standard-setting steps followed (e.g., activities the panelists completed, types of feedback and discussions, number of rounds), and how data (e.g., item level, test level), if any, were incorporated in the standard-setting process. The level of detail addressing each of these features should be sufficient to provide policymakers and users of the standard-setting results with a clear understanding of what occurred and why, how it occurred, and what the results were and what they mean. Additionally, the detail should be at a level such that others would be able to implement a similar process.

Evidence of procedural validity should be augmented by feedback from the panelists about the standard-setting process. In most instances, panelists respond to two surveys, one immediately before the first round of judgments and one at the end of the study, although there may occasions when obtaining feedback after each round of standard-setting ratings is useful. At a minimum, before making their first round of judgments, panelists should be asked to verify their understanding of the purpose of the standard-setting study, that they have been adequately trained to carry out the standard-setting task, and that they are ready to proceed (Cizek, Bunch, & Koons, 2004; Hambleton & Pitoniak, 2006). If one or more panelists reports "not being ready to perform the standard-setting task," then additional training should be provided. At the conclusion of the standard-setting process, panelists should provide feedback on the clarity of instructions and explanations, the ease with which they followed the process of making their standard-setting judgments, the factors that influenced their judgments (e.g., PLDs, between-round discussions, cut score recommendations of other panelists), and their evaluations about whether each cut score is too low, about right, or too high (Hambleton & Pitoniak, 2006; Zieky et al., 2008).

Internal evidence. Internal validity evidence focuses on three aspects of consistency: consistency of the standard-setting method, consistency within each panelist, and consistency among the panelists. Method consistency addresses the likelihood that the recommended cut scores would be replicated. As Kane (1994) stated, "no matter how well designed

the standard-setting study is or how carefully it is implemented, we are not likely to have much faith in the outcomes if we know the results would be likely to be very different if the study were repeated" (p. 445). One way to estimate the replicability of cut scores is to (a) implement the standard-setting process with two different panels, or split one large panel into two subpanels, that are comparable in expertise, experience, and representation (Hambleton, 2001; Hambleton & Pitoniak, 2006), and then (b) calculate the standard error of the cut scores (see Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006, also discussed using generalizability theory to estimate a standard error).

It is not always feasible to conduct two standard-setting sessions or to assemble a large enough panel to form subpanels. If only one panel completes the standard-setting process, a standard error of judgment (SEJ) may be computed from the panelists' ratings; the SEJ is analogous to a standard error of the mean, whereby the standard deviation of panelists' ratings is divided by the square root of the number of panelists (Cizek & Bunch, 2007). Cohen, Kane, and Crooks (1999) suggested that an SEJ should be no more than half the value of the standard error of measurement to reduce the impact on misclassification rates (false positives and false negatives).[1]

Internal evidence also includes the variability of a panelist's ratings across rounds and a panelist's ability to discern the relative difficulty of items within rounds (intrapanelist consistency; Hambleton & Pitoniak, 2006). When more than one round of judgment occurs, panelists typically receive feedback about their ratings and about the difficulty of the items for test takers; the panelists then engage in discussion about their rating rationales and about their reactions to the item difficulty data. Panelists sometimes change their ratings in response to the feedback and discussion. No variance in ratings across rounds might signal that a panelist did not consider the feedback and discussion, and so discounted important information (Hambleton & Pitoniak, 2006). Another aspect of intrapanelist consistency is the ability of a panelist to recognize the relative difficulty of items. This ability might be evidenced by a positive correlation between ratings and actual item difficulty values. A low correlation would signal that a panelist may not understand what makes each of the test items more or less difficult, calling into question how much weight should be placed on his or her standard-setting judgments.

Internal validity evidence also includes the convergence in ratings across panelists (interpanelist consistency). Complete agreement among panelists is not expected or desired; by design, panelists represent multiple perspectives and are instructed to use their expertise and experiences to shape their standard-setting judgments (AERA et al., 1999; Hambleton & Pitoniak, 2006). Nonetheless, a large variation (standard deviation) in ratings may call into question the meaningfulness of the cut scores. For example, a lack of convergence might signal that panelists have not sufficiently considered the PLDs in their ratings. According to Hambleton and Pitoniak (2006), "acceptance of standards is obviously easier to defend when the panelists are in agreement about their recommendations" (p. 46).

External evidence. External validity evidence focuses on comparing the cut scores to other sources of evidence. In a general sense, one seeks evidence of convergence in the decisions from the different sources (Kane, 2001). An approach to collecting external evidence is to implement two different standard-setting methods to see whether they result in comparable cut scores. However, the research literature suggests that convergence is unlikely (Zieky, 2001). Different standard-setting methods direct the panelists to interact differently with the test content and to consider different types of data; it may not be realistic to expect a high degree of convergence (Kane, 2001). Nichols, Twing, Mueller, and O'Malley (2010) drew a parallel between the variance in test scores due to different item formats and the variability in cut scores due to different standard-setting methods. The field of measurement recognizes that different item formats may measure

[1]An SEJ assumes that panelists are randomly selected from a larger pool of panelists and that standard-setting judgments are independent. It is seldom the case that panelists may be considered randomly sampled, and only the first round of judgments may be considered independent. The SEJ, therefore, likely underestimates the uncertainty associated with cut scores.

somewhat-different aspects of the same construct. Thus, scores will likely be different, but this difference is accepted and valued because the multiple formats offer a more complete understanding of the construct. Nichols et al. (2010) suggested that standard-setting method variance and the sources of variance in standard setting should also be reconsidered.

An external reference may also take the form of data from another test measuring the same construct as the test used in the standard setting. One check on the reasonableness of cut scores is to compare the classification percentages of test takers on the test used during the standard setting and the percentages for these same test takers on the other test (Kane, 2001). Comparable classification percentages provide evidence of the meaningfulness and credibility of the cut scores. A related approach is to have persons who are most familiar with the test takers' knowledge and skills classify the test takers into the performance levels of interest and then compare these classifications to the classifications based on the test used in the standard setting (Kane, 2001), similar to the Contrasting Groups method.

It is a worthwhile goal to collect external validity evidence, but Kane (2001) has offered a useful caution, "none of the external checks on the validity of the standard is decisive, and, at best, each provides a rough indication of the appropriateness of the cut scores" (p. 76). In essence, this statement reminds us that setting a standard is about policies and policy formation, for which there is no objectively "right" answer (Kane, 2001).

Decide on the final cut scores. Policymakers should read the report of the standard-setting study, with its recommended cut scores, as their first step in deciding on final cut scores. As described, the report should contain both the recommended cut scores as well as some indication of the associated measurement error (e.g., the standard error of measurement of the test, the variability of the panelists' ratings). If feasible, a report should also contain estimates of the standard error around each cut score and classification errors associated with the cut scores (AERA et al., 1999). Classification errors include false positives (incorrectly ascribing a test taker's performance to a higher performance level than deserved) and false negatives (incorrectly ascribing a test taker's performance to a lower performance level than deserved).

The decision about a particular cut score should strike a balance between these two types of classification errors that coincides with the purpose of the test and the consequences of incorrect classifications. For example, for a licensure test of commercial pilots, the consequences of concluding incorrectly that a test taker is ready to be a pilot (false positive) are more severe than concluding incorrectly that a test taker is not ready (false negative). Generally, policymakers decide to reduce the more consequential error, even at the expense of the other error being raised. Of course, policymakers might decide that that each error is equally consequential and select cut scores that minimize both errors equally. Policymakers cannot forget that there will always be classification errors and that a decision regarding these errors needs to be part of the policy formation. Geisinger and McCormick (2010) presented a range of factors, including classification errors, that policymakers may consider as they decide on final cut scores.

EVOLVING ISSUES

This section addresses three up-and-coming issues related to standard setting. This discussion is not meant to be exhaustive, but rather it reflects key issues in the standard-setting literature as of the writing of this chapter. The first two issues relate to validity: (a) how standard-setting concepts, if applied in the early stages of test development, might provide a firmer foundation for the interpretation of test results, and (b) how an understanding of the cognition underlying standard-setting judgments might inform standard-setting practice and the evaluation of a study's quality. The third issue, the possibility of conducting a standard-setting study online, or virtually, is a natural outgrowth of advancing web-based meeting technologies.

Standard Setting as Part of Test Development

Several researchers (Bejar et al., 2007; Cizek, 2006; Kane 1994) have argued that performance standards

should inform assessment development to create a purposeful alignment between assessed content and the standards. If an assessment must differentiate between one or more performance levels within a content domain, test items should be written to maximize information at these levels. As Kane (1994) noted, "rather than apply the standard-setting procedures to an existing test, we would specify the performance standard and then develop the test to fit the standard" (p.430).

Unfortunately, most standard setting occurs *after* assessment development and administration (Cizek & Bunch, 2007); performance standards rarely inform test content. Yet each performance standard (PLD), through its description of knowledge and skills, represents the validity claim that test takers who reach that level should possess the described knowledge and skills. Bejar et al. (2007) have asked, "Is it sensible, considering the high stakes associated with educational assessments, to risk finding out, once the assessment has been administered, that the desired inferences are not well supported?" (p. 5).

Recently, Plake, Huff, and Reshetar (2009) introduced an approach incorporating performance levels into the test development process. The approach involves developing achievement-level descriptors (ALDs; these are the same as PLDs) from claim statements. Experts delineate a range of assessment claims and then assign claims to each ALD. Over multiple revision cycles, the experts consider each set of assigned claims, judging whether it reflects the appropriate level of expectation and the desired progression of expectations across the ALDs. Plake et al. argued that "because ALDs emerge from the same claims and evidence that shape item design and because ALDs inform form assembly specifications, the link between score performance and score interpretation is directly built into the assessment design, development, and interpretation" (pp. 18–19).

Panelist Cognition

Standard setting involves individual and group decision making. Panelists must balance their understanding of test content, the meaning of performance levels, and their own expectations and professional experiences to arrive at their decisions. Little is known, however, about the cognition underlying how panelists make their standard-setting judgments. Knowledge of panelists' reasoning can provide a theoretical basis for exploration of standard-setting practice and provide firmer grounding in how certain practices aid or hinder the validity of judgments.

What are panelists thinking when they make their judgments and how do elements of standard-setting practice influence this thinking? Within the standard-setting literature, research toward this question has taken an exploratory, qualitative approach (see Skorupski, 2012), as opposed to an experimental or theory-driven approach. For example, McGinty (2005) took notes during panelists' discussions and identified frequently mentioned difficulties. Panelists seemed particularly confused about the policy-level decisions associated with standard setting and how their expertise applied. Skorupski and Hambleton (2005) asked panelists to fill out questionnaires at specific points in their standard-setting study. They documented panelists' evolving understanding of the PLDs and the standard-setting process as well as changes in panelists' confidence in their standard-setting judgments. Ferdous and Plake (2005) conducted focus group interviews to understand what influenced panelists' decisions. All panelists reported paying attention to the just-qualified candidate definition. The panelists who assigned the lowest cut scores appeared to focus on their own students, whereas those who assigned higher cut scores tended to consider students (hypothetical or real) who fit the just-qualified candidate definition. Hein and Skaggs (2010) focused on panelists' understanding of the just-qualified candidate. Their panelists reported difficulty with visualizing a group of just-qualified candidates, preferring to consider one specific just-qualified candidate (cf. Impara & Plake, 1997). As a first step in developing a cognitive analysis of panelists' judgments, these studies provide clues as to the standard-setting factors that influence judgments.

A cognitive modeling approach might be a fruitful next step for this research area by developing explanatory mechanisms that describe not only the observable aspects of a situation (e.g., the information presented and the panelists' ratings) but also the unobservable cognitive processes that mediate

the judgments. As discussed earlier, many standard-setting methods ask panelists to estimate the likelihood of an event (e.g., a just-qualified candidate answering an item correctly), suggesting that research on probabilistic reasoning might be usefully applied to the question of panelist cognition. Much research has focused on how people make probability estimates both under laboratory conditions (Nilsson, Ohlsson, & Juslin, 2005) and in a variety of real-world settings, such as predicting success of a new product (Astebro & Koehler, 2007), evaluating military air threats (Bryant, 2007), and predicting various types of fiscal events (McCaffery & Baron, 2006). Other researchers have investigated the qualities of the cues that people attend to when making probability estimates and the effect of that information on estimation strategies. For example, people pay attention to more salient (vs. implied) information (McCaffery & Baron, 2006), are influenced more by information presented as frequencies than as proportions (Friedrich, Lucas, & Hodell, 2005), can become distracted by irrelevant information (Hall, Ariss, & Todorov, 2007), and more heavily weight cues that agree with each other (Karelaia, 2006). To the extent that people ignore certain information or focus overly much on others, it might diminish the quality of the estimates (i.e., standard-setting judgments) based on the information. By applying cognitive models of estimation to the task of making a standard-setting judgment, research can develop experimental investigations to address implications of the model for panelist selection, training procedures, development of PLDs, and other practical aspects of standard setting.

Virtual Standard Setting

Standard-setting studies typically convene experts at the same physical location. These face-to-face panels are time-consuming and relatively expensive. Experts who serve on the panel must commit several days of their time away from their school or office. The costs associated with travel, lodging, meals, and meeting facilities can be significant. These factors are compounded when, as is frequently the case, multiple panels need to be conducted concurrently or within the same general time period. Despite the growing popularity of virtual (web-based) meeting technologies throughout education and the workplace, the research literature contains few examples of virtual standard-setting studies.

How easily can a standard-setting study be adapted to a web-based environment? The research literature on virtual teams suggests that standard-setting studies possess characteristics that make them amenable to remote meeting technology. For one, standard-setting panels are temporary, existing only for the duration of the study compared with other business virtual teams that might exist for years. Because the "team" is temporary, the difficulties of building trust and cohesiveness are less challenging than for virtual teams for which these factors influence the team's long-term success (Espinosa, Slaughter, Kraut, & Herbsleb, 2007; Geister, Konradt, & Hertel, 2006). Although panelists should trust the expertise of each other to have confidence in the panel's recommendation, that needed level of familiarity with other panelists might be achieved through relatively simple sharing of information (Zheng, Veinott, Bos, Olson, & Olson 2002). In addition, distance-based environments tend to be more conducive to activities that are clearly defined and place greater emphasis on individual work (Gurtner, Kolbe, & Boos, 2007; Olson & Olson, 2000), as is the case for standard setting. Panelists make their standard-setting ratings independently, with group discussions occurring between rounds of judgments. The common elements of standard-setting studies contain only a few additional group discussions, the most interactive of which is the development of the PLDs.

Virtual standard setting also poses challenges. For example, reaching overall panel satisfaction with or acceptance of the recommended cut scores might require greater deliberation compared with a face-to-face meeting (Harvey & Way, 1999; Martins, Gilson, & Maynard, 2004) because of the need to communicate with fewer nonverbal cues. In addition, virtual standard setting poses the challenge of making secure test materials accessible to panelists (Harvey & Way, 1999; Katz, Tannenbaum, & Kannan, 2009; Schnipke & Becker, 2007). Finally, compared with face-to-face participants, participants in a virtual team might be more likely to become distracted or drop out of the team altogether (Harvey & Way, 1999; Olson &

Olson, 2000). Just as in the face-to-face meeting, the standard-setting facilitator in a virtual setting has the important role of keeping meeting participants engaged and focused on the task.

Virtual standard setting has been conducted using both the Bookmark method (Schnipke & Becker, 2007) and various modifications of the Angoff method (Harvey & Way, 1999; Katz et al., 2009). The Bookmark method may be particularly amenable because of the minimal data collected from each panelist—the identification of the bookmarked items corresponding to the performance levels. The Angoff method poses more challenge because each panelist produces one or more ratings for each test item. Web-based surveys (Katz et al., 2009) or specially designed web-based applications (Harvey & Way, 1999) ease the data collection burden considerably. Recently, Katz and Tannenbaum (2010) demonstrated that the convergence between virtual and face-to-face panels was comparable to the convergence between two independent, face-to-face panels on the same test using a modified Angoff approach.

Virtual standard setting holds much promise as an alternative, or supplement, to face-to-face meetings. Without the need to travel, panels may more readily accommodate a more diverse group of experts. Without the need for a physical meeting space, multiple panels (replications) can be conducted for the same budget, potentially providing more robust recommendations of cut scores. Thus, in addition to the cost savings, virtual standard setting may enhance the quality of the standard-setting process and resultant outcomes.

CONCLUSIONS

This chapter introduced the practical and measurement issues involved in conducting a standard-setting study as well as the larger validity and policy-making contexts in which standard setting occurs. Although based on expert judgment, standard-setting studies follow systematic procedures designed to provide evidence supporting (a) the validity of inferences about test-taker performance and (b) the goals of policymakers who make decisions based on test results. These systematic procedures consist of elements common to most standard-setting studies as well as specific methodologies appropriate for different types of tests. But these procedures are not sufficient for standard setting. A study must be designed appropriately, taking into account the eventual policy-based use of test results. A study must be implemented appropriately by a trained facilitator who can accommodate unexpected events while maintaining the integrity of the process. Beyond its procedures, standard setting is also an evolving field of research. Although much of the standard-setting research literature focuses on methodology and comparisons of alternative methodologies, recent work reconsiders the role of standard setting in test development, the cognition underlying standard-setting procedures, and the implementation of virtual standard-setting studies. As stated at the beginning of the chapter, standard setting is more than a set of procedures for recommending cut scores; it is an integral part of a testing program, influenced by and influencing policy formation, test development, and validity.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Astebro, T., & Koehler, D. J. (2007). Calibration accuracy of a judgmental process that predicts the commercial success of new product ideas. *Journal of Behavioral Decision Making, 20*, 381–403. doi:10.1002/bdm.559

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 1–30). Maple Grove, MN: JAM Press.

Brandon, P. R. (2002). Two versions of the contrasting-groups standard-setting method: A review. *Measurement and Evaluation in Counseling and Development, 35*, 167–181.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*, 59–88. doi:10.1207/s15324818ame1701_4

Bryant, D. J. (2007). Classifying simulated air threats with fast and frugal heuristics. *Journal of Behavioral Decision Making, 20*, 37–64. doi:10.1002/bdm.540

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*, 93–106. doi:10.1111/j.1745-3984.1993.tb01068.x

Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225–258). Mahwah, NJ: Erlbaum.

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). New York, NY: Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*, 31–50.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education, 12*, 343–366. doi:10.1207/S15324818AME1204_2

Egan, K. L., Ferrara, S., Schneider, M. C., & Barton, K. E. (2009). Writing performance level descriptors and setting performance standards for assessment of modified achievement standards: The role of innovation and importance of following conventional practice. *Peabody Journal of Education, 84*, 552–577. doi:10.1080/01619560903241028

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York, NY: Routledge.

Espinosa, J. A., Slaughter, S. A., Kraut, R. E., & Herbsleb, J. D. (2007). Familiarity, complexity, and team performance in geographically distributed software development. *Organization Science, 18*, 613–630. doi:10.1287/orsc.1070.0297

Ferdous, A. A., & Plake, B. S. (2005). Understanding the factors that influence decisions of panelists in a standard-setting study. *Applied Measurement in Education, 18*, 257–267. doi:10.1207/s15324818ame1803_4

Friedrich, J., Lucas, G., & Hodell, E. (2005). Proportional reasoning, framing effects, and affirmative action: Is six of one really half a dozen of another in university admissions? *Organizational Behavior and Human Decision Processes, 98*, 195–215. doi:10.1016/j.obhdp.2005.06.002

Geisinger, K. F., & McCormick, C. A. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice, 29*, 38–44. doi:10.1111/j.1745-3992.2009.00168.x

Geister, S., Konradt, U., & Hertel, G. (2006). Effects of process feedback on motivation, satisfaction, and performance in virtual teams. *Small Group Research, 37*, 459–489. doi:10.1177/1046496406292337

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237–261. doi:10.1111/j.1745-3984.1978.tb00072.x

Gurtner, A., Kolbe, M., & Boos, M. (2007). Satisfaction in virtual teams in organizations. *Electronic Journal for Virtual Organizations and Networks, 9*, 9–29.

Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes, 103*, 277–290. doi:10.1016/j.obhdp.2007.01.003

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Erlbaum.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/Praeger.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*, 41–55. doi:10.1207/s15324818ame0801_4

Hambleton, R. K., & Slater, S. C. (1997). Reliability of credentialing examinations and the impact of scoring models and standard setting policies. *Applied Measurement in Education, 10*, 19–38. doi:10.1207/s15324818ame1001_2

Harvey, A. L., & Way, W. D. (1999, April). *A comparison of web-based standard setting and monitored standard setting.* Paper presented at the annual conference of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Hein, S. F., & Skaggs, G. (2010). Conceptualization the classroom of target students: A qualitative investigation of panelists' experiences during standard setting. *Educational Measurement: Issues and Practice, 29*, 36–44. doi:10.1111/j.1745-3992.2010.00174.x

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366. doi:10.1111/j.1745-3984.1997.tb00523.x

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six methods with an application to tests of reading in EFL.* Arnhem, the Netherlands: CITO.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Erlbaum.

Kane, M. T. (2002). Conducting examinee-centered standard-setting studies based on standards of practice. *Bar Examiner, 71*, 6–13.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/ Praeger.

Karantonis, A., & Sireci, S. G. (2006). The Bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*, 4–12. doi:10.1111/j.1745-3992.2006.00047.x

Karelaia, N. (2006). Thirst for confirmation in multi-attribute choice: Does search for consistency impair decision performance? *Organizational Behavior and Human Decision Processes, 100*, 128–143. doi:10.1016/j.obhdp.2005.09.003

Katz, I. R., & Tannenbaum, R. J. (2010, April). *Comparison between face-to-face and web-based standard setting using the Angoff method.* Presented at the annual conference of the National Council on Measurement in Education, Denver, CO.

Katz, I. R., Tannenbaum, R. J., & Kannan, P. (2009). Virtual standard setting. *CLEAR Exam Review, 20*, 19–27.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the Body of Work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 219–248). Mahwah, NJ: Erlbaum.

Kingston, N. M., & Tiemann, G. C. (2012). Setting performance standards on complex assessments: The Body of Work method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 201–223). New York, NY: Routledge.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 175–217). Mahwah, NJ: Erlbaum.

Martins, L. L., Gilson, L. L., & Maynard, M. T. (2004). Virtual teams: What do we know and where do we go from here? *Journal of Management, 30*, 805–835. doi:10.1016/j.jm.2004.05.002

McCaffery, E. J., & Baron, J. (2006). Isolation effects and the neglect of indirect effects of fiscal policies. *Journal of Behavioral Decision Making, 19*, 289–302. doi:10.1002/bdm.525

McGinty, D. (2005). Illuminating the "black box" of standard setting: An exploratory qualitative study. *Applied Measurement in Education, 18*, 269–287. doi:10.1207/s15324818ame1803_5

Mills, C. N., & Jaeger, R. M. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. Hansche (Ed.), *Handbook for the development of performance standards: Meeting the requirements of Title I* (pp. 73–88). Washington, DC: Council of Chief State School Officers.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.

Morgan, D. L., & Michaelides, M. P. (2005). *Setting cut scores for college placement* (No. 2005–9). New York, NY: The College Board.

Nichols, P., Twing, J., Mueller, C. D., & O'Malley, K. (2010). Standard-setting methods as measurement processes. *Educational Measurement: Issues and Practice, 29*, 14–24. doi:10.1111/j.1745-3992.2009.00166.x

Nilsson, H., Olsson, H., & Juslin, P. (2005). The cognitive substrate of subjective probability. *Journal of Experimental Psychology Learning, Memory, and Cognition, 31*, 600–620. doi:10.1037/0278-7393.31.4.600

Norman, R. L., & Buckendahl, C. W. (2008). Determining sufficient measurement opportunities when using multiple cut scores. *Educational Measurement: Issues and Practice, 27*, 37–45. doi:10.1111/j.1745-3992.2008.00113.x

Olson, G. M., & Olson, J. S. (2000). Distance matters. *Human-Computer Interaction, 15*, 139–178. doi:10.1207/S15327051HCI1523_4

O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing-specific passing standard for the IELTS examination. *Language Assessment Quarterly, 4*, 295–317.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27*, 15–29. doi:10.1111/j.1745-3992.2008.00135.x

Plake, B. S. (2008). Standard setters: Stand up and take a stand. *Educational Measurement: Issues and Practice, 27*, 3–9. doi:10.1111/j.1745-3992.2008.00110.x

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The Modified Angoff, Extended Angoff, and Yes/ No standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods,*

and innovations (2nd ed., pp. 181–199). New York, NY: Routledge.

Plake, B. S., Huff, K., & Reshetar, R. (2009, April). *Evidence-centered assessment design as a foundation for achievement level descriptor development and for standard setting.* Paper presented at the annual conference of the National Council on Measurement in Education, San Diego, CA.

Raymond, M. R., & Reid, J. R. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Erlbaum.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159–173). Mahwah, NJ: Erlbaum.

Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard-setting methods. *Educational Measurement: Issues and Practice, 25*, 4–18. doi:10.1111/j.1745-3992.2006.00052.x

Schnipke, D. L., & Becker, K. A. (2007). Making the test development process more efficient using web-based virtual meetings. *CLEAR Exam Review, 18*, 13–17.

Skorupski, W. P. (2012). Understanding the cognitive processes of standard setting panelists. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 135–147). New York, NY: Routledge.

Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education, 18*, 233–256. doi:10.1207/s15324818ame1803_3

Subkoviak, M. J. (1988). A practitoner's guide to computation and interpretation of reliability for mastery tests. *Journal of Educational Measurement, 25*, 47–55. doi:10.1111/j.1745-3984.1988.tb00290.x

Tannenbaum, R. J. (2011). Standard setting. In J. W. Collins & N. P. O'Brien (Eds.), *Greenwood dictionary of education* (2nd ed.). Santa Barbara, CA: ABC-CLIO.

U.S. Department of Education. (2007). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of* 2001 [Revised December 21, 2007]. Washington, DC: U.S. Department of Education.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*, 7–25.

Wiliam, D. (1996). Meanings and consequences in standard setting. *Assessment in Education, 3*, 287–307. doi:10.1080/0969594960030303

Zheng, J., Veinott, E., Bos, N., Olson, J. S., & Olson, G. M. (2002). Trust without touch: Jumpstarting long-distance trust with initial social activities. *Proceedings of the Human Factors in Computing Systems, 4*, 141–146.

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–51). Mahwah, NJ: Erlbaum.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

CHAPTER 23

REPORTING TEST SCORES IN MORE MEANINGFUL WAYS: A RESEARCH-BASED APPROACH TO SCORE REPORT DESIGN

Ronald K. Hambleton and April L. Zenisky

What do test scores tell us about a respondent's educational achievement or psychological state? Providing an answer to this question is the central purpose of score reporting in the context of educational and psychological assessment, and for test developers, reporting scores (and other relevant performance data) to intended users is a critical part of the test development process. Clear and useful score reports support users in making appropriate score inferences and have an important role to play as part of efforts to explain the validity argument for a test to key stakeholders. Quite simply, reporting scores in clear and meaningful ways to users is critical, and when score reporting is not handled well, all of the other extensive efforts to ensure score reliability and validity are diminished.

Tests (or more specifically, results from tests) have taken on an increasingly consequential role in many aspects of our existence, including decisions about educational trajectories and access to opportunities for individuals, career paths, and medical and psychological diagnoses. For example, intelligence and personality tests are standard tools for clinicians, quality-of-life measures are critical to medical researchers, and educational tests serve as tools for monitoring students' educational progress and identifying their academic strengths and weaknesses. Giving stakeholders information about what test scores mean in understandable and relevant ways helps to support appropriate test use and can encourage test results being viewed as data that are actionable rather than something to be passively viewed (Hattie, 2009).

Across testing contexts the usability and quality of score-reporting materials has historically varied considerably from a simple listing of a total score to extensive breakdowns of test performance that are instructive for future improvement, and although initiatives such as the No Child Left Behind Act of 2001 (NCLB, 2002) have sparked greater awareness of educational tests and how their results can be used, the range and quality of reports being disseminated remains variable and can often result in score meaning being an area of confusion for intended users of test data (e.g., see Goodman & Hambleton, 2004; Hambleton & Slater, 1995). To this end, the area of score reporting within the broader domain of test development is growing in importance, and the responsibility of developing reports that are accessible, useful, and understandable is increasingly a priority for testing agencies and program managers.

Consider, for example, educational testing in the United States. With testing at Grades 3 to 8, and at least one grade in high school in every state under NCLB (2002), more than 25 million score reports are being sent home to parents each year. To the extent that these reports are not understood by parents and even teachers, the value of this costly endeavor is greatly diminished.

The purpose of this chapter is to provide an overview of reporting options and considerations in a range of assessment settings, with the goal of introducing a formal process for report development, including a review sheet with specific questions about various reporting elements that agencies can use to guide their report development and review.

DOI: 10.1037/14049-023
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

The focus of the chapter is mainly applicable to educational testing, but many of the issues raised and suggestions for improving score reports are equally relevant for tests measuring such psychological variables as personality, attitudes, and intelligence, and the reader is encouraged to review chapters in this handbook related to those topics, including Volume 1, Chapter 19, and Volume 2, Chapters 8 and 24; see also Volume 2, Chapter 3, on communicating individual test results.

A secondary goal of this chapter is to build on the growing literature about score reporting to provide test developers and testing programs with some tools for creating and maintaining high-quality score reports that are rooted in both current psychometric practices and the psychometric literature. Just as various processes have been developed and become accepted to ensure psychometric quality over the years for other topics in psychometrics, reporting test results is not (or should not be) an ad hoc activity that is appended to test development, but rather it is a recognized aspect of test development (see Volume 1, Chapter 9, this handbook) composed of multiple steps designed to promote quality reporting (and, ultimately, appropriate understanding of test performance among test takers and other stakeholders).

To this end, this chapter also presents a seven-step model for score report development that outlines a reasonable process for persons responsible for developing score reports to follow and provides literature-based guidance and empirical examples throughout the seven steps. The chapter begins with background on score reporting and then moves to a review of score reports with the goal of defining what score reports are across testing contexts and settings and uses this background to advance a deliberate approach to conceptualizing score reports for both individuals and groups that are mindful of test purpose, the intended users, and appropriate score inferences.

BACKGROUND ON SCORE-REPORTING PRACTICES TO DATE

Developing high-quality assessments typically follows a logical and orderly process, although as described by Downing's (2006) model outlining 12 steps for effective test development, a degree of variation certainly is present across tests depending on the type of test, the format, the administration mode, the test purpose, the intended inferences and the stakes associated with the scores, and the expertise of the test developers. All test development activities (across Downing's 12 broad areas, which include but are not limited to specifying the construct domain, item development, test design, test administration, and test scoring) are rooted in test theory and best practices drawn from the psychometric literature and the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999).

With regard to score reporting, it is highly encouraging to see that reporting test results is indeed listed as a distinct step among the 12 aspects of test development in Downing's (2006) model. This inclusion is not common. Downing makes this important point, but many others do not. As a topic within psychometrics, score reporting historically has not been positioned as an essential part of test development, and it is not done consistently with the rigor associated with test development activities, such as equating or standard setting. For example, until recently, score-reporting design and development has rarely been a line item in state test contract budgets, and the topic is rarely addressed in technical manuals for tests. Instructions for interpreting scores and explanations of score report attributes are common (e.g., explanation of stanines and percentiles) but little else is normally addressed on the topic. Research findings on score reports or score interpretations are rarely reported in technical manuals. But that situation is changing—see, for example, policy documents and state assessment technical manuals being produced by Massachusetts (Massachusetts Department of Elementary and Secondary Education, 2008) and Pennsylvania (Data Recognition Corporation, 2010).

A research base for persons charged with developing score reports has lagged behind other areas of psychometrics. For example, whereas many researchers over many years have weighed in with

many *empirical* studies on multiple aspects of psychometrics, such as detecting and interpreting differential item functioning, and these findings are regularly published, summarized, and accessible in the psychometric literature (e.g., Holland & Wainer, 1993; Osterlind & Everson, 2009), there are many fewer studies of score reporting than just about any important topic in the test development process (notable exceptions are Goodman & Hambleton, 2004; Hambleton & Slater, 1995; Wainer, Hambleton, & Meara, 1999). Furthermore, although some resources exist—including the bibliography of reporting resources by Deng and Yoo (2009) commissioned by the National Council on Measurement in Education and papers by Jaeger (2003); Ryan (2006); and Zenisky, Hambleton, and Sireci (2009)—these are often specific to educational testing or even individual testing programs, and score-reporting guidance is clearly needed across a much broader range of testing contexts and applications.

Simply put, how to proceed in the development of score reports that are useful and understandable is not always clear-cut. One example of some of the issues that can influence the extent to which score reports are understood by intended users is the nature of scores and score scales themselves. From testing program to testing program and test purpose to test purpose, most tests adopt score scales that are distinct, which can be confusing for users across test contexts; however, this limits the extent to which stakeholders can make inappropriate assumptions about score meaning because of prior experience. The SAT Critical Reading, Mathematics, and Writing test scores are reported on a scale from 200 to 800 in intervals of 10 score points, whereas the American College Test (ACT) composite score and four content scores range from 1 to 36, and the ACT subscores (e.g., Rhetorical Skill) are reported on a scale from 1 to 18 (more information on these assessments can be found in Chapter 14, this volume). Among the 50 U.S. states and their assessments used for NCLB (2002) accountability reporting, more than 50 different criterion-referenced score scales are in effect, taking into account testing across elementary and secondary grades as well as tests used for special populations such as English language learners and students with disabilities. Many norm-referenced tests report scores on a percentile scale, whereas others may use percent correct, and these two types of scores are *not* interchangeable (although the differences may not be made evident to users on the actual reports).

The Uniform Certified Public Accountant (CPA) Exam (administered to candidates for professional credentialing as certified public accountants) reports on a 0 to 99 scale with 75 as the passing score, but it is made explicit in documentation that those values are not percentages and should not be interpreted in that way. Many candidates talk about the need to obtain 75% of the score points to pass the exam, which is inaccurate. The key issue for reporting (beyond the CPA exam example) is that the idea of the percent of score points required to pass is a conceptually accessible value and—without additional context—a passing score of 75 on a 0–99 score scale is an abstraction. Examinees want to be in the group that scores more than 75, so the challenge for reporting (across tests and testing programs) is how to give actionable meaning to scores and limit score preconceptions.

Other scores that are commonly seen include intelligence quotient (IQ) scores, *T* scores, sten scores, and stanines. Ultimately, each testing program chooses a score scale to communicate a certain kind of information, but the challenge is in how to impart meaning to a number that too often is presented devoid of context to stakeholders. What does a 550 mean on the verbal section of the SAT? How should a 76 on the Uniform CPA Exam be interpreted? In isolation, these numbers are just numbers. Good reporting is at least in part about contextualizing test results in a meaningful way so that stakeholders can attach real meaning to them, in a practical way. Another demand from many users of educational tests is that the scores be linked to actions. Lexile scores are a prominent example of an effort to link reading performance to reading materials (Lennon & Burdick, 2004).

A second challenge as a significant problem for score reporting involves the contents of many reports. These errors include errors of omission (leaving out critical information) and errors of "including the kitchen sink." Years ago, some reports of student achievement were called "data dumps." The reports simply contained too much

irrelevant information for users that was typically presented in a small font to accommodate the extensive amount of information being provided on a page or two. In the former case, some reports did not provide basic information about the test being reported (including its purpose) or the purpose of the report. Other data were left out, including necessary details about the score scale, the range of performance levels, measurement error in scores, or other pertinent data (and details about that data). The other end of that continuum is the decision to include too much information on a report (working on the theory that the more data, the better), without due consideration to ensure that the report is laid out in a well-reasoned fashion, that each element is done well, as part of a coherent story describing someone's performance on a test, and that the material is appropriate for the intended audience. This problem has been referred to as *chart clutter* (see Tufte, 2001), which can lead to a report being quite overwhelming for an intended user.

Stakeholders (and their variability in understanding assessment data within and across groups) can be another important and challenging consideration for score reporting. It is critical that score reports should be crafted in a way that takes both their purpose and the users into account: These documents should be intended to be actionable (Hattie, 2009). If someone needs to use these reports, report developers should ensure that the score report can be understood in the appropriate context by the intended user and not only by a test developer, program administrator, psychometrician, or statistician who works with such data all day. The language of score reports too often is firmly situated in the jargon of statistics or assessment and educational or psychological theory, and such terms may not always carry over to users (see Goodman & Hambleton, 2004; Hambleton & Slater, 1995). This can be a particular problem with technical footnotes, but it also has been seen as a problem in other elements of reports, such as text, charts, graphs, and tables.

REPORTING AND THE STANDARDS

There is much to consider in developing reports, and reports in use span from bare bones to complex and technical documents to highly polished materials. As variable as testing programs' approaches to score reporting may be to date, the 1999 version of the *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*; AERA et al., 1999; see also Volume 1, Chapter 13, this handbook) does provide guidance for test developers across testing contexts regarding the importance of score reporting with direction for elements of report development. This direction begins with the first Standard (1.1), which reads, "A rationale should be presented for each recommended interpretation and use of test score, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation" (AERA et al., 1999, p. 17). This standard emphasizes the extent to which score reporting is not merely scores on a page but rather should be considered as the well-supported outgrowth of test validity activities. According to the *Standards*, a score report is a product that can be used to convey how scores can be understood appropriately in the context of the assessment and what are the supported actions that can be taken using the results. Indeed, after Standard 1.1, Standard 1.2 is quite explicit on this point about interpretation and use:

> The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described. (AERA et al., 1999, p. 17)

These two standards concerning validity in particular have clear application in score-reporting efforts and reinforce the idea that purposeful test development includes consideration of issues that are central to report development and score communications.

A third standard with relevance to the present topic is 5.10, which states that

> when test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations.

> The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used. (AERA et al., 1999, p. 65)

This standard echoes the validity-focused Standards 1.1 and 1.2, and also guides the person charged with developing score reports to the specifics of report contents and how those elements should be included to form the basis of a usable and understandable report.

It may be helpful to mention a few additional standards. Standard 13.14 focuses on errors and the communication of the imprecision of test scores: "In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores" (AERA et al., 1999, p. 165), which historically has been a unique challenge for reporting efforts because of difficulties in communicating test score error both conceptually and technically to different stakeholders. Interestingly, often stakeholders indicate that they do not care about the measurement imprecision information, in part, because they think the information clutters the reports. So clearly there are challenges to educating users, and ways must be found to communicate the measurement imprecision information. One solution is the use of percentile band reporting with norm-referenced tests. Through the use of text, graphics, and numbers, the concept of measurement imprecision can be communicated.

Two standards remind test developers to consider certain issues that have greater relevance for the reporting of scores in aggregate and for making comparisons across groups of test takers. Standard 13.15 states: "In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of those differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation" (AERA et al., 1999, p. 148). And Standard 13.19 states: "In educational settings, when average or summary scores for groups of students are reported, they should be supplemented with additional information about the sample size and shape or dispersion of score distributions" (AERA et al., 1999, p. 149).

WHAT ARE SCORE REPORTS?

The term *score report* is one that appears deceptively simple to understand. In its most common form, it is a test score printed on a page for a test taker, along with basic administration data such as test date, examinee name and contact information, and perhaps a performance-level classification such as pass–fail, or a description of a psychological state into which a respondent has been classified. Score-reporting efforts, however, span from a range of reports summarizing individual test performance and perhaps detailing diagnostic-type information, to group reports that are likewise varyingly summative, diagnostic, or comparative in nature. This section provides an overview of the different types of reports and report elements and how they are being used in operational testing programs.

Score reports for individuals are perhaps most prominently what stakeholders envision at the mention of the topic. These reports typically contain one or more elements that are intended to communicate information that is summative, diagnostic, or normative in nature. Report elements that are *summative* in nature include numerical test scores and proficiency or performance classifications. These components benefit from additional information, such as a notation as to the full range of the score scale, the demarcation of cut scores on the score scale between performance levels, text that defines the meaning of the performance level achieved (and perhaps other levels as well), and the inclusion of some mechanism for communicating measurement error in scores. These summative data can be represented using text, tables, and graphics. For an interesting article on the downside of using graphs and tables, and what can be done to improve them, see the work of Wainer (1992).

For individual reports, *diagnostic* report elements are increasingly being included as a significant portion of information being communicated. Diagnostic

reporting is typically imagined and implemented as a strategy for presenting test performance data that a test taker can put to use to inform preparation for a future test administration. Subscores are among the most common ways of displaying diagnostic information (typically communicated as points earned relative to the total possible points available in each of the content subareas of interest; Haberman, 2008; Haberman et al., 2009; Haladyna & Kramer, 2004; Lyrén, 2009; Monaghan, 2006; Sinharay et al., 2007). Measurement specialists have great concerns about subscale reporting because of concerns about low subscale score reliability (and sometimes the relatively high correlations among purportedly different subscales). Considerable research is ongoing to overcome the problem by locating auxiliary information, such as candidate information on related subscales, prior information provided by a teacher or counselor, or more complex scoring (such as with multiple-choice items, scoring for the suitability of each possible answer choice).

Yet another diagnostic display of achievement test performance is an item-level breakdown of performance on individual test questions, which is particularly valuable when the actual items (or a subset thereof) are released to the public. With achievement tests, for example, the item-level performance view often includes information on performance by content classification, cognitive level, and item type. This information is particular valuable at the group level (e.g., class, school, district, or even state) for which the reliability of the information can be high. At the individual candidate level, parents and teachers need to be cautioned about the relatively unreliable score information at the item level. There is also the fear that users will focus their attention only on the particular items that appear on the test. Users need to be instructed to take a broader perspective on the data (address the broader objectives measured by the test items) and need to be alerted to unreliable item-level information. This fear is manifested in the "teach to the test" criticism, which is often leveled at school achievement tests.

Diagnostic information can also take the form of text-based feedback, in which test performance is used to inform the development of short bullet points summarizing areas of relative strength and weakness for individual test takers. Recent work on the SAT by Hambleton, Sireci, and Huff (2008) used item response modeling, scale anchoring, and item mapping (methods described by Beaton & Allen, 1992; Zwick et al., 2001) to provide diagnostic information to illustrate the meaning of different score intervals on the SAT score scale and to develop performance category descriptions rooted in actual student performances. This strategy was implemented in the SAT Skills Insight online tool (College Board, 2012), in which sample questions keyed to various score intervals are provided for the SAT Reading, Mathematics, and Writing sections.

The third type of data that may be found on individual score reports is *normative* in nature. Examinees not only are interested in knowing their performance (summative information) and how to improve for the future (diagnostic information) but also typically want to know how they stand relative to other test takers. This interest cuts across both criterion-referenced and norm-referenced tests, and on the norm-referenced test, it is often handled by reporting trait scores using percentiles or percentile ranks or other types of normative scores. For criterion-referenced tests, comparisons are typically included in table or graph form, highlighting the performance of the individual test takers set against the average score obtained for relevant groups (in an educational test setting, this is often the test taker's school, district, and the state average).

Turning to group-level reports, several common strategies are employed in creating these reports. Group reports typically take the form of either list-style records of individual student performance (such as a class or grade) or results presented in aggregate on the basis of geographic units (class, school, district, state, region, and/or nation), or by other demographic variables of interest (race or ethnicity, gender, socioeconomic status, English proficiency level, participation in special education, and even migrant status). With many psychological measures, age, occupation, or health status may be prominent score-reporting categories. The choice of how to define these groups for reporting are dependent on individual testing contexts and on the reporting demographics of interest among stakeholders.

List-style reports typically are intended for use by educators within a school or district setting and generally include student-level data, such as scale score or performance-level classification for easy reference. Other score data (such as from norm-referenced tests) may be listed as well as points earned by content subdomains or an item analysis for each student in the group. These reports typically are structured as tables with each test taker's data occupying its own row.

Summative group reports can be structured several ways. Depending on the information being presented and the intended use, these reports can be formulated as tables or may incorporate graphs as well (e.g., for information on preparing graphs, see Lane & Sandor, 2009; Wainer, 1992). Tabular group reports often list such data as average score for the group or the percent of students within various performance classifications (and these may be further subdivided to reflect the geographic or demographic groupings mentioned previously).

The National Assessment of Educational Progress (NAEP) has become a leader in graphic score reporting for groups at the state and national level, including the use of line graphs to illustrate trends in performance over time and bar charts to summarize percents within achievement levels (i.e., performance categories). In 1990, NAEP reports were designed by a statistical agency in the U.S. Department of Education. In the 21st century, policymakers have a major role to play in the NAEP score report design and considerable amounts of resources are used to improve the quality of the reports for a wide range of audiences. The improvements are obvious, and the NAEP reports serve as a model for group-level reporting. For example, the methods and approaches for reporting results from the Trends in Mathematics and Science Study (TIMSS; National Center for Education Statistics, 2012d) reflect many of the strategies used in NAEP reporting, including the availability of the Data Explorer online analysis tool.

NAEP also uses line graphs to illustrate trends in score gaps for comparison groups (such as male and female examinees) over time. In recent years, technology has become more prominent in NAEP reporting of cross-state comparisons, and web-based tools (National Center for Education Statistics, 2012c) enable users to view a map of the United States color coded by performance level. Users can select different states as the unit of reference for cross-state comparison and the display changes accordingly.

Beyond these types of documents and resources that are expressed based on ways of communicating test scores for individuals or groups, score reporting can more broadly be thought of as including various ancillary materials that provide additional context for understanding test performance. As noted in Goodman and Hambleton (2004), interpretive guides are one such reference that might be made available by testing programs, as are item maps (such as those used by the NAEP; see National Center for Education Statistics, 2012b); frequently asked questions documents; technical manuals; and sample, annotated student and group-level reports.

A number of different types of score reports and report elements currently are used in practice for educational and psychological testing. Producing a report is an activity that entails both awareness of the options and careful thought about the data that are of interest to support appropriate inferences and uses of test scores.

THE MODEL

Our model for score report development is defined by seven guiding principles for score report design and validation. These principles stem from best practices in various aspects of test development, experiences in score report design, and knowledge of the score-reporting literature. Because score reporting has not always been given formal treatment, this model introduces formality into the score report development process and provides an empirically based structure for how report creation might best proceed, as informed by the psychometric literature.

A key feature of this approach to score report development is that it is empirically based and grounded in both experimental work and practical experience. Although the psychometric literature on score reports is small but growing, score reporting can be viewed as both an art and a science. This activity in test development not only draws on technical skills in terms of psychometric methods to

focus on data and produce ways to describe test performance that are reliable and justifiable, but it also benefits from collaboration with experts in public relations, graphic design, information technology, and cognitive science. Best practices in reporting are not best represented by numbers haphazardly printed on a page without context, but rather by consideration of effective and attractive ways to communicate to intended users by incorporating graphs, tables, and text, and by exploring the use of color and layout. To this end, this model has wide application in a variety of testing contexts because it builds on diverse literatures that are highly generalizable across testing contexts.

Briefly, as shown in Figure 23.1, the seven-step score report development process begins with the type of consideration that is afforded any other test development task with respect to some level of needs assessment for the activity. On the basis of the intended test score uses and interpretations, what are the needs of the key stakeholders (Step 1) who will be viewing or using it (Step 2), and what has been done before this report with examples of features that may work or may be elements to avoid (Step 3)? With that kind of information in hand, only then should active report development begin (Step 4). At that point, the score report developers should have gathered background knowledge sufficient to compile reports that are purposeful and informed by the nuances of the specific testing program, the testing population, and the report examples in relevant disciplines and testing contexts. During this step, developers also should review score report drafts carefully according to the checklist for report elements (included in the extended discussion of Step 4).

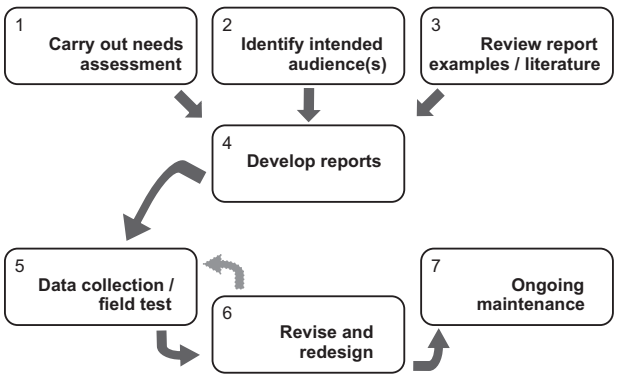


FIGURE 23.1. Model of the score report development process.

With reports drafted, the score report developers then can seek feedback from intended users about their understanding and perception of the usability of the reports, broadly consistent with the concept of instituting a mechanism for gathering feedback on test specifications or for setting standards or for how items are field-tested in item development (Step 5). With such feedback in hand, it is incumbent on the report developer(s) to filter comments into the reporting process (Step 6) and then make a judgment call about instituting a second round of external review after these changes are made. After final revision of the report(s), some program for ongoing monitoring and maintenance should be instituted (Step 7) to evaluate the ongoing use and understanding of the report material(s) and to ensure that they have the practical utility that is intended.

The Seven-Step Score Report Development Process

The description of each step begins with the guiding question that defines the step and includes the research and practical rationale for its inclusion. Steps 1, 2, and 3 can occur concurrently, as those processes feed into the report development portion of the model. In addition, where relevant, illustrative report examples are used to help in visualizing various report elements and aspects of the report development process.

Step 1: What are the information needs of key stakeholders used to guide score report development? Reports should be designed to address specific questions that stakeholders may have about the appropriate uses and interpretations of test scores. The information needs serve as the target to which all of the other steps in report design are linked. This portion of the score report development process thus seeks to link the score report and its contents explicitly to the test purpose and the validity evidence collected to support the proposed uses of the assessment, as documented by the publisher.

A wide range of reports and report elements exists, as noted. Developing reports seems to

require a clear and consistent vision of what a specific report is intended to do and the kind of information required to accomplish that purpose. Different testing organizations may have different policies and priorities about communicating test score data, and these factors certainly can influence how reporting is handled in different test settings. For example, K–12 educational testing in the United States has moved toward providing more score interpretation information to parents and educators in the years since the passage of NCLB (2002), with a focus on diagnostic information that could be used to improve (or minimally, maintain) performance on a yearly basis. As noted by Ysseldyke and Nelson (2002), one important issue concerns reporting with special populations: When preparing reports that involve students with disabilities, special care must be taken in terms of appropriate interpretations, and it may be necessary to include confidentiality and additional cautionary statements.

Many certification and licensure programs have begun to build more informative reports as well, and one key example of this is the Uniform CPA Exam. Examples of the these reports are available online (AICPA, 2012). In credentialing, when diagnostic reporting is provided, performance relative to the cut score is typically the focus of reporting efforts, and the information usually is being generated for those examinees who did not pass. It is not common to provide diagnostic information for candidates who pass a credentialing exam.

Step 2: Who are the audiences for the score report and what audience characteristics should be considered to support the choice of information and level of detail needed? Simply put, test stakeholders and the intended users of score reports matter. Different audiences have different data needs and familiarity with the assessment being reported, a critical aspect of reporting that needs to be considered as reports are conceptualized. Thus, this step should be completed early in the report development process. In pulling together score reports that have relevance and practical utility, the report development team should think critically about stakeholder characteristics and background to communicate appropriate test score use and interpretation in meaningful ways (Jaeger, 2003).

The NAEP's online reporting resources has differentiated pages for groups, such as selected schools, parents, students, researchers, the media, and educators (and each of those distinct pages is accessible from a single website; see National Center for Education Statistics, 2012a). By tailoring score-reporting resources to different users by design, testing programs can incorporate a measure of efficiency into reports and promote their use by the different groups.

Step 3: What does the literature contain regarding examples of student and parent reports or whichever reports are of interest? Increasingly, more and more agencies are making sample reports and interpretive guides available for users to consult through online resources. Although this addition has obvious benefits for the stakeholders of those testing programs, other agencies can include such materials in their own report development review process and see how different groups represent different pieces of score information, such as scores and performance-level descriptions. The type of literature review that is important at this step is somewhat less about empirical research studies on reporting (although these do exist and can be quite useful) and instead are premised on (a) identifying elements of reports that may or may not work in a given context and (b) how such elements are included in the reports from a layout and design perspective.

Reports for individuals typically are expected to include a final or total score for each test or subject area being reported, and it can be useful to be explicit in text or with visuals as to what the score scale is and how the individual performance stands relative to that scale in terms of achievement and errors of measurement. Many tests also report performance levels using terms keyed to quality descriptions (such as *advanced*, *proficient*, *basic*, or *below basic*), or they report performance relative to an absolute standard (such as pass or fail) or to a reference point (such as the performance of passing candidates or possibly the average performance of candidates). When performance levels or perhaps psychological states for psychological tests are

present, some reports will provide a text-based description of what constitutes that level of performance, perhaps in paragraph form or in a bullet-style list. A breakdown of the total score into subscores is another option that many tests are increasingly using, and an even finer level of grain size in reporting performance is to include an item analysis of an individual's test performance, notating questions answered correctly and incorrectly as well as item attributes such as content and cognitive classifications. Norm-referenced performance indicators such as how an individual's performance compares with other groups like a class, a school, a district, or a state are also common. Ultimately, a review of existing individual reports will identify these and other score report elements that may be necessary in a given reporting context.

Many of the subtest or subscale reports available should be viewed with skepticism because subscores are typically imprecise and communicating unreliability is a difficult task typically compounded by the desire to avoid negative assumptions of assessment quality. Even when score bands are used to address the unreliability issue, they can be confusing, and often they are not included, for just that reason, further complicating interpretation of differences across subscales. A useful document on this topic is provided by Monaghan (2006). The harm resulting from misinterpretation is one of the strongest reasons against some subtest reporting. At the same time, more research can address the misinterpretation of subscores because these scores appear to be popular in practice.

Likewise, different agencies make different decisions about layout and design, and a literature review of the type recommended in this section will provide considerable insight into how different reports are structured. Some reports are oriented vertically whereas others are horizontal. Some reports integrate graphics as a way to communicate score information, whereas others primarily use text for the same purpose. In the context of student reports, the websites of many U.S. states include links to sample reports, and Goodman and Hambleton (2004) included an informative series of examples in their paper. Considerably fewer sample reports are available in the credentialing domain, although the AICPA does have sample reports available, as does the National Council of State Boards of Nursing (2012).

Step 4: In developing score reports, how can information from Steps 1, 2, and 3 be integrated into the process, and how are diverse talents involved (e.g., psychometricians, graphic designers, policymakers, curriculum specialists, public relations specialists)? During this step of report development, the report design team begins to draft reports in light of considerations such as the information needs of stakeholders (Step 1), characteristics of the intended users (Step 2), and an identified list of report elements to be included or omitted based on a review of literature and previously disseminated reports (Step 3). Each of these dimensions plays a critical role in informing the initial drafts. At this point, it is important to reinforce the concept of a collaborative approach to report development. By bringing a range of people to the table in this draft development step, agencies can facilitate integration of elements of the reports by prioritizing the interrelatedness of report contents and its appearance.

Also, a review form such as that presented in Table 23.1 should be used. The 34 questions on the review form are divided among eight report element areas: Area I: Needs Assessment; Area II: Content, Report Introduction and Description; Area III: Content, Scores and Performance Levels; Area IV: Content, Other Performance Indicators; Area V: Content, Other; Area VI: Language; Area VII: Design; and Area VIII: Interpretive Guides and Ancillary Materials. Although considerations related to report contents are integral to the process, the design team must be aware of issues that may arise relative to design, language, and the production of supplementary reporting materials.

Area I is intended to give the checklist user an opportunity to reflect holistically on the report given the first three steps in the report development process. Developers should take a moment to critically examine report drafts to consider the extent to which they have fashioned a report that is aligned with what the intended users of the report call for. Area II encompasses content considerations that concern the report introduction and description.

TABLE 23.1

Review Sheet for Score Report Development and Evaluation

Report element	Score report review questions
Area I: Needs assessment	A. Does the score report reflect the reporting interests of key stakeholders?
Area II: Content—Report introduction and description	A. Does the report have a title clearly identifying what it is? B. Are details provided about the test(s) being reported? C. Is there information describing the unit of analysis being reported? D. Are the purpose(s) of the test described? E. If present, does the introductory statement from the sponsoring agency (e.g., governor, commission, president, psychologist, etc.) set a positive tone for the report?
Area III: Content—Scores and performance levels	A. Is the range of the score scale communicated? B. Are the performance categories or psychological states being used (e.g., failing, basic, proficient, advanced, passing) described sufficiently for the intended audience? C. Is information provided for how all of the numerical scores and classifications should be used and should not be used? D. Are concrete examples provided for the use of the test score information? E. Is the topic of score imprecision handled for each score that is reported? Descriptions, graphics, or numbers are all possibilities. F. Have "probabilities" or "conditional probabilities" been avoided? If they are used, is the explanation clear? G. Have footnotes been avoided, but if they are used, are they clearly written for the reader? H. Is there sufficient information for the reader, without being overwhelming?
Area IV: Content—Other performance indicators	A. Is there any linking of test results to possible follow-up activities? For example, with educational tests, are the results linked to possible instructional follow-up? B. If present, are relevant reference group comparisons reported with information on appropriate interpretations? C. If present, are results of performance on individual test questions reported with a key for understanding the item attributes and the performance codes? D. If subscale reporting is included, are users informed about the level of score imprecision? If norms are provided, is the reference group described in sufficient detail? Are the meanings of *T* scores, *z* scores, normalized *z* scores, stanines, stens, percentiles, and grade equivalent scores made clear? E. If present, are reports of scores from other recent and relevant tests (norm-referenced tests, etc.) explained?
Area V: Content—Other	A. Does the report provide telephone numbers, website addresses, or mailing addresses for resources to which questions can be directed?
Area VI: Language	A. Is the report free of statistical and other technical jargon and symbols that may be confusing to users? B. Is the text clearly written for users? C. Is the report (or ancillary materials) translated or adapted into other languages, and if so was the translation or adaptation carried out by more than a single person and was an effort made to validate the translated or adapted version?
Area VII: Design	A. Is the report clearly and logically divided into distinct sections to facilitate readability? B. Is a highlight or summary section included to communicate the key score information? C. Is the font size in the different sections suitable for the intended audience? D. Are the graphics (if any) presented clearly to the intended audience? E. Is there a mix of text, tables, and graphics to support and facilitate understanding of the report data and information? F. Does the report look friendly and attractive to users? G. Does the report have a modern "feel" to it with effective use of color and density (a good ratio between content and white space)? H. Is the report free of irrelevant material or material that may not be necessary to address the purposes of the report? I. Is the "flow" for reading the report clear to the intended audience starting with where the eyes should go first?
Area VIII: Interpretive guides and ancillary materials	A. Is there an interpretive guide prepared, and if so, is it informative and clearly written? Has it been field-tested? Are multiple language versions available to meet the needs of intended readers? B. If there is an interpretive guide, is there an explanation of both acceptable and unacceptable interpretations of the test results?

Across all contexts of reporting (individual and group, educational testing, and otherwise), reports should be clearly labeled and defined *on the report document*. This includes information about what test(s) data are included and the unit of analysis being reported. For individual reports, personalization of the report often adds a perception of accessibility to the report, and details such as grade, school, district, and state help to set the context as well. Some testing programs may include some reference to the purpose of the scores or the testing program items to be proactive about appropriate use and interpretation of test scores. Some reports also include an introductory statement about the assessment, its goals, and appropriate uses from persons in the sponsoring agency, and they should be reviewed for content and tone.

Area III addresses the specifics of reporting scores and performance levels, which as noted are (logically) the data points most prominently featured on a score report. These elements are intended to help agencies ensure that the scores, score scales, and performance levels are unambiguously defined and can be readily understood, with a focus on being concise.

Area IV is related to content but converges on other aspects of report elements, with an emphasis on context. The link of assessment results to instruction is a key area for reporting achievement, but it should be handled with considerable attention to detail and caution to ensure that users are aware of how those results can and cannot be used. A similar level of care should be taken when including item analysis results, subscale reporting, and other test data.

Area V is brief but critical. In disseminating a report, it is necessary to give users a mechanism by which they can communicate with the testing agencies concerning questions about score interpretation and use.

Area VI involves the language of the report. Review of the report for use of appropriate terminology, especially given intended users, is necessary. Too often (e.g., see Hambleton & Slater, 1995), score reports incorporate language that although technically correct, is unfamiliar or even daunting to the individuals who are interested in the data being reported (Hattie, 2009). Report developers need to be aware of this and strive to be simple, clear, and direct in all aspects of reporting. In many education settings, there is a significant (and positive) move toward translating these documents into multiple languages as necessary based on local populations. This is an encouraging development with respect to the kind of outreach that agencies should do to support score interpretation activities given high levels of diversity within different education settings (and with respect to the students and their parents or guardians). This inclusion means that score reports and ancillary materials disseminated to intended users must go through a level of scrutiny with respect to the quality of translation and the communication of score use across languages and cultures. Massachusetts, for example, currently translates its test score interpretation guide into 10 languages.

The elements in Area VII include design considerations. The intent of the questions is to prompt report developers to consider the report from a logical and orderly design perspective. By sectioning the page, highlighting certain key results, and using other design elements such as color, the development team can put together some reasonable options for consideration. The use of mixed methods for communicating results should be considered, with the possibility of using text, graphics, and tables to illustrate different reporting elements as appropriate. Substantial research in the literature supports this principle (e.g., see Hattie, 2009). Area VIII is not targeted at the score report document itself but rather is a nod to the increasing presence of auxiliary materials, which is a positive development. With the production of score interpretation guides and the like, those materials should likewise be thoroughly reviewed and evaluated to the standards of the reports themselves.

Step 5: How are reports field-tested? As noted, field-testing is routine in many aspects of test development. The historical tendency of agencies to disproportionately allocate fewer resources to score reporting has correspondingly resulted in reports not being appropriately field-tested. In recent years, a number of research studies have been undertaken that illustrate the extent to which results from various

field-test activities can lead to significant improvements in reports. Such approaches are recommended by both Jaeger (2003) and Allalouf (2007).

Wainer, Hambleton, and Meara (1999) provided an instructive study involving experiments. A set of potentially troublesome NAEP reports were identified, and the reports were redesigned. Participants in the study were randomly assigned to see the reports in either their original or revised form. Identical questions were asked of participants. Participants were shown various displays of results using NAEP data and asked specific, probing questions about the information presented. These researchers were able to illustrate that the redesign of certain displays led to much higher rates of understanding among the study's participants. Although their focus was on NAEP reporting of group-level data and was important for NAEP reporting efforts, the methodological contribution of this study has special relevance. When in doubt about the best choice of design for a report, an experimental study involving the optional reports could be revealing.

Focus groups are another technique that should be considered. Zenisky, Hambleton, and Smith (2006) and Zenisky, Delton, and Hambleton (2006) have provided two examples of research studies that use focus group techniques to study understanding of score reports. In these studies, the focus was again on NAEP group-level reports, but the process of bringing together groups of stakeholders and developing questions that probe issues related to the utility and understanding of a report is useful. Many states have used similar methodologies with their own score reports.

Furthermore, other research methods can be used to evaluate score reports. As noted in Zenisky and Hambleton (2007), certain online resources associated with NAEP were evaluated using think-aloud protocols that took place while users of reports were seated at computer terminals and navigating the online tools. This methodology shows considerable promise. This approach, sometimes called *cognitive interviewing*, is widely used in survey development.

Step 6: How are the results from the field-test used in the redesign of the reports? Given the focus on field-testing score reports, these results should be filtered into the design process and (as necessary) should result in revisions to the draft reports. Depending on the extent of the changes, it may be worth an agency's time to pilot the reports again on the same or a smaller scale to gather additional data about report quality. This feedback loop is marked in gray in Figure 23.1.

Step 7: What is the plan to evaluate the reaction to the score report or reports when they are used operationally so that more revisions can be made for the next operational use? Once reports are in operational use, their use and utility must be monitored. Agencies should consider ways to connect with intended users and identify not only if reports and other score-reporting documents are being used but also *how* they are being used and likewise should probe the quality of the inferences being made. As more innovative approaches to reporting are being included as report elements (e.g., item analysis breakdowns and growth models), the statistical intricacies of these report elements have the potential to be sources of confusion for stakeholders.

CONCLUSION

Score reports can take on many forms and display a wide range of data; following formal steps in report development as outlined in this chapter is one way to standardize what at times has been a nonstandard process and can provide guidance on a testing task that is not typically the province of psychometricians. Score reporting begins with being purposeful in every aspect. The purpose of the score report must be a foremost consideration. Whether the data being included are summative, diagnostic, normative, or some combination of the three, the information and how it is communicated should be prepared to promote its use, which springboards from the test purpose. Whatever the test purports to measure and be used for, the information on the score report should complement and further those aims. Along these lines, what score report recipients need and their characteristics must be prioritized. Some report designers want to impress one another with their creativity, whereas the goal should be effective

communication with the intended audiences, in terms of information needs. Being explanatory, specific, and clear are priorities in this regard. This goal is aided by checking what has been done before in terms of both exemplar reports in particular testing contexts (educational, psychological, certification, and licensure) and the growing literature in the field of empirical score reporting and review articles that summarize professional experience and practice.

The first three aspects of the model explicated in this chapter provide background as well as the necessary footing for the actual report development. Early editions of NAEP reports, for example, missed the policymaker audience almost completely because policymakers did not have the statistical background to understand these reports. Indeed, report length was at times prohibitive for even the most skilled policymakers. Many states have had similar experiences with their student and parent reports.

As with many elements of test development, report development is a process and being purposeful, aware of audience(s), and knowledgeable of practice makes the process go much more smoothly, especially in terms of the preparation of report drafts. These first reports should indeed be considered drafts, in that input from others within the testing agency and intended users will be beneficial. Comments received should be integrated into reports, and revisions should be circulated again as necessary. It is surprising how rarely score reports are actually field-tested in any form. Monitoring reports over time also can provide helpful information about the use and understanding of reports.

Communicating test score information matters. Stakeholders want to know what scores are and what they mean. This places a significant responsibility on the shoulders of testing agencies to prioritize reporting on par with other test development activities, which in most cases may be a departure from practice, requiring a shift in how agencies view their relationship with examinees and score users. Increasingly, users want to be able to act on test score information, and agencies should embrace the opportunity to take the lead in guiding users to appropriate test score interpretations rather than perpetuating ambiguity or failing to clarify erroneous information as such (mis)understandings arise. The trend toward including additional and more useful diagnostic information is a major force in this evolution of score-reporting practices. By working with stakeholders to develop reports to address their needs in a principled and research-based way, per the model presented in this chapter, agencies can meet stakeholders' needs more effectively.

References

Allalouf, A. (2007). An NCME instructional module on quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice*, *26*(1), 36–43. doi:10.1111/j.1745-3992.2007.00087.x

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Institute of Certified Public Accountants. (2012). *CPA exam FAQs: Candidate performance report and samples*. Retrieved from http://www.aicpa.org/BecomeACPA/CPAExam/PsychometricsandScoring/CandidatePerformanceReport/DownloadableDocuments/Sample_Diagnostic_Reports.pdf

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*, 191–204. doi:10.2307/1165169

College Board. (2012). *Improve your SAT scores with the SAT Skills Insight*. Retrieved from http://sat.collegeboard.org/practice/sat-skills-insight

Data Recognition Corporation. (2010). *Technical report for the 2010 Pennsylvania system of school assessment*. Maple Grove, MN: Author. Retrieved from http://www.portal.state.pa.us/portal/http;//www.portal.state.pa.us;80/portal/server.pt/gateway/PTARGS_0_123031_1040547_0_0_18/PSSA_Technical_Report_2010.pdf

Deng, N., & Yoo, H. (2009, April). *Resources for reporting test scores: A bibliography for the assessment community*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA. Retrieved from http://h96-60-107-202.mdsnwi.tisp.static.tds.net/resources/blblio1/NCME.Bibliography-5-6-09_score_reporting.pdf

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–301). Mahwah, NJ: Erlbaum.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research.

Applied Measurement in Education, *17*, 145–220. doi:10.1207/s15324818ame1702_3

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229. doi:10.3102/1076998607302636

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*, 79–95. doi:10.1348/000711007X248875

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions*, *27*, 349–368. doi:10.1177/0163278704270010

Hambleton, R. K., Sireci, S., & Huff, K. (2008). *Development and validation of enhanced SAT score scales using item mapping and performance category descriptions* (Final report). Amherst: University of Massachusetts, Center for Educational Assessment.

Hambleton, R. K., & Slater, S. C. (1995). *Are NAEP executive summary reports understandable to policy-makers and educators?* Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

Hattie, J. (2009, April). *Visibly learning from reports: The validity of score reports.* Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Mahwah, NJ: Erlbaum.

Jaeger, R. M. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress* (Working Paper 2003–11). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Lane, D. M., & Sandor, A. (2009). Designing better graphs by including distributional information and integrating words, numbers, and images. *Psychological Methods*, *14*, 239–257. doi:10.1037/a0016620

Lennon, C., & Burdick, H. (2004). *The Lexile framework as an approach for reading measurement and success*. Durham, NC: MetaMetrics. Retrieved from http://www.lexile.com/m/uploads/whitepapers/Lexile-Reading-Measurement-and-Success-0504_MetaMetricsWhitepaper.pdf

Lyrén, P.-E. (2009). Reporting subscores from college admission tests. *Practical Assessment, Research, and Evaluation*, *14*(4). Retrieved from http://pareonline.net/pdf/v14n4.pdf

Massachusetts Department of Elementary and Secondary Education. (2008). *Ensuring technical quality: Policies and procedures guiding the development of the MCAS tests*. Malden, MA: Author. Retrieved from http://www.doe.mass.edu/mcas/tech/technical_quality.pdf

Monaghan, W. (2006). *The facts about subscores* (ETS R&D Connections No. 4). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RD_Connections4.pdf

National Center for Education Statistics. (2012a). *NAEP: Information about NAEP for you*. Retrieved from http://nces.ed.gov/nationsreportcard/infofor.asp

National Center for Education Statistics. (2012b). *NAEP mathematics: Map of selected item descriptions on the NAEP Mathematics Scale—Grade 4*. Retrieved from http://nces.ed.gov/nationsreportcard/itemmaps

National Center for Education Statistics. (2012c). *The nation's report card: National Assessment of Educational Progress (NAEP)*. Retrieved from http://nces.ed.gov/nationsreportcard

National Center for Education Statistics. (2012d). *Trends in International Mathematics and Science Study (TIMSS)—Overview*. Retrieved from http://nces.ed.gov/timss

National Council of State Boards of Nursing. (2012). *Candidate performance report*. Retrieved from https://www.ncsbn.org/1223.htm

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115. *Stat*, 1425 (2002).

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: Sage.

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 677–710). Mahwah, NJ: Erlbaum.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21–28. doi:10.1111/j.1745-3992.2007.00105.x

Tufte, E. R. (2001). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher*, *21*, 14–22.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, *36*, 301–335. doi:10.1111/j.1745-3984.1999.tb00559.x

Ysseldyke, J., & Nelson, J. R. (2002). Reporting results of student performance on large-scale assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 467–480). Mahwah, NJ: Erlbaum.

Zenisky, A. L., Delton, J., & Hambleton, R. K. (2006). *State reading content specialists and NAEP reporting: Use and understanding of selected data displays* (Technical report for the Comprehensive Evaluation of NAEP; Center for Educational Assessment Report

No. 596). Amherst: University of Massachusetts, School of Education.

Zenisky, A. L., & Hambleton, R. K. (2007). *Navigating "The Nation's Report Card" on the World Wide Web: Site user behavior and impressions* (Technical report for the Comprehensive Evaluation of NAEP; Center for Educational Assessment Report No. 625). Amherst, MA: University of Massachusetts, School of Education.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education*, *22*, 359–375. doi:10.1080/08957340903221667

Zenisky, A. L., Hambleton, R. K., & Smith, Z. R. (2006). *Do math educators understand NAEP score reports? Evaluating the utility of selected NAEP data displays* (Technical report for the Comprehensive Evaluation of NAEP; Center for Educational Assessment Report No. 587). Amherst: University of Massachusetts, School of Education.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*(2), 15–25. doi:10.1111/j.1745-3992.2001.tb00059.x

CHAPTER 24

MULTIPLE TEST FORMS FOR LARGE-SCALE ASSESSMENTS: MAKING THE REAL MORE IDEAL VIA EMPIRICALLY VERIFIED ASSESSMENT

Neil J. Dorans and Michael E. Walker

Decisions based on high-stakes tests have important consequences for most test takers. Promotion to the next grade in school, graduation from high school, admission to a college of choice, admission to a graduate or professional school, and licensure in a profession are all examples of high-stakes uses of test scores. When decisions like these are being made, all examinees should be given an equal opportunity to maximize their performance on a level playing field. Some aspects of this opportunity lie outside the locus of control of test developers, such as opportunity to learn, quality of teaching, and test-taker motivation. Other components can be controlled by the test developer. Multiple forms of the test should be as parallel as possible. Scores across administrations should be as equivalent as possible. Score reliability should be large enough to inspire trust in the scores. Test instructions should be clear and widely accessible to prospective test takers.

This chapter focuses on testing programs that develop multiple forms of tests to produce scores used in high-stakes assessments. Given these high stakes, these testing programs should place a premium on empirical verification of their assessments. This chapter focuses on tests of ability, although the framework discussed here may be adapted to other venues. This chapter also limit its discussion to tests composed of several test questions, excluding pure, subjectively scored, performance assessments (e.g., juried competitions). Tests covered by this discussion may include subjectively scored components—for example, essays scored by readers. These subjective components should be governed by the same principle of empirical verification as other components of the assessment.

The title of this chapter reflects its approach: *Multiple Test Forms for Large-Scale Assessments: Making the Real More Ideal via Empirically Verified Assessment*. The principle of empirically verified assessment (EVA) may be stated as follows:

> Decisions about every aspect of testing—design, assessment, administration, scoring, reporting, and recommended use—should be based on empirical evidence that the chosen testing aspects maximize the utility of scores for their intended use while maintaining fairness and protecting the rights of test takers.

This principle, which could be articulated in many ways, is consistent with the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999). EVA practitioners should work to provide concrete, empirical proof that they are doing their best to adhere to these standards.

Our intent in this chapter is to move existing practices toward principled EVA. This chapter argues that any scores reported for a test used in high-stakes settings should offer reliable and nonredundant information. These scores should be interchangeable with scores that come from different editions of that test. These requirements pertain

DOI: 10.1037/14049-024
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

to both objectively and subjectively scored portions of the test. The scores should be reported on properly maintained scales that are aligned with the purposes of the test. Steps need to be taken to ensure that these scores are properly interpreted and that they are used validly. All facets of a testing program, whenever possible, should be backed by empirical evidence of their relevance with respect to the intended purpose of the test.

The chapter has four sections. A fundamentals section discusses the importance of specifying purposes, populations, and products to shape the large-scale assessment. The fundamentals section also discusses the need for well-defined content and statistical specifications and the need to control for the influence of measurement conditions as well as the importance of establishing an appropriate score scale. In the next section, the focus is on essential properties of the scores that are reported, including high reliability, interchangeability across forms, and validity. The importance of monitoring scale stability over time is also addressed. The third section emphasizes that proper data collection is essential to pool replenishment, score equating, and quality assurance on a regular basis. A summary constitutes the final section.

FUNDAMENTALS

Several basic questions must be addressed by large-scale assessments that use multiple test forms to produce scores that have high-stakes consequences. Foremost among these: Why is there a need for the assessment? To answer this "why" question, it is essential to identify who is being assessed. Also, the products of an assessment, typically scores from tests, need to be aligned to the purposes of the assessment.

Why Assess?

It is essential to identify the purposes for the assessment. Why is it necessary to measure some individual attribute? To what use will this information be put? The purpose may be highly focused. A test designed to certify readiness to teach mathematics in elementary school is one such example. Sometimes the purpose is couched within a complex assessment design framework such as evidence-centered design (ECD), as elucidated by Mislevy and his colleagues (Mislevy, Almond, & Lukas, 2003; Mislevy, Steinberg, & Almond, 2002). Once the purpose of the test has been established, this information should inform the particular knowledge, skills, and abilities (KSAs) targeted by the assessment. A rationale will show the connection between these KSAs and observable, measurable behavior (Mislevy & Haertel, 2006). In some cases, test makers begin the process with the construct to be measured. In most cases, identifying the construct to be measured cannot take the place of articulating the purpose of the test. As form should follow function, so too should purpose drive assessment design.

In defining the purposes of the assessment, it is essential to consider all plausible uses of the scores. For example, college admissions tests are often designed to predict performance the first year of college, but the scores from these tests may be used for decisions regarding course placement and scholarship award decisions as well. A test designed for one particular purpose is unlikely to serve another purpose as well as tests designed specifically for that second purpose, but the test's appropriateness for that secondary use can and should be evaluated empirically. Professional standards maintain that tests that produce scores that are designed for multiple uses must be validated for each of those uses (AERA et al., 1999).

Target Population

The *target population* is defined as the group for whom the test is designed. Members of the target population are a subset of the *test-taking population*, the people who actually take the test. Typically, the nontarget portion of the test-taking population is small in nature. For example, college-bound juniors and seniors compose the SAT target population. A small group of precocious individuals in junior high school also take the test, although they are not part of the target population.

For some tests, there are major differences in the target and testing populations. For example, a test designed to measure proficiency in a second language acquired in high school classes may be taken

by native speakers of the language and these native speakers may represent the greatest percentage of the test taking population. The test may be simple for them but hard for those learning the language in high school. Depending on the purpose of the assessment, this advantage that native speakers possess can be viewed as construct relevant or inappropriate. For example, it is appropriate to use a score on a Chinese test for placement out of a course in advanced Chinese regardless of whether the test taker is a native speaker of Chinese. If, on the other hand, a test score in Chinese contributes to a composite score that is used to determine admissions to a university, then non-Chinese test takers are being disadvantaged relative to Chinese test takers.

Alignment of a Test to Its Intended Uses

Having determined what the test is supposed to measure and why and for whom, the next step is to convert these concepts into a concrete reality. The content domain needs to be clearly specified and linked to the purpose of the assessment and its intended use with the target population. Kane (2006) discussed how to link content to intended uses in a valid manner. The final product of content specification is a blueprint indicating the number of items of various item formats in each content subdomain that will appear on the test.

Two general classes of items are typically used: selected response (e.g., multiple-choice items; see Volume 1, Chapter 18, this handbook) and constructed response (e.g., essays). Selected-response items have the advantage of being quick to answer and machine scorable. Thus, a test can contain large quantities of selected response items, covering a large content area. For a discussion of the pros and cons of constructed versus selected response, see Wendler and Walker (2006)

Practitioners of EVA should critically evaluate the test blueprint, including the particular blend of selected- and constructed-response items on the test, in various ways. Does the test as designed achieve the purpose of the test? Are scores maximally useful to test score users? Does the test design maintain fairness among test takers, and does it protect their rights? Production cost does not necessarily enter into an EVA argument (except insofar as those costs, when passed onto the test taker, may be seen as having unfair consequences). Cost, however, is a factor test makers will consider when designing a test.

Innovations of the past decade, such as ECD (Mislevy et al., 2002, 2003), offer conceptual tools to help assessment designers frame the content specifications of the test. As more test makers discover how to incorporate ECD methods into their own production processes, these methods promise to gain even more widespread use (Snow et al., 2010). ECD approaches the construction of assessments in terms of evidentiary arguments, and the validity argument for the test becomes part of its formal development. ECD allows the test developer to (a) consider the skills to be measured, (b) identify the evidence that indicates that the skills are present, and (c) construct questions that reflect this evidence. The full ECD framework has several phases (Mislevy & Haertel, 2006; Mislevy et al., 2002). This framework ensures that all versions of the test follow the same test specifications.

Much of the evidence to which ECD refers is theoretical, not empirical. It is essential to build in empirical checks on the validity of the validity arguments themselves, which should not be taken on faith in theory alone. Furthermore, ECD provides a framework for ensuring the content specificity of a test, with task models allowing for the understanding and control of evidential variation in test forms. The danger exists that people will misinterpret ECD task models as capable of producing interchangeable items (e.g., Hendrickson, Huff, & Leucht, 2009), although this is not the purpose for which ECD was intended (Mislevy et al., 2003). The statistical properties of items produced using ECD methodology still need to be empirically determined.

Statistical specifications state the desired statistical properties of the test that should guide a test assembly process that ensures parallelism across multiple versions of the test. The difficulty of a test question is a function of the percent of examinees who responded correctly to it (i.e., *p* value). These *p* values are influenced by the group of examinees who responded to the question. Ideally, the *p* values should be estimated for the target population. At the very least, average test difficulty and some measure

of spread of item difficulty around the average in the target population must be established.

Where possible, the distribution of item difficulty (i.e., spread of item difficulty and shape of the item distribution) should be specified. The shape of the distribution of item difficulty needs to reflect the intended uses of the test. For example, for narrow-range tests such as a licensing test, the greatest precision needs to be at the point on the reporting scale at which certification decisions are made. In this case, the distribution of difficulties would be peaked in this region. For a test used for broad-range assessment, where test takers' abilities fall across the entire scale range, the distribution of item difficulty will be spread out. It is important to consider both question content and format when determining the appropriate distribution of item difficulty.

Item discrimination, expressed most simply as a correlation between the item and the total test score, is another statistic to consider. Correlations range from +1 to −1. The more a question distinguishes examinees with high scores from those with low scores, the higher the correlation. Test questions with low or negative correlations should be avoided. Items should relate to the total test score the same way for all examinees at the same score level; that is, individual items are expected to perform similarly for examinees with the same score.

Item response theory (IRT) offers an array of tools for the assembly of tests, such as test information functions and conditional standard error curves. The foundations of classical IRT[1] approaches can be found in Lord (1980). These procedures make strong assumptions about the data that may or may not be met. Violations of these assumptions may not adversely affect the utility of test assembly methods that are based on these assumptions. Recent research has indicated that tests built to IRT specifications and tests built to specifications based on more classical item statistics (i.e., *p* values and item-total correlations) tend to be of comparable quality (Davey & Hendrickson, 2010).

Yen and Fitzpatrick (2006) provided an up-to-date review of IRT applications. For more information on IRT, see Volume 1, Chapter 6, this handbook; for more discussion of item analysis issues, see Volume 1, Chapter 7, this handbook; and for details on test development, see Volume 1, Chapter 9, this handbook.

Scoring Rules

The method used to score the test is another psychometric consideration. It is essential in high-stakes settings that test takers understand the scoring rules because knowledge of these rules could affect how they approach the test.

The simplest method for scoring a test, and the most familiar one to most people, is the total number of correct items, also known as rights scoring. Under rights scoring, an examinee's incorrect and omitted responses are scored as wrong. Thus, the best way to maximize the test score is to answer every item, even if the examinee must guess on some items. If an examinee randomly guesses the answers to items about which the examinee knows nothing, then these guesses add noise to the measurement of knowledge or ability. Arguments against rights scoring and the accompanying directions exist on philosophical grounds as well. According to Thorndike (1971b), "many educators argue that to encourage guessing on the part of examinees is poor educational practice, since it fosters undesirable habits" (p. 59).

As a correction for random guessing, some tests use a method called *formula scoring* (Thurstone, 1919). A score derived from rights scoring, for example, only uses information from those items an examinee answered correctly. Formula scoring further distinguishes between items that the examinee attempts but gets incorrect and items the examinee chooses to omit.

IRT offers an array of scoring methods that may use more information in the data. An IRT score (an item pattern score) uses not just the number of correct responses but also information about the

[1]*Classical* is used on purpose here. Like classical test theory, unidimensional IRT is now a long-standing established practice, and as such, it is a classical procedure. There is a misconception that IRT and classical test theory are competitors. As Lord (1980) implied at the end of Chapter 1 of his IRT book, they are complementary, not competitive. There is also a tendency to call anything that is non-IRT classical. This reflects improper training more than anything else. Calling IRT classical challenges people to think about their use of language and puts them in touch with IRT classics like Lord.

particular items that were answered correctly to estimate an examinee's ability. This scoring leans heavily on the underlying IRT model. If the model is wrong, then the scoring is suspect. In addition, it is harder to explain how these items are scored.

Lord (1980) expressed reservations about the use of IRT models with tests administered under rights-scored instructions: "*In principle, item response theory can be applied to number-rights scored tests only if they are unspeeded*" (pp. 225–226). Despite Lord's concern that time limits on tests given under rights-scoring instructions would lead to responding that is no longer related to the construct of interest, the use of IRT in such situations is ubiquitous. The claim that IRT models are not appropriate for formula-score tests and should be used only with rights-scored tests is simply incorrect. This belief appears to be based on the presumption that use of IRT scoring is consonant with rights-scoring instructions, which is true only for the simplest IRT model. From the test-taker's perspective, a test that is administered under rights-scoring instructions but then scored in a different manner (e.g., pattern scored) may violate the rights of test takers to be informed about scoring criteria and general test-taking strategy (see Standard 8.2 of AERA et al., 1999).

There are disagreements about which scoring method is preferable (see Wendler & Walker, 2006, for a discussion of the pros and cons of rights and formula scoring). Note however, that $y' = R + O / k$, where R is the number of correct responses (i.e., the rights score), O is the number of omitted items, and k is the number of response options per item, is a formula score that is mathematically equivalent to the *standard correction for guessing* version. As with rights scoring, test takers receive no points for incorrect responses to questions. With the *partial credit for omitting* version of formula scoring, however, test takers receive a partial score if they choose not to answer a question for which they have not had the opportunity to learn the material. Under rights scoring, these examinees are forced to answer or they will receive a zero on the item. Thaler and Sunstein (2008) have demonstrated that choice architecture matters.

Although this chapter cannot provide definitive answers about which scoring method is best, it can suggest that a practitioner following the EVA principle would base scoring decisions on answers to such fundamental questions as, "Which instructions and scoring methods lead to optimal test-taker performance?" "Which instructions and scoring methods yield scores most highly reflective of test-taker ability (e.g., as indicated by high correlations with criteria of interest)?" "Which scoring method is most appropriate given the nature of the test and its intended uses?"

Scale Definition

Scores are the most visible and widely used products of a testing program. The score scale provides the framework for the interpretation of scores. The choice of score scale needs to be consonant with test specifications and test reliability and has implications for equating and validity as well as for test interpretation. The utility of a score scale depends on how well it supports the inferences attached to its scores and how well it facilitates meaningful interpretations and minimizes misinterpretations (Petersen, Kolen, & Hoover, 1989).

Kolen and Brennan (2004) and Kolen (2006) provided a broad perspective on scale definition. Included in Kolen (2006) are sections on procedures for incorporating normative informative information and score precision information into score scales (Dorans, 2002; Kolen, 1988; Kolen, Hansen, & Brennan, 1992; McCall, 1939; Zwick, Senturk, Wang, & Loomis, 2001).

What should a good scale look like? The scale should be well aligned with the intended uses of the scores. Dorans (2002) described desirable scale properties for broad-range tests, such as admissions tests, that have been introduced and used elsewhere. Dorans, Liang, and Puhan (2010) described desirable scale properties for narrow-range tests, such as certification exams, pointing out how they relate to the properties for scales of the broad-range tests. Both types of test should adhere to these two principles:

1. The number of scale units should not exceed the number of raw score points, which is usually a simple function of the number of items. Otherwise, unjustified differentiation of examinees may occur.

2. The score scale should be viewed as infrastructure that is likely to require repair.

The first property is based on the fundamental *one-item, one-scale point* property.[2] A gap occurs when a 1-point difference in raw scores translates to a multiple-point (2 or more) difference in scaled scores. A clump occurs when two or more raw scores convert to the same-scaled score. Gaps are worse than clumps. Gaps exaggerate differences, whereas clumps can hide them. To the extent that the score is unreliable, the exaggeration of differences is undesirable.

The second principle reminds us that the properties of score scales degrade over time and it is important to monitor these properties, a topic addressed in a later section. The perspectives listed here and elsewhere are only guidelines. The EVA practitioner should examine many possible score scales to determine which set of scores was most related to the construct of interest and which set of scores the score users would be most likely to interpret appropriately. The most important thing is to give serious thought to choice of scale and to monitor scale properties over time to ensure that the scale is still useful.

Measurement Conditions

Measurement conditions play an important role in assessment. This realization seems to have been forgotten over past 20 years with applications involving subjective scoring and the administration of tests tailored to the perceived ability of the examinee. The first two editions of *Educational Measurement* (Lindquist, 1951; Thorndike, 1971a) included chapters dedicated to the nature of measurement that emphasized the importance of attending to measurement conditions.

Lorge (1951) noted that "in scientific observations, whether direct or indirect, the conditions for measurement are carefully specified in terms of time, place, and circumstance. . . . The statement about observations, necessarily, must contain specification of condition (p. 536). Jones (1971) cited the same point and added,

> In general, science demands a reproducibility of observations. Whenever conditions and methods are identical, the observations should be identical (within range of measurement error) unless the object underwent some changes during observation or subsequent to it. The development of apparatus as an aid to measurement has promoted commonality of perceptual judgments among observers and has enhanced opportunities for agreement and reproducibility. (p. 338)

Context is a condition of administration. Despite the fact that context effects have been well documented (Brennan, 1992; Harris, 1991; Harris & Gao, 2003; Leary & Dorans, 1985), standard psychometric models ignore the context of assessment. For example, context effects can occur in equating when common items are administered in different locations in the test (e.g., a common item is item position 5 in one form and position 25 in the other form), under different testing conditions (e.g., paper and pencil vs. computer delivered), or when they are adjacent to different kinds of items in the two tests. It is essential to examine whether item parameters depend on context, especially in settings (e.g., adaptive testing) in which item properties must be robust to location and administration conditions (Kingston & Dorans, 1982). Kolen and Brennan (2004) listed common measurement characteristics and conditions as one of the necessary requirements of equating. Inequities in assessment may occur if the need for equivalent measurement conditions is ignored.

RELIABLE, INTERCHANGEABLE, AND VALID REPORTED SCORES

Scores have consequence in high-stakes settings. Hence they need to be reliable, trustworthy packets of information. Multiple test forms must produce scores that can be used interchangeably. The scores must be validated for their intended uses. Different chapters in this handbook delve into reliability (Volume 1, Chapter 2), equating (Volume 1, Chapter 11), and validity (Volume 1, Chapter 4) in detail.

[2]Alternative approaches, as well as alternative prescriptions, are described in Kolen and Brennan (2004, Chapter 8).

This chapter discusses the importance of these concepts to high-stakes large-volume test settings.

Reliability can be assessed in a variety of ways. Chapters by Thorndike (1951), Stanley (1971), Feldt and Brennan (1989), and Haertel (2006) in the four editions of *Educational Measurement* provided an interesting historical perspective on the evolution of thinking about reliability as well as a variety of formulae. Practitioners prefer specific prescription, not vague platitudes. Most treatments of reliability, however, shy away from addressing the question of level of reliability needed for specific purposes. As one early exception, Thorndike (1951, p. 609) cited Kelley (1927) who listed minimum levels of correlation needed for different purposes. With high-stakes tests, such prescriptions are helpful.

Uncertainty Reduction Justification for High Reliability

Dorans (2004) used the concept of uncertainty reduction to assess the degree of agreement between two scores, **X** and **Y**, and to describe reliability in more accessible terms. Reduction in uncertainty (RiU) is defined as follows:

$$RiU = 1 - \sqrt{1 - \rho_{xy}^2}, \tag{24.1}$$

where ρ_{xy}^2 is the squared correlation between scores on **X** and **Y**, Alternatively, the equation can be written,

$$100 \times RiU = 100 \times \left[\frac{\sigma_{YP} - \sigma_{YP}\sqrt{(1 - \rho_{xy}^2)}}{\sigma_{YP}} \right], \tag{24.2}$$

where σ_{YP} is the standard deviation of **Y** in population **P**. This standard deviation represents the total uncertainty associated with predicting a score on **Y** given no other information. The right-hand term in the numerator is the familiar standard error of prediction (SEP), which indicates the amount of uncertainty in **Y** that remains after **X** is used to predict **Y**. The difference between the two terms gives the amount by which uncertainty in predicting **Y** has been reduced by using **X** as a predictor. In this form, *RiU* is seen to be the percentage of uncertainty in **Y** that is eliminated with knowledge of **X**.

The reliability coefficient can be interpreted as the squared correlation between observed scores on **Y** and the true scores that **Y** estimates. Placing reliability in the context of RiU, Dorans and Walker (2007) argued that the reduction in uncertainty should be at least 50%, which corresponds to a reliability of .75, a value that is probably too low for high-stakes settings. Kelley's (1927) recommendation of a reliability of .94 is close to a reduction in uncertainty of 75%, whereas the typical reliability of .90 achieved by many high-stakes assessment corresponds to a reduction of uncertainty of 70%. (As Dorans & Walker, 2007, pp. 186–187, noted when discussing their Table 10.1, in terms of the measurement of sound, the sound of rustling leaves is akin to a reliability of .91, which is more appropriate for a high-stakes test.) These values, although vastly superior to complete uncertainty, still leave room for improvement.

Reliability is relatively easy to assess given the proper data collection design. When possible, a variety of estimates of the same type of reliability, for example, the squared correlation between observed and true score, should be computed, especially those based on different assumptions. If they agree, then one has a good grasp of the reliability of the scores. If they disagree, reasons for the differences should be sought.

Score Augmentation as a Palliative for Low Reliability of Scores

According to Standard 5.12 of the Standards for Educational and Psychological Testing (AERA et al., 1999), consumer needs are pushing test makers to try to extract more and more information from tests. The No Child Left Behind Act of 2001 (2002), for example, stated that students should receive diagnostic reports that allow teachers to address their specific academic needs. Subscores might be used in such diagnostic reports. Testing standards require that scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established (AERA et al., 1999). Thus, the appetite for additional information needs to be tempered by consideration of psychometric considerations. This section considers reliability.

Haberman (2008b) and Haberman, Sinharay, and Puhan (2009) suggested statistical methods that

help determine whether subscores add any value to what is already reported in the total score. Their methods are based on the relatively unrestrictive assumptions of classical test theory. In essence, these methods partially assess how reliable and unique the additional information is relative to the total score. Haberman (2008b) and Sinharay, Haberman, and Puhan (2007) demonstrated that a subscore is more likely to have added value when it has high reliability, the total test has relatively low reliability, and the subscore is sufficiently distinct from the total tests score and other subscores. Sinharay (2010) summarized results obtained from the analysis of operational data that are relevant to the value-added question. He also used data simulated from a multivariate item response theory model to assess the valued-added potential of subscores. He found that it is hard to find realistic conditions in which subscores have added value. Subscores need to have at least 20 items and be sufficiently distinct from each other. This finding contradicted the common practice of reporting subscores based on 10 or fewer questions.

Weighted averages of total scores and subscores (Haberman, 2008b; see also Wainer et al., 2001, for a slightly different approach), on the other hand, did exhibit some added value. For the most part, they had added value as long as the disattenuated correlation between the subscore and the total score was less than .95. Even with 10 items on the subscore, the weighted averages were primarily found to have added value when the disattenuated correlation was .85 or less. If users insist on subscores, these weighted averages that augment the subscore with additional information on the test may be psychometrically defensible when subscores are not.

When Can Scores be Used Interchangeably?

In addition to being highly reliable, scores on high-stakes assessments need to be interchangeable. Test score equating is essential to any testing program that continually produces new test forms and for which the uses of these test forms require that the meaning of the score scale be maintained over time. Even though different editions or forms of a test are designed to measure the same constructs and are usually built to the same test specifications or test blueprint, they almost always differ somewhat in their statistical properties. Test score equating strives to eliminate the effects on scores of these unintended differences in test form difficulty. Test score equating is performed in large-scale testing programs to be fair to examinees taking different test forms and to provide score users with scores that mean the same thing across test forms.

Reported scores are usually the most visible part of a testing program, even though they represent the endpoint of a large test production, administration, and scoring enterprise. An error in the equating function or score conversion function can affect the scores for all examinees. For this reason, the credibility of a testing organization is called into question over test-equating problems to a greater degree than when, for example, flawed test questions are discovered in operational tests. In high-stakes testing programs, the importance that score equating be done carefully and accurately cannot be overemphasized.

A *link* between scores on two tests is a transformation from a score on one to a score on the other. In the field of educational measurement, there is rich literature on test score equating. There are several important books on the topic of score linking and equating. Holland and Rubin (1982) included a set of conference papers covering a variety of test-equating topics. The most complete coverage of the entire field of test equating and linking is provided by Kolen and Brennan (2004). The book by von Davier, Holland, and Thayer (2004) introduced several new ideas of general use in equating, although its focus is on kernel equating. The two book-length reports from the National Research Council, *Uncommon Measures: Equivalence and Linkage Among Educational Tests* by Feuer, Holland, Green, Bertenthal, and Hemphill (1999) and *Embedding Questions: The Pursuit of a Common Measure in Uncommon Tests* by Koretz, Bertenthal, and Green (1999), are accessible summaries of informed, professional judgment about the issues involved in linking scores on different educational tests. Livingston's (2004) training manual for those who will actually do score equating provided his perspective on many of the major issues and procedures encountered in practice. Dorans, Pommerich, and Holland (2007) provided an integrated look at various types of score linking, including score equating, concordance, and vertical linking.

A wealth of material has appeared in the four editions of *Educational Measurement*. Flanagan (1951), in the first edition, covered some of the test-equating techniques that were available at that time. He also discussed issues and problems that are still relevant, which shows how pervasive they are. Angoff (1971), in the second edition, provided a comprehensive introduction to scales, norms, and test equating. Petersen, Kolen, and Hoover (1989) introduced new material developed since the Angoff chapter. Holland and Dorans (2006) provided a historic background for test score linking. In addition to test equating, Holland and Dorans (2006) discussed other ways that scores on different tests are connected or linked together, presenting a framework that served as the foundation for the Dorans et al. (2007) book on linking. Dorans, Moses, and Eignor (2011) described best practices for equating from their combined perspectives. Volume 1, Chapter 11, this handbook, also considers equating and scaling in some detail.

Five requirements are widely viewed as necessary for a linking to be an equating:

1. The Equal Construct Requirement: The two tests should both be measures of the same construct (latent trait, skill, ability).
2. The Equal Reliability Requirement: The two tests should have the same reliability.
3. The Symmetry Requirement: The equating transformation for mapping the scores of **Y** to those of **X** should be the *inverse* of the equating transformation for mapping the scores of **X** to those of **Y**.
4. The Equity Requirement: It should be a matter of indifference to an examinee to be tested by either of the tests that have been equated.
5. The Population Invariance Requirement: The equating function used to link the scores of **X** and **Y** should be the same regardless of the choice of (sub)population from which it is derived.

Both formal and informal statements of subsets of these five requirements have appeared in a variety of earlier sources, including Angoff (1971), Holland and Dorans (2006), Kolen and Brennan (2004), Lord (1950, 1980), and Petersen et al. (1989). These five requirements have value as criteria for evaluating whether or not two tests can be, or have been, successfully equated. Dorans and Holland (2000) added the second requirement to the set of requirements (1, 3, 4, and 5) provided by Lord (1980).

Brennan (2010) examined the role of reliability in the equity requirement. He reported that for curvilinear equating, both high reliability, as advocated by Dorans and Walker (2007), and equal reliability, suggested by Dorans and Holland (2000), are important to equating. He demonstrated that relatively high reliability is necessary for approximating the requirements of equity, which is universally viewed as a desirable property for equating. Support for the importance can be gleaned from Holland and Hoskens (2003), who have demonstrated the possibility of subpopulation dependence of relationships whenever reliability is low (cf. Requirement 5).

Dorans et al. (2011) have described several threats to the quality equating. The *amount* of data collected (sample size) has a substantial effect on the usefulness of the resulting equating. Because it is desirable for the statistical uncertainty associated with test equating to be much smaller than the other sources of variation in test results, it is important that the results of test equating be based on samples that are large enough to ensure this. Ideally, the data should come from a large representative sample of motivated examinees that is divided in half either randomly or randomly within strata to achieve equivalent groups. When an anchor test is used, the items are evaluated via differential item functioning (DIF) procedures to see whether they perform the same way in both the old and new form samples. The anchor test needs to be highly correlated with the total tests. All items on both tests are evaluated to see whether they are performing as expected. It is valuable to equate with several different models, including both linear and equipercentile models. The use of multiple methods provides one with a sense of the stability of the equating. Dorans et al. (2011) have discussed this in more detail.

An equating should be checked for its reasonableness by comparing the raw-to-scale conversion for the new form with those that have been obtained in the past. In testing programs with large volumes

and relatively stable populations, it may be reasonable to expect that the new form sample will have a similar scale score distribution to that obtained at the same time the prior year. If the test is used to certify mastery, then the pass rates should be relatively stable from year to year, although not necessarily across administrations within a year. All equated tests should exhibit population invariance across subpopulations (Dorans & Holland, 2000; Dorans & Liu, 2009).

As mentioned, IRT makes strong item-level assumptions. When these assumptions are met, they allow psychometricians to compute score-equating functions before administration (a process known as preequating). If the assumptions are not met, however, their violation may adversely affect the scores assigned to test takers. Due diligence indicates that preequating should be avoided unless ample evidence can be garnered to justify its use.

How Do We Know That Score Uses Are Valid?

High reliability is a prerequisite for validity and, as noted, the score interchangeability sought by score equating. Both are forms of internal validity. Both can be assessed empirically. In large-scale testing settings in which stakes are high, this empirical validation is essential.

Assume that it has been demonstrated that scores from different editions of a test are interchangeable by being highly reliable and that they are exhibiting population invariance linking functions. The difficult task of external validation remains. Messick (1989, 1994) departed from previous treatments of validity by Cureton (1951) and Cronbach (1971) by describing validity as being about appropriate evidential and consequential test interpretation and test use. Kane (2006) treated validity as an argument, building on the Toulmin (1958) model of inference.

This chapter focuses on empirical verification of assessments—that is, statements about an assessment must be subject to falsification (Popper, 1959) as part of the validation process. Claims about reliability and equatability can be assessed empirically. Claims about valid score use, and the validity of inferences based on those uses, are more difficult to assess because many sources of irrelevant variance may affect empirical evaluations of these uses. Validation of score uses is a Sisyphean task. Due diligence requires that the validation effort must occur. In the chapter titled "Validation," Section 7, Fallacies in Validity Arguments, Kane (2006) addressed several of the ways in which validation can be subverted, including begging the question of consequences, overgeneralization or spurious prediction, surrogation, and reification.

Threats to the Maintenance of Score Meaning Over Time

Validation is an unending process. So is scale maintenance. Once a high-stakes test is producing reliable, interchangeable scores that serve a useful function, there might be a temptation to presume that the job is done. It is essential, however to monitor the quality of the scores and other products associated with the testing program. As noted, a data-based orientation that emphasizes sound data collection designs that strive to achieve adequate pool replenishment with items that meet fairness criteria and the production of interchangeable scores through equating is necessary. In addition, it is important to demonstrate that the test forms continue to measure the same construct and that scores continue to have their intended meaning and can support their intended uses. Haberman and Dorans (2011) considered one of the often-overlooked threats to the quality of the testing operation—drift in the meaning of scores.

Scale drift is defined as a change in the interpretation that can be validly attached to scores on the score scale. Among its potential sources are population shifts, inconsistent or poorly defined test-construction practices, estimation error associated with small samples of examinees, and inadequate anchor tests. In addition, a sequence of sound equatings can produce nonrandom drift.

Practitioners routinely use two indicators to provide evidence of scale stability: average test scores and the relationship between raw and scaled scores. If both indicators remain fairly constant, practitioners take this as evidence that the meaning of the score scale remains intact. Conversely, whenever score distributions shift in one direction over time, there is a tendency to wonder whether the score

scale has remained intact. A shift in score distributions does not necessarily mean scale drift, however.

With respect to the relationship between raw and scaled scores, when tests are built successfully to rigorous specifications, it is reasonable to expect that the conversion that takes a raw score onto a score scale is the same for all versions of a test form. If factors such as item availability or staff inexperience interfere with strict adherence to specification, some variability may be seen in raw to scale conversions. This variability may or may not indicate evidence of scale drift.

The size and composition of the equating sample can contribute to scale drift as well. For example, when the equating sample is not representative of the population, systematic error can be induced, especially in the absence of an anchor. Even with representative but finite samples of examinees, a random error of equating will appear in the estimation process. The standard deviation of this error is approximately proportional to the reciprocal of the square root of the sample size. This random error introduces random noise into the equating process.

Accumulation of random error over many successive administrations can produce random scale drift. This accumulation of error is the bane of continuous testing. Haberman (2010) noted that multiplying the number of administrations by a factor M is typically accompanied by a decrease in the typical sample size by 1/M; these two factors together lead to an increasing random equating error by a factor of M. Ignoring this important relationship among total volume, the number of administrations, and scale drift can lead to practices that rapidly undermine the scale of a test.

Random scale drift can have effects similar to those of systematic scale drift. In typical data collection, results equated within a small time interval are much more similar to each other than they are to results derived in the distant past. When placed in the context of the nonrandom error associated with a given test form, the conditional error of measurement (or its average, the standard error of measurement), the amount of drift induced by any and all sources can seem small. For example, when the standard error of measurement for a test on a 200- to 800-point scale is 40 points, then a drift of 10 points might seem small by comparison. The comparison is inappropriate, however. The former is random error, which means that on average across all people, it is expected to be close to 0. Drift, on the other hand, is in one direction. Across all people, it is 10 points rather than 0 points. This distinction is important. Another issue to consider is the affect of drift on the group average. Group averages are more reliable than the individual scores on which they are based because errors of measurement across individuals can cancel out. Consequently, a scale drift of 10 points has a greater impact on the group statistics than it has on any individual. In sum, drift that may look small relative to the standard error of measurement can look quite large relative to the random error associated with the average score.

Scale drift is a shift in the meaning of score scale that alters the interpretation attached to score points along the scale. Trends in score distributions and inconsistent raw-to-scale conversions do not necessarily indicate scale drift. In addition, continuous testing with small volumes at each administration significantly shortens the life span of a meaningful score scale. From the EVA perspective, the testing professional would closely monitor various aspects of the score scale and would look far beyond the obvious average score and raw-to-scale relationships.

PROPER DATA COLLECTION

EVA requires data to evaluate the appropriateness of all decisions related to test design. Proper data collection is most critical to the successful maintenance and evolution of large-volume testing programs. The right data allow the test maker to assess the quality of new items and new tests, to equate test forms, to assess the psychometric properties of tests, to maintain score scales, and to make improvements to the measure.

The design of a testing program should include allowances for collecting the needed data to evaluate the quality of that program, a program that is expected to produce comparable tests that are properly administered and scored to produce interchangeable scores. Test specifications provide information about the composition of a test—number and type of items, content covered, and so on—designed for particular purpose. In the same way,

test administration plans, which specify how many and which forms are given during a testing cycle (typically a year), should reflect the need for adequate data to link test forms and to test their psychometric properties. In principle, both test specifications and administration plans should support assessment quality. In practice, market constraints and other realities sometimes may conflict with these principles. For example, customers increasingly request short, computerized tests delivered on demand, which does not bode well for measurement quality.

In general, data collection should be guided by three principles: representative samples, adequate sample sizes, and secure data. Examinee samples should represent the population of interest. Likewise the items should mirror specifications. The samples should be large enough to support the intended use of the data. This means enough examinees are employed to provide stable estimates of item and test statistics and that enough items are employed to provide reliable measures of examinee attributes. Short tests and tests administered to small numbers of candidates may possess data collection deficiencies that cannot be surmounted. Data need to be collected under secure conditions. Otherwise, the integrity of the testing operation may be subject to subversion. Exposure of material needs to be minimized. This necessity suggests large-scale administrations, not the numerous small-scale administrations that seem to be driven more by technological innovation than by test quality.

An Adequate Item Inventory

Testing programs with multiple administrations in a given time period, especially when scores are used for high-stakes decisions, must determine the optimal number of new test forms to be created along with the acceptable level of form reuse. New test forms are needed for test security reasons. The development and use of new test forms has a number of disadvantages, however. The cost of item development and tryout can be prohibitive, especially for some types of items. As the number of individual versions of the test increases, aligning new forms with test specifications becomes more difficult, if not impossible. In addition, new forms must be equated.

Once the number of new test forms required for the test inventory is established, the number of test items needed to support the development of new test forms can be determined. Although the most pressing need for new items comes from the number of new test forms required in the test inventory, other requirements also affect the number of new items available for development needs.

Item pretesting. In high-stakes settings, some type of item pretesting should be done before the item appears as a scored item on the test form. The goal of pretesting is to ensure that items are functioning as expected. Information about item difficulty and item discrimination can be determined if the tryout sample is large enough and adequately represents the test-taking population. Pretesting can identify flaws in an item (e.g., a negative relationship with other items or a very high omission rate). With constructed-response items, pretesting can facilitate refinement of the scoring rubric. Pretests also allow the determination of the statistical equivalence of items generated from the same task model. The statistical information obtained during pretesting helps to ensure the construction of multiple parallel forms according to specifications. Wendler and Walker (2006) discussed three basic models for pretesting items as well as the pros and cons associated with each.

Postadministration item analysis. After the administration, item analysis similar to that performed during pretesting will help to ensure that the test and the items are performing properly. In particular, the item difficulty distribution for the test or the test characteristic curve can be compared with the statistical specifications. Item–total correlations will help identify items not adequately measuring the construct of interest. Section and test reliabilities can be computed from the data. When constructed-response items are included in the test, it would be desirable to have at least a portion of the papers for each item scored by two raters so that rater reliability can be computed.

Verifying test and item quality at the postadministration stage requires a fair amount of data. Many large-scale assessments with multiple test forms have large administrations, with a minimum of a few thousand examinees. Not all large-scale, high-stakes

assessments have large volumes, however. In those cases, some measures must be taken to ensure adequate precision of the obtained statistics. Even with a moderately large sample, precision of item statistics may not be adequate. On the other hand, smaller sample sizes may suffice for some test-level information, such as reliability (Walker, Zhang, & Zeller, 2002).

Differential item functioning. DIF is a form of secondary analysis employed by testing programs with high-stakes uses (Holland & Wainer, 1993). The focus of a DIF analysis is to ensure that items function in the same manner across different subgroups. Data requirements for DIF, as with many item-level analyses, such as those associated with IRT, are greater because the focus is on group-by-item interactions rather than simply item properties in a group. Penfield and Camilli (2007) provided an extensive review of DIF procedures. Mapuranga, Dorans, and Middleton (2008) summarized and classified DIF methods and procedures that have appeared since Holland and Wainer (1993) and assessed their appropriateness for practical use. Widely used DIF methods are evaluated alongside newer methods for completeness, clarity, and comparability.

DIF techniques require adequate sample sizes (in the several hundreds) not just for the total test-taking population but for subgroups as well. For many programs, gender DIF may be easily tested, as there may be fairly equal numbers of males and females in the test-taking population. For other DIF evaluations (e.g., ethnic groups), some subgroups may be a relatively small proportion of the total population. In these cases, special studies may be necessary to test for DIF. Empirical verification of item fairness requires adequate data. Substitution of assumption for data in DIF begs the need for empirical verification.

Data Collection for Test Score Equating

Testing programs engaged in high-stakes assessments should have well-designed score-equating plans and well-aligned score scales that increase the likelihood that scores on different forms can be used interchangeably. A score-equating plan that links a new form to multiple old forms is preferable to a plan with a link to a single old form.

Equating with data from an operational test administration. Data collection is one of the most important aspects of best practices in equating. Holland and Dorans (2006) have provided an extensive discussion of the pros and cons of designs associated with score linking. Strictly from a score-equating standpoint, new forms ideally would be administered alongside old forms to equivalent groups from the same motivated population of examinees. Additionally, each group would receive a common anchor block. In practice, this ideal is rarely achieved, primarily because of concerns about keeping test items secure. Often, then, test score equating involves linking scores form tests given to two groups that differ in ability. In these circumstances, score-equating procedures need to control for this differential examinee ability. In examining the distributions of the resulting scores, two confounded factors can complicate the interpretation of results. One is the relative *difficulty* of the two tests and the other is the relative *ability* of the two groups of examinees on these tests. This unknown difference in ability needs to be eliminated so that the equating process can adjust for differences in test difficulty. In score equating, the goal is to adjust for differences in test characteristics, while controlling for any differences in examinee ability that might complicate this adjustment.

There are two distinct means of addressing the separation of test difficulty and differential examinee ability. The cleanest and preferred approach uses a common population of examinees, and the other approach uses an anchor measure. Differential examinee ability is explicitly controlled when the same examinees take both tests. More often, equivalent groups are nonoverlapping samples of examinees from a common population. When the samples are nonequivalent, performance on a set of common items or an anchor measure is used to quantify the ability differences between two samples of examinees. The use of an anchor measure permits the use of more flexible data collection designs than the use of equivalent examinees. Methods that use anchor measures, however, require making various assumptions that are not needed when the examinees are

either the same or from equivalent samples. When ability differences exist, different statistical adjustments for these differences often produce different results.

The role of the anchor test is to quantify the differences in ability between samples that affect their performance on the two tests to be equated. An anchor should measure the same thing as the total tests and produce adequately reliable scores. The statistical role of the anchor is to remove bias in the equating function that would occur if the groups were presumed to be equivalent as well as to increase precision in the estimation of the equating function. An anchor can be expected to do a good job of removing any bias due to the nonequivalence of the two samples, provided the correlations between anchor and tests scores are high. Results by Dorans (2000) suggested that a correlation of .87 be considered a minimum target for the correlation between the anchor and total tests. If the anchor does not adequately represent both tests, even a high anchor–total correlation will not guarantee a successful equating (e.g., see Cook & Petersen, 1987).

The anchor test design is subject to more sources of drift than a well-executed equivalent group design. Much can go wrong with this design. The groups may be too far apart in ability. The anchor may not have a strong enough correlation with the total tests to compensate for the lack of equivalence between the samples for the old and new forms. The anchor may possess different content than the tests. All of these factors can result in raw-to-scale conversions that vary as a function of the equating method. These variations can induce scale drift, and the set of anchor–test influences may be the largest contributing factor to scale drift.

Mixed-format tests, those containing both selected-response and constructed-response items, present special problems for equating. Because the two item types may measure somewhat different constructs, the use of selected-response items only in the anchor may not yield accurate results (Kim, Walker, & McHale, 2010a; Walker & Kim, 2012). Use of constructed-response items in the anchor presents a problem because scoring standards change from administration to administration, even for mathematics items (Fitzpatrick, Ercikan, Yen, & Ferrara, 1998). Ignoring these shifting standards can lead to bias; correcting for them requires special effort in scoring and linking (Kim et al., 2010a, 2010b; Tate, 1999, 2000).

Preequating: Estimating the scaling function before test administration. *Preequating* offers an alternative to equating a test after administration of the test. The most common method of preequating, or more accurately *precalibration*, involves calibrating a pool of test items using IRT so that a new test may be placed on the same scale as the old test(s) before the new test is ever administered (Lord, 1980). This method has several advantages. Among other things, this method facilitates the small test administrations common with continuous testing programs that make postequating problematic. Examinees can receive scores immediately after testing rather than having to wait several weeks for score equating to take place. Furthermore, data for item calibration can be collected over a long period of time without disrupting the score-reporting schedule.

The quality of item precalibration relies on the quality of the pretest data upon which the calibration is based. Some research suggests that the scoring tables obtained from precalibration are quite similar to those obtained from IRT calibration based on operational administration of the intact test, although the item parameter estimates may be quite different (Tong, Wu, & Xu, 2008), even when the pretest data come from a population separate from the testing population, such that measurement equivalence does not hold across the two groups (Domaleski, 2006). More sobering evidence suggests that equivalence of precalibration to postcalibration results requires that the pretest samples come from the testing population and that they be almost as large as the entire population (Taherbhai & Young, 2004). Precalibration of tests containing constructed-response items can also be problematic (Tate, 1999). At the very least, the mixed research results suggest that any preequating plan be preceded by research to determine whether it will yield satisfactory results for the testing program under consideration. It may well be the case that the longer the test, the more robust preequating may be to violations of assumptions.

Quality Assurance at Each Administration

At each administration, several steps should be taken to improve the quality of the data obtained via the data collection. These data processing steps deal with sample selection and item screening. Checking assumptions is a key aspect of EVA.

Examinee data. Tests are designed with a target population in mind, and often only members of that target population are included in many psychometric analysis. For example, admissions tests are used to gather standardized information about candidates who plan to enter a college or university. The SAT excludes individuals who are not juniors or seniors in high school from its equating samples because they are not considered members of the target population. This is done to remove any potential influence of these individuals on the equating results. Examinees who perform well below chance expectation on the test are sometimes excluded, although many of these examinees may have already been excluded if they were not part of the target group. Inclusion of repeat test takers in the equating sample also may affect results (Kim & Walker, 2012; Puhan, 2009). A controversial issue involves whether nonnative speakers of the language of the test should be included or excluded from DIF analyses (Sinharay, Liang, & Dorans, 2010).

Haberman (2008a) discussed how outliers can assist in quality assurance. Their frequency can suggest problems with form codes, scanning accuracy, ability of examinees to enter responses as they intend, or exposure of items.

For nonequivalent groups anchor test (NEAT) designs, statistical outlier analysis should be used to identify those examinees whose anchor test performance is substantially different from their performance on the operational test, for which the scores are so different that both scores cannot be plausible indicators of the examinee's ability. Removing these examinees from analysis samples prevents their unlikely performance from having an effect on the results of the analysis.

Anchor item. For equating involving anchor items, the statistical properties of the anchor items should be evaluated to ensure that they have not changed from the one test administration to the other. DIF methods (Holland & Wainer, 1993) may be used to compare the performance of the common items with the two test administrations treated as the reference and focal groups. The total score on the common items would serve as the matching criterion. Simple plots of item difficulty values and other statistics may be used to detect changes in items. Common items embedded with the old and new tests are susceptible to context effects because they are surrounded by different sets of items in the two tests.

Model fit and invariance. Evaluation of model fit is central to an EVA. Model fit needs to be assessed. One way to do this is to test claims made by the model. Much of the power of IRT comes from its claim that item parameter estimates are invariant. Computer-adaptive testing would not be possible without this restrictive assumption. Yet, this assumption is rarely tested because operational data collections do not generate enough data to test it. Given the proper data, goodness-of-fit procedures for common IRT models do exist (Haberman, 2009; Haberman, Holland, & Sinharay, 2006).

Different models make different assumptions and may lead to different outcomes. The elegant simulation study by Sinharay and Holland (2010) demonstrated this point quite well. The authors simulated data for a NEAT design in three ways. One way was based on the assumptions of poststratification equating, another was based on the assumptions of chain equating, and the third was based on the assumptions of IRT true score equating. The three ways of generating data produced three different winners; each method worked best with the data generated in accord with its assumptions.

The study also produced one big loser: The myth that simulations uncover the truth. In fact, a simulation study that generates data consistent with one model will subsequently demonstrate the superiority of that model over its competitors in these simulated data sets. Tucker, Koopman, and Linn (1969) noted long ago that simulations need to include model misfit components to make the simulated data more realistic. The Tucker–Koopman–Linn approach provided a mechanism for introducing noise into made-up data to make the data more realistic. These more

realistic data could be used to evaluate different models. The problem of inferring substantive conclusions from simulated data still exists. Test makers should exercise appropriate caution when using a simulation to justify their use of a particular model.

One empirical check of model fit involves score-equating functions. Score-equating functions should be subpopulation invariant. Holland and Dorans (2006) maintained that assessing the degree of population invariance is an empirical means of determining whether or not a score linking qualifies as an equating, with the concomitant inference of score interchangeability. Although tests like the SAT have been assessing this assumption for some time (Dorans & Liu, 2009), many testing programs do not. Fear of detecting a lack of invariance that would force a testing program to ask unsettling but necessary questions about the quality of their test blueprints may account for some of this reluctance.

Continuous Improvement

Alterations of blueprints for assessment are often done in response to external pressures. The redesign of the SAT, which led to the adoption of a holistically scored essay and a writing section in 2005, was a response to the perceived needs of major test score users. Since that change was implemented, SAT tests have been built to a set of specifications that have remained constant. Research was conducted (Liu & Walker, 2007) to see whether the content changes altered the meaning of the scores. The changes to the SAT were in many respects reactive. Other models of change have been proposed. Brennan (2007) made a distinction between two types of test changes: gradual versus abrupt. Liu and Dorans (2010) made a different kind of distinction: inadvertent deviations from specifications versus planned modifications of specifications. The gradual and abrupt changes described by Brennan may be viewed as planned modifications that differ in degree.

Liu and Dorans (2010) demonstrated how score equity assessment (SEA) indices from subpopulation invariance studies can be used to quantify the degree to which planned modifications in specifications affect score comparability. Instead of assuming that any change in specifications or measurement conditions automatically induces a change in score meaning, SEA can be used to assess how much of an affect has occurred. In addition, SEA can be used as a tool in the modification process. If maintaining the old scales or minimizing the impact on the scales is a primary concern for testing programs that undergo modifications in test specifications, consideration should be given to making a number of small changes in sequence over time, instead of introducing many changes all at once. SEA indices can evaluate the impact of these small steps on score equatability, indicating which steps or combination of steps could be taken without jeopardizing comparability of score meaning.

SEA can be used to help guide a program of continuous improvement that does not appreciably alter the meaning of scores over a prescribed period of time. Large-volume assessments that are used in high-stakes settings could engage in a program of continuous improvement that is designed to increase score validity while not appreciably altering test score meaning over the short run.

CONCLUSION

Large-scale assessments involving multiple forms that produce scores that are used for high-stakes settings are complex. Many of the chapters in this handbook pertain to these kinds of tests. This chapter has pointed out where these connections lay. The chapter has addressed specification of the purposes of the assessment, the test-taker population, and the products of the assessment, and it has discussed the need for well-articulated content and statistical specifications as well as the importance of controlling the influence of measurement conditions on scores and hence inferences associated with them.

The principle of empirical verification has played a role throughout this chapter. It manifests itself in the section that emphasizes the importance of highly reliable, valid scores that maintain their meaning over test editions. Proper data collection directed to item pool replenishment, score equating, quality assurance, and program improvement is essential for the empirical verification of score meaning.

Restricting this chapter to large-volume, high-stakes tests has circumscribed assessment to a

relatively specialized setting for which ample data exist to test models and to validate aspects of the assessments. Professionals who work in settings in which high-stakes tests are administered are well aware that much of the testing taught in academic settings is taught from a perspective that views examinees as passive objects in the measurement process (Dorans, 2012). Consequently, psychometric models based on the passive examinee view may have failed to achieve the potential they promised. These failures have been most evident in settings in which the stakes associated with testing were high and adequate data existed to assess whether the promises of assessment were empirically verified.

The tripods of sound high-stakes assessment are high reliability, interchangeability of scores from different test assessments, and validity. Reliability is the easiest to achieve, and guidelines exist for how to determine how much reliability is enough (Dorans & Walker, 2007). Equating is also achievable in principle, and procedures for assessing how well it is met exist. High reliability, according to Brennan (2010), helps achieve equatability, which can be assessed via population invariance studies as well (Dorans & Holland, 2000). Both high reliability and high equatability improve the chances of producing useful internally valid score inferences. Intended uses imply external validity claims that are best tested directly.

Mislevy et al. (2002, 2003) recommended an integrated framework (ECD) to assessment. Applications of this framework have focused on integrating the test development process with the conceptual framework underlying the testing process. Less attention seems to have been given to empirical verification of the frameworks that are used to guide the test-assembly process. ECD applications need to focus on collecting data in a manner that will enable the user to empirically evaluate and improve the test construction framework.

What works on paper does not always work in practice. Empirical verification is essential to sound assessment. EVA is close to Ayer's (1936) notion of empirical verification in that empirical data can be used to verify the plausibility of empirical propositions that are posited by a model or framework. Specific predictions are made about observables and data are collected and assessed. Empirical data are central to assessing reliability, equitability, and validity.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Ayer, A. J. (1936). *Language, truth and logic*. London, England: Gollancz.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, *5*, 225–264. doi:10.1207/s15324818ame0503_4

Brennan, R. L. (2007). Tests and transitions. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 161–175). New York, NY: Springer-Verlag. doi:10.1007/978-0-387-49771-6_9

Brennan, R. L. (2010). *First-order and second-order equity in equating.* (CASMA Research Report No. 30). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225–244. doi:10.1177/014662168701100302

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Davey, T., & Hendrickson, A. (2010, May). *Classical versus IRT statistical test specifications for building test forms.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO. Retrieved from http://professionals.collegeboard.com/profdownload/pdf/Davey_Hendrickson_NCME_2010_Test_specs.pdf

Domaleski, C. S. (2006). *Exploring the efficacy of pre-equating a large scale criterion-referenced assessment with respect to measurement equivalence* (Doctoral dissertation, Georgia State University). Retrieved from http://digitalarchive.gsu.edu/cgi/viewcontent.cgi?article=1002&context=eps_diss

Dorans, N. J. (2000). *Distinctions among classes of linkages* (College Board Research Note RN-11). New York, NY: The College Board.

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: How and why. *Journal of Educational Measurement*, *39*, 59–84. doi:10.1111/j.1745-3984.2002.tb01135.x

Dorans, N. J. (2004). Equating, concordance and expectation. *Applied Psychological Measurement, 28*, 227–246. doi:10.1177/0146621604265031

Dorans, N. J. (2011, April). *The contestant perspective on taking tests: Emanations from the statue within.* Invited 2010 Career Award Address at the 2011 annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice, 31*(4), 20–37.

Dorans, N. J., & Holland, P. J. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281–306. doi:10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., Liang, L., & Puhan, G. (2010). *Aligning scales of certification tests* (ETS Research Report No. RR-10-07). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Liu, J. (2009). *Score equity assessment: Development of a prototype analysis using SAT Mathematics test data across several administrations* (ETS Research Report No. RR-09-08). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Moses, T., & Eignor, D. E. (2011). Equating test scores: Towards best practices. In A. A. von Davier (Ed.), *Statistical models for scaling, equating and linking* (pp. 21–42). New York, NY: Springer-Verlag.

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales.* New York, NY: Springer-Verlag.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 179–198). New York, NY: Springer-Verlag. doi:10.1007/978-0-387-49771-6_10

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York, NY: Macmillan.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington, DC: National Academies Press.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11*, 195–208. doi:10.1207/s15324818ame1102_5

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). Washington, DC: American Council on Education.

Haberman, S. J. (2008a). *Outliers in assessment* (ETS Research Report No. RR-08-41). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2008b). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*, 204–229. doi:10.3102/1076998607302636

Haberman, S. J. (2009). *Use of generalized residuals to examine goodness of fit of item response models* (ETS Research Report No. RR-09-15). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2010). *Limits on the accuracy of linking* (ETS Research Report No. RR-10-22). Princeton, NJ: Educational Testing Service.

Haberman, S. J., & Dorans, N. J. (2011). *Sources of scale inconsistency* (ETS Research Report No. RR-11-10). Princeton, NJ: Educational Testing Service.

Haberman, S. J., Holland, P. W., & Sinharay, S. (2006). *Limits on log cross-product ratios for item response models* (ETS Research Report No. RR-06-10). Princeton, NJ: Educational Testing Service.

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*, 79–95. doi:10.1348/000711007X248875

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education/Praeger.

Harris, D. J. (1991). Effects of passage and item scrambling on equating relationships. *Applied Psychological Measurement*, *15*, 247–256. doi:10.1177/01466216910150030 4

Harris, D. J., & Gao, X. (2003, April). *A conceptual synthesis of context effect.* In *Context effects: Implications for pretesting and CBT*. Symposium conducted at the 2003 annual meeting of the American Educational Research Association, Chicago, IL.

Hendrickson, A., Huff, K., & Leucht, R. (2009, April). *Claims, evidence and achievement level descriptors as a foundation for item design and test specifications*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA. Retrieved from http://professionals.collegeboard.com/profdownload/pdf/Hendrickson_ECD_Item_Test_Specs_NCME09.pdf

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational*

measurement (4th ed., pp. 187–220). Westport, CT: American Council on Education/Praeger.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika, 68*, 123–149. doi:10.1007/BF02296657

Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New York, NY: Academic Press.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Jones, L. V. (1971). The nature of measurement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 335–355). Washington, DC: American Council on Education.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York, NY: World Book.

Kim, S., & Walker, M. E. (2012). Investigating repeater effects on chained equipercentile equating with common anchor items. *Applied Measurement in Education, 25*, 41–57. doi:10.1080/08957347.2012.635481

Kim, S., Walker, M. E., & McHale, F. (2010a). Comparisons among designs for equating mixed-format tests in large scale assessments. *Journal of Educational Measurement, 47*, 36–53. doi:10.1111/j.1745-3984.2009.00098.x

Kim, S., Walker, M. E., & McHale, F. (2010b). Investigating the effectiveness of equating designs for constructed response tests in large scale assessments. *Journal of Educational Measurement, 47*, 186–201. doi:10.1111/j.1745-3984.2010.00108.x

Kingston, N. M., & Dorans, N. J. (1982). *The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory* (ETS Research Report No. RR-82-22). Princeton, NJ: Educational Testing Service.

Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement*, *25*, 97–110. doi:10.1111/j.1745-3984.1988.tb00295.x

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Westport, CT: American Council on Education/Praeger.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Kolen, M. J., Hansen, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scaled scores. *Journal of Educational Measurement, 29*, 285–307. doi:10.1111/j.1745-3984.1992.tb00378.x

Koretz, D. M., Bertenthal, M. W., & Green, B. F. (Eds.). (1999). *Embedding questions: The pursuit of a common measure in uncommon tests* (Report of the Committee on Embedding Common Test Items in State and District Assessments, National Research Council). Washington, DC: National Academies Press.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: An historical perspective on an immediate concern. *Review of Educational Research, 55*, 387–413.

Lindquist, E. E. (1951). *Educational measurement*. Washington, DC: American Council on Education.

Liu, J., & Walker, M. E. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109–134). New York, NY: Springer-Verlag. doi:10.1007/978-0-387-49771-6_7

Liu. J., & Dorans, N. J. (2010). *Using score equity assessment to measure construct continuity when tests deviate from specifications or test specifications change* (ETS Statistical Report No. SR-10-41). Princeton, NJ: Educational Testing Service.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1950). *Notes on comparable scales for test scores* (ETS RB-50-48). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lorge, I. (1951). The fundamental nature of measurement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 533–559). Washington, DC: American Council on Education.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). *A review of recent developments in differential item functioning* (ETS Research Report No. RR-08-43). Princeton, NJ: Educational Testing Service.

McCall, W. A. (1939). *Measurement*. New York, NY: Macmillan.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (ETS Research Report No. RR-03-16). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing.

Educational Measurement: Issues and Practice, *25*(4), 6–20. doi:10.1111/j.1745-3992.2006.00075.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice* (pp. 97–128). Hillsdale, NJ: Erlbaum.

No Child Left Behind Act of 2001, 20 U.S.C. § 6319 (2002).

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125–167). Amsterdam, the Netherlands: Elsevier.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York, NY: Macmillan.

Popper, K. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.

Puhan, G. (2009). *What effect does inclusion or exclusion of repeaters have on test equating?* (ETS Research Report No. RR-09-19). Princeton, NJ: Educational Testing Service.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*, 150–174. doi:10.1111/j.1745-3984.2010.00106.x

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21–28. doi:10.1111/j.1745-3992.2007.00105.x

Sinharay, S., & Holland, P. W. (2010). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement*, *47*, 261–285. doi:10.1111/j.1745-3984.2010.00113.x

Sinharay, S., Liang, L., & Dorans, N. J. (2010). First language of examinees and empirical assessment of fairness. *ETS Research Spotlight* (No. 3), 10–18.

Snow, E., Fulkerson, D., Feng, M., Nichols, P., Mislevy, R., & Haertel, G. (2010). *Leveraging evidence-centered design in large-scale test development* (Large-Scale Assessment Technical Report No. 4). Menlo Park, CA: SRI International. Retrieved from http://ecd.sri.com/downloads/ECD_TR4_Leveraging_ECD_FL.pdf

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356–442). Washington, DC: American Council on Education.

Taherbhai, H. M., & Young, M. J. (2004). Pre-equating: A simulation study based on a large scale assessment model. *Journal of Applied Measurement*, *5*, 301–318.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36*, 336–346. doi:10.1111/j.1745-3984.1999.tb00560.x

Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*, 329–346. doi:10.1111/j.1745-3984.2000.tb01090.x

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.

Thorndike, R. L. (1971a). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Thorndike, R. L. (1971b). The problem of guessing. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 59–61). Washington, DC: American Council on Education.

Thurstone, L. L. (1919). A method for scoring tests. *Psychological Bulletin*, *16*, 235–240. doi:10.1037/h0069898

Tong, Y., Wu, S.-S., & Xu, M. (2008, March). *A comparison of pre-equating and post-equating using large-scale assessment data.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge, England: Cambridge University Press.

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*, 421–459. doi:10.1007/BF02290601

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York, NY: Springer.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., . . . Thissen, D. (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Hillsdale, NJ: Erlbaum.

Walker, M. E., & Kim, S. (2010). *Examining two strategies to link mixed-format tests using multiple-choice anchors* (ETS Research Report No. RR-10-18). Princeton, NJ: Educational Testing Service.

Walker, M. E., Zhang, L. Y., & Zeller, K. E. (2002, April). *Estimating internal consistency reliability of tests for ethnic and gender subgroups within a population.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wendler, C., & Walker, M. E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 445–467). Mahwah, NJ: Erlbaum.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education/Praeger.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*(2), 15–25.

CHAPTER 25

LEGAL ISSUES IN EDUCATIONAL TESTING

Christopher P. Borreca, Gail M. Cheramie, and Elizabeth A. Borreca

Educational testing is a global or generic term that refers to any type of testing done for educational purposes. The term *testing* is often used interchangeably with the term *assessment*. A distinction has been made between testing and assessment and the assessment process (Kubiszyn & Borich, 2007), in which testing is considered a more specific activity to quantify an attribute, and the process of assessment is broader based, is more comprehensive, and includes many components. This is an important distinction. For purposes of this chapter, however, these concepts are used interchangeably as educational testing is often done in the context of an assessment process that yields specific results. The implications of educational testing and the assessment process are far reaching, and they range from the ability to progress within the educational setting and graduate, the identification of disabilities, and the admission to and exclusion from programs. Basically, educational testing is used for decision making about individuals, educators, and systems. The National Board on Educational Testing and Public Policy (2002) has stated that tests are powerful tools and can provide useful feedback on teaching and learning, but they also can lead to serious negative consequences. This conclusion was reiterated by Kubiszyn and Borich (2007), who stated that tests are tools and can be both used and misused.

In an invited paper presented 15 years ago at the Joint Committee on Testing Practices, Bersoff (1996) indicated that tests are neutral, do not inherently discriminate, and have been used to "admit, advance, and employ," but also have been used to "segregate, institutionalize, track, and deny access" (p. 1). Bersoff stated,

> as . . . uses of tests multiplied so did their potential for causing unjustified negative consequences. When those consequences led to legally cognizable injuries, tests began to be examined by the legal system . . . there is probably no current activity performed by counselors, educators and psychologists so closely scrutinized and regulated by the legal system as testing. (p. 1)

Although testing and assessment is typically done appropriately, with valid and useful purposes (for a thorough description of test validity, see Volume 1, Chapter 4, this handbook), there are classic cases of misuse, especially in areas of discrimination (e.g., *Larry P. v. Riles*, 1972/1979/1984), and testing needs legal mandates and protections to prevent misuse and ensure appropriate use.

More critical decisions about individuals and systems are being made through testing results, and the U.S. law, the No Child Left Behind (NCLB) Act of 2001, mandates standardized testing nationwide. Controversy continues over what is now termed high-stakes testing, which can deny educational progression and high school graduation. Testing and test use has not only multiplied, as Bersoff (1996)

The Department of Education's Family Policy Compliance Office (FPCO) administers FERPA, and the Office of Special Education Programs (OSEP) is largely responsible for IDEA. Most of the FPCO and OSEP letters cited are available at http://www.ed.gov.

DOI: 10.1037/14049-025
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

stated, but also has grown exponentially in quantity as well as importance. It is through the establishment of and adherence to professional standards, ongoing professional development, appropriate uses of tests, and legal scrutiny that educational testing can and will maintain its validity and utility in decision making.

LAWS APPLICABLE TO EDUCATIONAL TESTING

The purpose of this chapter is to present relevant laws and their relationship to certain issues in educational testing. Several sources of law govern education, including the U.S. and individual state constitutions, statutes enacted by Congress and state legislatures and their implementing regulations, and case law, which is the body of judicial decisions that interpret provisions as they apply to specific situations (Russo & Osborne, 2009). This latter source of law for which interpretations are made is complex; laws in different states may differ greatly, and a hearing in New York may take a different position than one in Texas. State departments of education have state rules, regulations, and policies that conform to the purposes of the federal laws, but they may not be identical and they must be followed by the local education agencies (LEAs) within each state. It is not within the scope of this chapter to review all decisions that may relate in some way to educational testing, and it is not within the scope to thoroughly review specific federal laws or specific laws that pertain to each state. The focus, rather, is placed on those laws that have applications to educational testing in general, the components of such laws that relate directly to educational testing, and laws that have effects across a wide variety of issues. To meet that purpose, five laws, including their respective implementing regulations, have been selected for this review: the Family Educational Rights and Privacy Act of 1974 (FERPA), Section 504 of the Rehabilitation Act of 1973 (Section 504), the Americans with Disabilities Act of 1990 as amended in 2008 (ADA), the Individuals with Disabilities Education Improvement Act of 2004 (IDEA), and the No Child Left Behind Act of 2001 (NCLB). These five laws collectively address the major issues in testing ranging from confidentiality and privacy, to discrimination and access, to educational diagnoses and overall educational accountability.

FAMILY EDUCATIONAL RIGHTS AND PRIVACY ACT OF 1974

FERPA is a U.S. federal law codified at 20 U.S.C. §1232g. The implementing regulations of FERPA are in title 34, part 99 of the *Code of Federal Regulations* (34 CFR 99). FERPA is also referred to as the Buckley Amendment after its principal sponsor, Sen. James Buckley of New York. FERPA has been in existence for more than 30 years and has been amended several times, with the latest revision occurring in 2008. FERPA provides privacy protections for education records when such records are held by federally funded educational institutions (elementary, secondary, and postsecondary educational institutions). Although it is rare for an educational institution not to receive federal funding of some sort, parochial and private schools at the elementary level do not generally receive federal funds and are not subject to FERPA (however, state laws and regulations may apply). The regulations of FERPA involve three major issues: access to education records (34 CFR Subpart B §99.10–99.12), an opportunity to have records amended (34 CFR Subpart C §99.20–99.22), and disclosure of information from the records (34 CFR Subpart D §99.30–99.39).

Key to FERPA and its regulations is the definition of education records (34 CFR Subpart A §99.3) because information about a student is protected by FERPA depending on whether it meets the statute's definition of an education record. When first enacted, FERPA provided a list of data that could be reviewed, such as academic work, grades, standardized achievement and intellectual test scores, psychological tests, interest inventory results, and teacher or counselor ratings and observations, but then they substituted the term *education record* for the list of data and provided a broad definition of an education record. Currently, FERPA defines education records as "those records that are directly related to a student and maintained by an educational agency or institution or by a party acting for

the agency or institution" (§99.3). Although there are exceptions to the term *education record*, generally this term relates to any record that includes personally identifiable information (e.g., name, date of birth, a social security or student number, biometric).

One recent case illustrates interpretation of the term "education record" and brings in the current issue of computer media. In *S. A. by L. A. and M. A. v. Tulare County Office of Education* (2009), the student's parents requested copies of all e-mails sent or received by the district concerning or personally identifying their 10-year-old son, a student with autism and a speech and language delay. The district sent the parents hard copies of e-mails that had been placed in their son's permanent file. The parents filed a complaint arguing that all e-mails that specifically identify the student whether printed or in electronic format are education records. The issue in this case rested on the interpretation of the FERPA requirement that an education record is one that is "maintained" by an educational agency. The U.S. District Court, Eastern District of California upheld the district's interpretation that it was only required to disclose e-mails maintained in the student's file; the parents argued that all e-mails are maintained in the district's electronic mail system and could be retrieved through technology even if previously deleted. The court noted that FERPA does not contemplate that education records be maintained in numerous places, that e-mails have a fleeting nature, and Congress contemplated that education records be kept in one place. The court explained that only those documents in the student's permanent file are considered "maintained."

In *Washoe County School District, Nevada State Education Agency* (2009), a different decision was reached by a state education agency (SEA) hearing officer. The parents of a student with autism exchanged e-mail correspondence with teachers throughout the school year regarding their son's education. When they requested a complete copy of their child's education records, these e-mails were not included. The district explained that e-mails, unless archived by the staff, were deleted from its server within 60 days. The district was found in violation of FERPA (and also IDEA) by not making available for the parents' inspection and review all the student's education records, and the SEA cited FERPA regulation 34 CFR 99.10, which defines an education record to mean any information recorded in any way, including computer media. In this case, the Nevada education agency indicated that violations arose in part because the district lacked policies regarding managing e-mails that are education records and how staff would inform parents when personally identifiable information was no longer needed.

Rights of Inspection and Review

In 34 CFR Subpart B (§99.10–99.12), the parent or eligible student is given the opportunity to inspect and review the student's education records. Again, there are regulations related to how and where such reviews can be conducted, limitations to this right (e.g., if the record contains information about more than one student), and whether copies can be made. The primary issue here is related to the right of the parent or eligible student (age 18 or attending a postsecondary education institution) to review or inspect an education record, and testing records, such as protocols administered for diagnostic assessment purposes, would be included in this definition. In October 1997 the Family Policy Compliance Office of the U.S. Department of Education issued a memorandum regarding access to test protocols and answer sheets identifying them as education records consistent with the FERPA definition (Rooker, 1997). A test protocol includes the answers provided to tests, and the scores are generated from the answers and performance of the student. Allowing access to protocols would allow for the detection of any errors in scoring and administration, and there is no doubt that test protocols are education records and can be inspected and reviewed.

There is much case law about education records and their review, and the Nevada case specifically addressed the right of inspection and review in e-mail correspondence. Although test protocols can be inspected and reviewed, several cases have addressed the issue of whether a parent can have a copy of the protocol. One of the issues in an Illinois school district case (School District U-46, 2005) was whether the student had been properly evaluated. The school district took the position that test protocols

were exempt from disclosure under the state mental health code or under federal copyright laws. Both positions were rejected by the hearing officer who determined that the parent had a right to review test protocols. In California, a decision was reached in the same year in *Newport-Mesa Unified School District v. State of California Department of Education* (2005) allowing copies of test protocols to be given to parents of students with disabilities. The court recognized that the existing body of laws considered it a copyright violation to distribute copies of standardized test protocols but cited "fair use" under copyright law. Fair use (section 107 of federal copyright law) is an exception to allow single copies of copyrighted documents and refers to four considerations: the nature of the use of the material (e.g., for nonprofit educational use), the nature of the material (e.g., test security issues), the amount of copyrighted work used, and the effect of using the material in a potential market.

Although court decisions have been in favor of copying test protocols, test publishers are explicitly against this practice. A legal policies document from Pearson Assessments (2006) addressed this issue from the test publishers' perspective. According to this document, the company "strongly oppose[s] the release of copies of protocols" as such release "could threaten the validity of the tests and therefore their value as a measurement tool" (Pearson Assessments, 2006, pp. 2–3). Pearson believed that copying test protocols would violate issues 2 to 4 of the fair use exemption. The legal policies document went on to say that when parents wish to obtain a second opinion regarding their child's testing and scores, the protocol can be copied and sent to another professional for review, but that "materials should pass directly from professional to professional and not through the hands of the parents or their attorney" (Pearson Assessments, 2006, p. 3).

FERPA does not require that copies of education records be given, but rather it allows for copies when in-person review is not feasible. According to 34 CFR 99.10(d), if parents cannot arrange to conduct an inspection during school hours, one option is to provide a copy of the record in lieu of scheduling an evening or weekend appointment. In a recent decision, the U.S. District Court in the Northern District of Alabama (*Bevis ex rel. D. B. v. Jefferson County Bd. of Educ.*, 2007) ruled that a parent of a 15-year-old student had access to her son's educational records and therefore she was not entitled to copies of the student's 999-page file.

Rights Regarding Amending Records

Subpart C of 34 CFR 99 (§99.20–99.22) relates to the right of a parent or eligible student to ask an educational agency to amend the education record if that record is believed to be inaccurate, misleading, or in violation of the student's rights of privacy. The agency or institution has to decide whether or not to amend the record as requested, and if it is decided not to amend the record, then the right to a hearing is granted. Norlin (2008) indicated that *inaccurate* "probably should be interpreted as incorrect"; *misleading* would mean that "the content of the record through overgeneralization, omission, or other intentional or unintentional means of expression, casts the student in a false light or gives an erroneous impression"; and *violation of the student's privacy* involves the "consideration of whether the school has any legitimate educational interest in recording or maintaining information about a student that a reasonable student or parent would consider offensively intrusive" (p. 8).

The most common use of this FERPA provision relates to grade challenges. Grades can be challenged and changed

> only if . . . the grade was supposed to be something other than what was shown on the student's education record; i.e., that it had been inaccurately recorded; that a mathematical error was made in computing the grade; or that there was a scoring error on a test that affected the grade. Absent proof of inaccuracy, FERPA cannot be used as a vehicle to dispute the validity of report cards, tests, or other evaluations. (Norlin, 2008)

Disclosure of Information From Education Records

Subpart D (§99.30–99.39) of FERPA deals with disclosure of information. Signed and dated written consent is the general principle under this subpart

regarding the disclosure of any personally identifiable information contained within the student's record. The written consent must specify who, what, and why (identify to whom the disclosure will be made, what the records will be disclosed, and the purpose for the disclosure). There are 15 exceptions or limitations to this consent at 34 CFR 99.31(a)(1)–(15), and they involve such conditions as disclosure (a) to other school officials within the agency or institution determined to have legitimate educational interests, (b) to officials of another school system where the student is seeking enrollment, (c) in conjunction with financial aid applications, (d) in compliance with a subpoena, and (e) under emergency situations.

Because many educational institutions employ professionals in a contractual arrangement, there has been some clarification of this disclosure to such individuals. In *Letter Re: Greater Clark County School District* (2006), the privacy protections are extended to persons acting for an educational agency or institution. The key here is that the school or district must have a policy broad enough to include a party with which it has contracted so that the party would be considered a school official with legitimate educational interests in education records to the extent that the party needs to review the education records to fulfill his or her professional responsibility. Regarding transmittal of education records to another school, including a postsecondary institution where the student is seeking enrollment, it should be noted that FERPA states that parents have a right to a copy of what was disclosed if they request it. Test scores are usually disclosed under this exception and in some cases required as in transmittal of transcripts.

SECTION 504 OF THE REHABILITATION ACT OF 1973 AND THE AMERICANS WITH DISABILITIES ACT OF 1990

The rights of children and adults with disabilities are protected by two major federal laws: Section 504 and the ADA. Section 504 of the 1973 Rehabilitation Act (P.L. 93-112, 87 Stat. 394) is codified at 29 U.S.C. §701 with implementing regulations at 34 CFR Part 104. Section 504 is considered to be the first civil rights statute for persons with disabilities. Section 504 was the last sentence in the 1973 Rehabilitation Act, and it took several years for implementing rules to be issued. Section 504 took effect in 1977 and has been amended several times, most recently in 2008 by virtue of the Americans With Disabilities Act Amendments Act of 2008 (ADAAA) effective January 1, 2009. Section 504 is designed to eliminate discrimination on the basis of disability in any program or activity receiving federal funds. Section 504 has a broad reach, including K–12 public education, colleges and universities, and any other entities receiving federal financial assistance (e.g., airports, public libraries).

The ADA was signed into law in 1990 and is codified at 42 U.S.C. §12101. The ADA has five major parts known as Titles and these relate to employment, public services of state and local governments, public services and accommodations, telecommunications, and miscellaneous provisions. According to Russo and Osborne (2009), the ADA "extends the reach of Section 504 to the private sector and programs that do not receive federal financial assistance" (p. 26). Recently, Congress amended the ADA (ADA Amendments Act of 2008, P.L. 110-325) to correct what it believed were inappropriate Supreme Court decisions involving employment that limited the ADA's coverage. These decisions (*Sutton v. United Airlines, Inc.*, 1999; *Toyota Motor Manufacturing, Kentucky, Inc. v. Williams*, 2002) were deemed as narrowing the scope of protection intended to be afforded by the ADA because of their interpretations of the definition of "disability" under the ADA. The ADA Amendments reaffirm the intent that the ADA's definition of disability is interpreted broadly and inclusively. The amendments to the ADA became effective on January 1, 2009. The purpose of the ADA is to "provide a clear and comprehensive national mandate for the elimination of discrimination against individuals with disabilities" (ADA, 1990).

Given that Section 504 and the ADA are closely related, they are discussed together in this section. Section 504 and the ADA both relate to educational testing in three major areas: evaluation procedures, determination of disability, and provision of accommodations.

Evaluation

The Section 504 regulations for initial or preplacement evaluations are found at 34 CFR 104.35 and are similar, but not identical, to those under IDEA (see the section Individuals with Disabilities Education Improvement Act). Generally, tests are selected to ensure that they accurately reflect the student's aptitude or achievement or other factor being measured rather than the student's disability (except in cases in which those are the factors being measured). In addition, the tests and other evaluation materials include those tailored to evaluate specific areas of educational need, are validated for the specific purpose for which they are used, and are selected and administered in a manner that is not racially or culturally discriminatory (for further discussion of validity and bias, see Volume 1, Chapter 4, this handbook, and Chapter 27, this volume). Compliance with IDEA is satisfactory under Section 504. Both IDEA and Section 504 address reevaluations, but in Section 504 it is noted that reevaluations are done periodically and before changes in placement. There is no specified time limit for completion of the initial evaluation or the reevaluation under Section 504, but they must be done in a reasonable time period. Parental consent is required for initial evaluation. The ADA does not list specific evaluation procedures.

One specific issue under evaluation involves the destruction of test protocols. Destruction of protocols is in violation of Section 504 (34 CFR 104.35 and 34 CFR 104.36) in that this would deny parents access to the records that were used to formulate an educational program. Districts should not maintain a policy of categorically destroying psychological protocols (*Allegheny (Pa.) Intermediate Unit*, 1993).

Definition and Determination of Disability

To be protected under Section 504 and the ADA, the individual must show that he or she has a disability. The definition of an individual with a disability under both Section 504 (34 CFR § 104.3) and the ADA (42 U.S.C. §12102) indicates that an individual has to have a physical or mental impairment that substantially limits one or more major life activities, has a record of having such an impairment, or is regarded as having such impairment. The Section 504 regulatory provision at 34 CFR 104.3(j)(2)(i) defines a physical or mental impairment as any physiological disorder or condition, cosmetic disfigurement, or anatomical loss affecting one or more bodily systems (e.g., neurological, musculoskeletal, respiratory, cardiovascular, digestive, etc.) or any mental or psychological disorder (e.g., mental retardation, emotional or mental illness, learning disability). There is no exhaustive list of specific diseases or conditions.

An impairment in and of itself (e.g., diagnosis) is not a disability because to meet the definition for disability, the impairment must substantially limit one or more major life activities. Under Section 504, these activities include functions such as caring for oneself, performing manual tasks, walking, seeing, hearing, speaking, breathing, learning, and working; again, this list is not exhaustive or exclusive. Once determined to have a disability, an individual then is protected against discrimination based on that disability. The ADA major life activities are similar (e.g., caring for oneself, seeing, hearing, walking, speaking, learning, reading, concentrating, working, and operation of a major bodily function), and the concept of being regarded as having such an impairment is further defined (i.e., the individual has been subjected to an action prohibited under the ADA because of an actual or perceived physical or mental impairment).

The definition of disability that prompted the ADA Amendments of 2008 is in reference to the phrase *substantially limits*. Section 504 does not operationally define this term, but the ADA Amendments Act of 2008 has provided further clarification of this term in §12102, Definition of Disability (4)(E). *Substantially limits* is now interpreted "without the regard to the ameliorative effects of mitigating measures" (e.g., medication, equipment, hearing aids, mobility devices, use of assistive technology, reasonable accommodations or auxiliary aides or services, or learned behavioral or adaptive neurological modifications). Thus, the definition of disability under the ADA has now been broadened in scope and the amendments expand the eligibility of K–12 students under Section 504.

Perhaps the most poignant case to illustrate the impact of such an expanded definition of disability is *Garcia v. Northside Indep. Sch. Dist*. decided in 2007

in the U.S. District Court, Western District of Texas. Alexander Garcia was a 14-year-old student with a clearly documented condition of asthma, yet he participated in a wide range of physical activities. He had periodic asthma flare-ups that were treated by use of an inhaler. In 2003, Alexander participated in physical education (PE), and during a running exercise, he began to exhibit breathing problems. His PE teacher accompanied him to the gym to retrieve his inhaler, but Alexander collapsed before reaching the building and died at the hospital later that day. The parents filed a liability claim under Section 504, and the district filed a motion for summary judgment claiming that Alexander was not eligible as an individual with a disability because he did not have an impairment that substantially limited a major life activity. The summary judgment was granted in favor of the school district with the judge citing *Sutton v. United Airlines, Inc.* (1999). The judge noted that because Alexander was able to control his breathing through the use of his inhaler and that he participated in multiple sports, he did not have a disability under Section 504; the use of an inhaler mitigated his asthma condition, and thus asthma did not substantially limit his breathing. Considering the 2008 ADA Amendments Act as applied to this case, it is highly likely that such a summary judgment would not have been in favor of the district, and Alexander would be considered a student with a disability. This does not mean that the parents would prevail in the liability suit, only that the liability claim would be subject to investigation.

A student who is not eligible under IDEA may be eligible under Section 504 or the ADA because these latter laws are more inclusive. If a school district finds a student ineligible under IDEA, that will not excuse its failure to evaluate the student's eligibility under Section 504 (*Yukon (OK) Pub. Schs.*, 2007).

It has long been taken for granted that a student who is eligible under IDEA will always meet eligibility under Section 504 and the ADA. A recent decision indicated that a student's eligibility for special education does not mean that he or she has a disability for purposes of Section 504. Sarah Ellenberg was denied admission to a military academy, and she filed suit claiming that because she received an individualized education program (IEP) for a disability under IDEA, she was automatically disabled under Section 504. The court disagreed (*Ellenberg v. New Mexico Military Institute*, 2009). The decision indicates that although most IDEA eligible students would also be eligible under Section 504, the impairment for Section 504 eligibility must substantially limit a major life activity. According to the 10th Circuit, Section 504 is broader in scope than the IDEA, but all disabilities under the IDEA do not automatically qualify for coverage under Section 504 (e.g., a learning disability may be substantially limiting for one student and not another). Ellenberg did petition the Supreme Court to review her case, but the Court denied the request.

Otherwise Qualified

An additional term in Section 504 and the ADA needs to be defined for the individual to receive the legal protection of these laws. *Otherwise qualified* means that the student is eligible to participate in a program or activity despite the existence of a disability (e.g., a student who meets the academic and technical standards for admission to an education program or activity). In the ADA, the term predominately refers to employment (an individual who with or without reasonable accommodations can perform the essential functions of the employment position), but under both laws *otherwise qualified* also relates to students applying for programs. Although the testing that would take place for entry into such a program can be accommodated, the individual must still be able to perform the essential functions of the position or meet the minimal requirements for admission and continuation in educational programs. In *St. Johnsbury Academy v. D. H.* (2001), the Second Circuit ruled that a private high school in Vermont did not have to enroll a student who was unable to read at a fifth-grade level. The student was not otherwise qualified and the school had no obligation to lower the school's requirements and admit the student. Thus, an otherwise qualified individual meets certain standards and can participate with reasonable accommodations.

Reasonable Accommodation

The term *reasonable accommodation* applies to many things, ranging from making a facility accessible to

providing an accommodation on an examination. It is this latter component ("appropriate adjustment or modifications of examinations") that most specifically relates to educational testing (see Chapter 18, this volume). Most school districts require that students pass exit or comprehensive state examinations to graduate from high school (see the section NCLB), and most postsecondary institutions require admission examinations or interviews. Providing testing accommodations to students with disabilities is covered by both Section 504 and the ADA. An accommodation is reasonable if it does not compromise the nature, content, and integrity of the test. Accommodations are reasonable when they provide students with disabilities an equal opportunity to participate without lowering or fundamentally altering the academic standards. Decisions regarding testing accommodations are made by a group of people knowledgeable about the student, the evaluation data, and the placement options (34 CFR 104.35). There is no single list of acceptable accommodations for testing, and decisions need to be made on an individual basis and may differ given the purposes of the assessment. Accommodations should be documented for each student on a 504 plan. Some common accommodations include oral testing, environmental issues such as reduction of distraction or taking a test in a different location, and extended time. Under Section 504, the required mastery level, materials, and grading are the same as for nondisabled peers. Officials are not required to alter the content of examinations.

In *Plainedge (NY) Union Free Sch. Dist.* (2006), the student's 504 plan included extended time on tests, a separate testing location, and preferential seating. The U.S. Office of Civil Rights (OCR) concluded that although the district did not keep records of all exams given in alternative testing sites, they did not discriminate against the student in violation of Section 504. The district did have documentation that the student was scheduled to take final exams in a separate testing location. In *Lake County (FL) Sch. Dist.* (2008), the 504 plan of a student with an undisclosed disability allowed her to use a compact disc player and headset during tests. The student asked the school if she could use a music player during the Florida Comprehensive Assessment Test. The district did not provide her with a compact disc player and headphones and she did not pass the test. Although the district appropriately declined the use of the music player, the district failed to implement the student's plan by neglecting to provide an accommodation.

In a Massachusetts district (*Springfield (MA) Pub. Schs.,* 2008), the OCR investigated a complaint that a classroom teacher was prevented from providing her students with test-taking accommodations as set forth in their IEPs. Because of a software glitch, the accommodation requirements were listed under general accommodations rather than for state- or districtwide assessments, thus the specialist who reviewed the IEPs found that no accommodations were required. Although the teacher objected and tried to explain the software problem, the instructional specialist informed her that accommodations could not be provided. OCR concluded that nine of 10 students were denied the accommodations to which they were entitled. The district's failure to ensure these testing accommodations was a violation under Section 504 and Title II of the ADA. In *North Rockland (NY) Cent. Sch. Dist.* (2008), a ninth-grade student with an undisclosed disability did not receive the testing accommodations to which he was entitled when he took a preliminary college entrance exam, the PSAT. The accommodation was extended time and OCR determined that the district's policy precluded 9th- and 10th-grade students from receiving accommodations for the PSAT; only 11th-grade students could receive accommodations on the PSAT. The district also declined to assist the parent in completing the eligibility form which mandates that an official school representative complete and sign the form and send it directly to the College Board. Thus, the district did not provide nor assist the student with requesting the accommodations that were provided for in his 504 plan. In a Connecticut case (*Regional (CT) Sch. Dist. No. 17*, 2006), OCR found that the district did not discriminate against an 11th-grade student with a visual and auditory disorder when it failed to grant her unlimited time for taking school tests and quizzes. Noting that the student's 504 plan only entitled her to additional time on standardized tests such as the SAT or PSAT, OCR concluded that the district did not violate Section 504.

Unlike the mandate for public schools, colleges are not required to identify students with disabilities and it is the student directly who must inform the college of the existence of a disability and the need for adjustments. School districts must assist students with disabilities in requesting accommodations for the SAT or similar exams (see *North Rockland (NY) Cent. Sch. Dist.*, 2008). Many of the cases regarding reasonable accommodations for testing are in the postsecondary arena. One example is *Rush v. National Board of Medical Examiners* (2003) decided in the District Court of Texas, Northern District. The plaintiff was a medical student with a learning disability who requested and was denied double time in which to take the U.S. Medical Licensing Exam. The court found that Rush was an individual with a disability because he was substantially limited in the major life activities of reading and learning compared with most people. The court, critical of the Board's experts, granted an injunction requiring the National Board of Medical Examiners (NBME) to provide Rush with the accommodations of double time for the exam, stating that without such accommodations the exam would test his disability and not his mastery of the subject matter.

In a similar case, *Rothberg v. Law School Admission Council, Inc.* (2004) in the District Court of Colorado, a learning disabled plaintiff with a lengthy history of disability diagnosis and accommodations was denied extended time accommodations on the Law School Admission Test (LSAT). She was initially denied extended time because she had not completed the Nelson Denny Reading Test. She took that test subsequently and her score was consistent with earlier evaluations that indicated the need for extended time on the LSAT. She was again denied the accommodation because the Law School Admission Council (LSAC) evaluator relied on the fact that the plaintiff was able to perform in the average or low-average range on the SAT and LSAT without accommodations and that her deficiencies in written expression and mathematical ability would not affect her performance on the LSAT. The court further found the LSAC's proffered expert witness not to be credible on the issue of establishing plaintiff's disability. The court granted the plaintiff a preliminary injunction compelling the LSAC to grant her extended time. The court held that Rothberg was substantially limited in the major life activities of learning and reading. The court rejected LSAC's argument that she did not need accommodations based on her average SAT and LSAT performance because the court found that the plaintiff actually completed only one third of the exam without accommodations and then randomly filled in answers to questions she could not read and that this compensatory technique does not support a finding that she is not disabled. The court found the argument of the LSAC unpersuasive because it determined that without extra time the LSAT would be measuring the plaintiff's disability rather than her knowledge.

In *Powell v. National Board of Medical Examiners* (2004), decided in the U.S. Court of Appeals, Second Circuit, a learning disabled student sued the NBME and the University of Connecticut after she failed the Step 1 Medical Licensing Exam three times and was dismissed from medical school. The plaintiff requested a waiver of the Step 1 Exam requirement from the University of Connecticut, which it refused, and was subsequently denied accommodations of extended time on the exam by the NBME. The Second Circuit held that Powell failed to show that even if she was disabled, she was otherwise qualified to continue to be a medical student at the University of Connecticut, noting that she had a background of educational difficulty and an average to low-average intelligence quotient (IQ). The court also held that there was no proof the University of Connecticut discriminated because they had provided extensive accommodations to the plaintiff but were not required to offer accommodations that fundamentally altered the nature of the service, program, or activity.

INDIVIDUALS WITH DISABILITIES EDUCATION IMPROVEMENT ACT OF 2004

Another piece of legislation affecting testing and evaluation is the IDEA, which was first enacted by Congress in 1975 as P.L. 94-142, the Education for All Handicapped Children Act. Before passage of this law, states were not required to provide special education services to students with disabilities even

though a few states provided some level of services. The IDEA has been amended several times and was most recently reauthorized in December 2004 as IDEA (2004). The implementing regulations were published in the *Federal Register* on August 14, 2006, and were fully in effect 60 days later.

Under IDEA 2004, Part B, funds are provided to states to enable LEAs to carry out the mandates in this law by providing a free, appropriate public education (FAPE) to eligible students with disabilities between the ages of 3 and 21 through the provision of special education and related services designed to meet their unique needs. The provisions in IDEA 2004, Part B address the entire special education process; however, this section focuses only on testing and evaluation provisions related to legal issues. Although the legal issues related to evaluation and eligibility for special education have seemingly become more complex with the introduction of Response to Intervention and revised evaluation procedures, courts and hearing officers continue to use basic eligibility criteria established over the years to analyze these cases. This section highlights case law and administrative decisions regarding the following testing and evaluation provisions: (a) identifying and locating children with disabilities, (b) disproportionality, (c) initial evaluation and reevaluation, (d) evaluation procedures, (e) determination of eligibility, and (f) procedural safeguards.

Identifying and Locating Children With Disabilities

The IDEA includes a *child find* provision that requires states to have policies and procedures in effect to ensure that

> all children with disabilities residing in the State, including children with disabilities who are homeless children or are wards of the State, and children with disabilities attending private schools, regardless of the severity of their disabilities, and who are in need of special education and related services, are identified, located, and evaluated.

Child find also must include children who are suspected of being a child with a disability under 34 CFR 300.8 and in need of special education, even though they are advancing from grade to grade, and highly mobile children, including migrant children (34 CRF 300.111).

Locating and identifying a student as possibly having a disability or suspected of having a disability does not mean that the student is automatically eligible for IDEA services. Under IDEA 2004, the term "child with a disability" means a child evaluated in accordance with 34 CFR 300.304–300.311 and found to meet the eligibility criteria for one or more of the following disability categories: autism, deaf-blindness, deafness, emotional disturbance, hearing impairment, mental retardation (recently changed to intellectual disability [ID]), multiple disabilities, orthopedic impairment, other health impairment (OHI), specific learning disability (SLD), speech or language impairment, traumatic brain injury, and visual impairment (including blindness), *and who, by reason thereof*, needs special education and related services (34 CFR 300.8(a)(1)). Child find also includes children between the ages of 3 and 9 suspected of having a developmental delay if a state or district chose to adopt that term under 34 CFR 300.8(b).

Although the disability categories in 34 CFR 300.8(c) are exhaustive, the list of specific impairments included within the definition of each of the categories of disabilities is not meant to be exhaustive (*Letter to Fazio*, 1994). For example, OHI may include several types of conditions (e.g., attention deficit hyperactivity disorder [ADHD], epilepsy).

To be eligible for special education services, the student must be determined to have 1 of the 13 disability conditions and need *specially designed instruction*, at no cost to the parents, to meet his or her unique needs. In addition to instruction, special education also includes speech–language pathology services and other *related* services (e.g., occupational therapy). If a child has one of the disabilities identified at 34 CFR 300.8(a)(1), but only needs related services and not special education, the child is not a child with a disability under the IDEA (34 CFR 300.8(a)(2)(i)). If, however, the related service that the child requires is considered "special education" under *state* standards, the child will then be eligible under the IDEA (34 CFR 300.8(a)(2)(ii)).

For example, in Texas speech and language services are considered instructional and not related services.

The *child find* provision of the IDEA indicates that children suspected of being disabled *and* in need of special education be identified and evaluated. Before this identification, students should be provided with intervention. The use of a response to intervention (RtI) strategy in the identification process has been addressed in a memorandum from the Office of Special Education Programs (OSEP) dated January 21, 2011, to state directors of special education. In this memorandum, RtI was defined as "a multi-tiered instructional framework" and "a school-wide approach that addresses the needs of all students including struggling learners." This RtI process is designed to identify at-risk students, monitor progress, provide interventions, and adjust these interventions depending on the student's response to instruction. The memorandum indicated that it has come to OSEP's attention that some LEAs "may be using Response to Intervention (RTI) strategies to delay or deny a timely initial evaluation for children suspected of having a disability." The memo went on to indicate that "LEAs have an obligation to ensure that evaluations of children suspected of having a disability are not delayed or denied because of implementation of an RTI strategy." The memo further stated that "the use of RTI strategies cannot be used to delay or deny the provision of a full and individual evaluation, pursuant to 34 CFR §§300.304-300.311, to a child suspected of having a disability under 34 CFR §300.8."

Several cases deal with *child find* issues regarding whether evaluations should have been conducted. In a Pennsylvania case (*Richard S. v. Wissahickon Sch. Dist.*, 2009), a decision was reached that the district did not violate the IDEA by failing to evaluate a middle-school student with ADHD when his academic performance declined. The Third U.S. Circuit Court of Appeals affirmed a decision that the student's academic problems stemmed from his lack of motivation and poor attendance rather than his disability. According to Slater (2010), this case points out that districts should consider the source of a student's motivational difficulties when determining eligibility. In *Regional Sch. Dist. No. 9, Bd. of Educ. v. Mr. and Mrs. M. ex rel. M. M.* (2009), a U.S. District Court in Connecticut ruled that because a Connecticut district had notice that a high school student had been placed in a psychiatric hospital, it should have evaluated the student's need for special education and related services. The court found that the school district violated *child find*, found the student eligible under IDEA, and required the district to reimburse the parents for the student's therapeutic placements. These cases illustrate that many factors need to be considered in determining the need for an evaluation for IDEA eligibility; however, it is clear that if a disability is suspected, then the evaluation should take place.

Disproportionality

A major issue related to identification is the inappropriate overidentification or disproportionate representation by race or ethnicity of children with disabilities. This is a national issue that is addressed through the IDEA 2004. All states must have policies and procedures to prevent this from occurring. In an April 2007 memorandum to state directors of special education, Alexa Posny, OSEP, addressed this issue, stating that

> excerpts from findings in the Individuals with Disabilities Education Act (IDEA) 2004's statute note that: (1) greater efforts are needed to prevent the intensifications of problems connected with mislabeling minority children with disabilities; (2) African-American children are identified as having mental retardation and emotional disturbance at rates greater than their white counterparts; (3) more minority children continue to be served in special education than would be expected from the percentage of minority students in the general school population; . . . States are required to address disproportionality . . . in the State Performance Plan. . . . Failure to conduct this analysis will be cited as noncompliance . . . which requires that States monitor LEAs with regard to disproportionate representation of racial and ethnic groups in special education

> and related services, to the extent the representation is the result of inappropriate identification.

A new provision included in the IDEA regulations is Early Intervening Services (EIS) (34 CFR 300.226). These new requirements are designed to help students who are not identified as having a disability, but who need additional academic and behavioral support to succeed in general education. According to 34 CFR 300.646(b)(2),

> in the case of a determination of significant disproportionality with respect to the identification of children as children with disabilities . . . the State or the Secretary of the Interior must . . . require any LEA identified . . . to reserve the maximum amount of funds [15%] . . . to provide comprehensive coordinated early intervening services.

The EIS regulation pertains to all students from kindergarten through 12th grade, but it has a particular emphasis on students in kindergarten through Grade 3. Given this regulation, it can be interpreted that the EIS is designed to provide support for students and also to assist in the reduction of disproportionality.

Initial Evaluation and Reevaluation

If a district suspects or has reason to suspect that a student may have a disability, it must obtain informed consent of the child's parent before conducting an initial evaluation to determine whether a child qualifies as a child with a disability. The public agency must make reasonable efforts to obtain the informed consent from the parent before conducting the initial evaluation to determine whether the child is a child with a disability (34 CFR 300.300(1)(i)(iii)). A public agency has the option to pursue the initial evaluation of a child using the procedural safeguards if a parent does not provide consent or fails to respond to a request to provide consent for an initial evaluation (34 CFR 300.300(a)(3)). Although a school district may use due process to override lack of consent, they are not required to do so, and in the Federal Register (2006), *Analysis of Comments and Changes to 2006 IDEA Part B*, it is noted that public agencies should use consent override procedures only in rare circumstances: "State and local educational agency authorities are in the best position to determine whether, in a particular case, an initial evaluation should be pursued" (p. 46632). It should be noted that overriding lack of consent for evaluation applies only to students in public school.

A recent case illustrates the impact of lack of parental consent for an initial evaluation from one parent but not the other. In *J. H. v. Northfield Pub. Sch. Dist.* (2009), the court ruled that because one of the student's parents refused to consent to an initial evaluation, a Minnesota district could not evaluate the student's need for special education and related services. The Minnesota Court of Appeals held that the consent of the other parent was not enough to allow the district to proceed with the assessment. According to Slater (2010), when parents disagree between themselves about the need for an initial evaluation, the right to evaluate will depend on state law. In this case, Minnesota law provides that a district cannot proceed with an evaluation if a parent provides written refusal to consent. The IDEA does not address this issue regarding disagreement between a student's biological parents.

An initial evaluation of a student with a suspected disability occurs before the student's first special education placement. Districts must follow several procedures to ensure that the evaluation meets all of the legal requirements (see the next section Evaluation Procedures). After the initial evaluation, the student will undergo periodic reevaluations while remaining eligible under IDEA.

A public agency must obtain informed parental consent before conducting any reevaluation of a child with a disability. According to 34 CFR 300.300(c),

> if the parent refuses to consent to the reevaluation, the public agency may, but is not required to, pursue the reevaluation by using the consent override procedures. . . . The informed parental consent . . . need not be obtained if the public agency can demonstrate that it made reasonable efforts to obtain the consent and the child's parent failed to respond.

Because the IDEA mandates a reevaluation to occur at least once every 3 years, school districts are given due process rights to override lack of consent for the reevaluation as well as the initial evaluation.

In *G. J. v. Muscogee County School District* (2010), the parents of a 7-year-old student with autism withheld their consent for reevaluation by placing numerous conditions on the reevaluation (e.g., preferred evaluator, approval of each instrument, mother's presence for the testing). The court noted that with such restrictions, there was not consent. The court ordered the parents to consent to the reevaluation to continue to receive special education services.

Parental consent is not required before either reviewing existing data as part of an evaluation or a reevaluation, or administering a test or other evaluation that is administered to all children unless, before administration of that test or evaluation, consent is required of parents of all children (34 CFR 300.300(d)). Although issues of consent for evaluation and reevaluation can be confusing, the IDEA basically requires that school districts evaluate students suspected of having disabilities and in need of special education and reevaluate students at least once every 3 years once they are eligible. Because of these requirements, parental consent needs to be obtained to conduct the evaluations or reevaluations. Parents sometimes do not agree with the recommended evaluation or reevaluation request from the district, however, and thus the district has due process rights to override such consent. Although this is not a common practice, when a school district feels very strongly that identification and provision of services are needed, they can seek to evaluate by use of the due process hearing procedures.

In *Brazosport ISD v. Student* (2007), the sole issue was to override lack of parental consent for a reevaluation of a high school student with autism. The major purpose of the reevaluation was to provide information about the student's current functioning levels to develop an appropriate IEP. The student's teachers were concerned about the appropriate placement of the child and how to best educate him. The parents gave initial consent but then "thwarted the completion of the evaluation process" and withdrew their consent for the assessments to proceed. The deadline for the 3-year reevaluation passed "leaving the district open to legal exposure for failing to perform the duties owed to the child under law." The hearing officer ordered lack of parental consent to be overridden and that Brazosport ISD complete a series of evaluations needed to "gauge the child's current levels of performance."

The IDEA does not require educational agencies to test all children for whom evaluations are requested (*Pasatiempo v. Aizawa*, 1996). A district conducts an initial evaluation when it suspects that the student has a disability and by virtue of that disability needs special education and related services. If a district has no reasonable basis to suspect that a student has a disability, it may refuse to conduct an evaluation (*Letter to Williams*, 1993). If a public agency does not suspect that the child has a disability and denies the request for an initial evaluation, the public agency must provide written notice to the parents consistent with 34 CFR 300.503(b); the notice of refusal must explain why the public agency refuses to conduct an initial evaluation and the information that was used as the basis to make that decision. The parent may challenge such a refusal by requesting a due process hearing. In *Clark County Sch. Dist.* (2002), the hearing officer concluded that there was no reasonable basis for suspecting that the student had a disability that required special education and related services, and thus the district was not required to conduct an evaluation.

Before the provision of special education and related services, an SEA, other state agency, or an LEA must conduct a full and individual evaluation. This initial evaluation must be conducted within 60 days of receiving parental consent for the evaluation, or if the state establishes a timeframe within which the evaluation must be conducted, within that timeframe (34 CFR 300.301(a)(c)). If evaluations do not meet the timelines, a district is subject to remedial actions. Relief awarded for evaluation delays has included compensatory education as well as reimbursement for privately obtained services (*Department of Educ. v. Cari Rae S.*, 2001).

According to 34 CFR 300.303,

> a public agency must ensure that a reevaluation of each child with a disability is conducted . . . if the public

> agency determines that the educational or related services needs, including improved academic achievement and functional performance of the child warrant a reevaluation or . . . if the child's parent or teacher requests a reevaluation.

The regulations indicate that reevaluations must occur at least once every 3 years and may not occur more than once a year; this can be modified if the parent and the public agency agree.

In *Springfield Sch. Committee v. Doe* (2009), it was determined that a district denied FAPE to a student with cognitive and behavioral difficulties by failing to reevaluate him after he missed 32 school days in less than 2 months and had a history of missing class. According to Slater (2010), the IDEA does not explicitly require a district to reevaluate a student just because he has been truant for a specific number of days. Frequent absenteeism, however, may trigger a district's duty to respond depending on the content of the student's IEP and other circumstances. This student's IEP contained a goal to better manage his school responsibilities, including staying in class. Given that goal, the district should have determined that a reevaluation was necessary following the numerous unexcused absences.

Evaluation Procedures

The IDEA 2004 Federal Regulations, 34 CFR 300.304 contain the following evaluation procedures:

(a) Notice. The public agency must provide notice to the parents of a child with a disability, in accordance with Sec. 300.503, that describes any evaluation procedures the agency proposes to conduct.

(b) Conduct of evaluation. In conducting the evaluation, the public agency must—

(1) Use a variety of assessment tools and strategies to gather relevant functional, developmental, and academic information about the child, including information provided by the parent, that may assist in determining—

(i) Whether the child is a child with a disability under Sec. 300.8; and

(ii) The content of the child's IEP, including information related to enabling the child to be involved in and progress in the general education curriculum (or for a preschool child, to participate in appropriate activities).

(2) Not use any single measure or assessment as the sole criterion for determining whether a child is a child with a disability and for determining an appropriate educational program for the child; and

(3) Use technically sound instruments that may assess the relative contribution of cognitive and behavioral factors, in addition to physical or developmental factors.

(c) Other evaluation procedures. Each public agency must ensure that—

(1) Assessment and other evaluation materials used to assess a child under this part—

(i) Are selected and administered so as not to be discriminatory on a racial or cultural basis;

(ii) Are provided and administered in the child's native language or other mode of communication and in the form most likely to yield accurate information on what the child knows and can do academically, developmentally, and functionally, unless it is clearly not feasible to so provide or administer;

(iii) Are used for the purposes for which the assessments or measures are valid and reliable;

(iv) Are administered by trained and knowledgeable personnel; and
(v) Are administered in accordance with any instructions provided by the producer of the assessments.

(2) Assessments and other evaluation materials include those tailored to assess specific areas of educational need and not merely those that are designed to provide a single general intelligence quotient.
(3) Assessments are selected and administered so as best to ensure that if an assessment is administered to a child with impaired sensory, manual, or speaking skills, the assessment results accurately reflect the child's aptitude or achievement level or whatever other factors the test purports to measure, rather than reflecting the child's impaired sensory, manual, or speaking skills (unless those skills are the factors that the test purports to measure).
(4) The child is assessed in all areas related to the suspected disability, including, if appropriate, health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative status, and motor abilities.
(5) Assessments of children with disabilities who transfer from one public agency to another public agency in the same school year are coordinated with those children's prior and subsequent schools, as necessary and as expeditiously as possible, consistent with Sec. 300.301(d)(2) and (e), to ensure prompt completion of full evaluations.
(6) In evaluating each child with a disability under Sec. 300.304 through 300.306, the evaluation is sufficiently comprehensive to identify all of the child's special education and related services needs, whether or not commonly linked to the disability category in which the child has been classified.
(7) Assessment tools and strategies that provide relevant information that directly assists persons in determining the educational needs of the child are provided.

As seen in these regulations, the procedures reflect best practice in assessment and testing. These practices ensure that information is reviewed regarding classroom performance and that students are evaluated with sound (reliable and valid) instruments and that a variety of assessment tools and strategies are used. The regulations also indicate that assessments must be nondiscriminatory. Trained personnel must administer the assessments. Finally, the regulations directly address the need that the child be assessed in all areas related to the suspected disability.

In *N. B. and C. B. ex rel. C. B. v. Hellgate Elementary Sch. Dist.* (2008), a Montana district referred the parents of a preschool child to a child development center to obtain an evaluation. The Ninth Circuit held that the district violated its IDEA obligations and should have evaluated the child for autism. School districts cannot abdicate their affirmative duties under the IDEA. The district still has a responsibility to evaluate the child in all areas of suspected disability even if a child's parents have the ability to obtain an evaluation.

A recent case that illustrates a failure to assess in all areas related to the suspected disability is *W. H. by B. H. and K. H. v. Clovis Unified School District* (2009). In this case, the U.S. District Court, Eastern District of California noted that the district failed to assess the student in the area of written expression. The district had found that an evaluation of written expression was unnecessary because the student performed adequately on a single writing sample. According to Slater (2010), "the court observed that the student produced the sample only after the district psychologist changed the topic, marked the appropriate space on the paper, redirected the

student and provided incentives" (pp. 1–7). As a result, the only evidence of the student's written expression was obtained when the student was provided with one-on-one assistance. The court ruled that the student's inflated grades and test scores were an inaccurate assessment of his achievement because they were the result of accommodations. As a result, the judge found that the student was eligible for IDEA services (Slater, 2010). On the basis of this case, it is evident that a school district should consider a student's performance without accommodations and modifications to show an accurate estimate of abilities.

Of the 13 categories of disabilities recognized by the IDEA, SLD is the only disability category for which the IDEA has additional evaluation procedures beyond the general evaluation requirements for all students with disabilities. IDEA 2004 now requires states to adopt criteria for determining whether a child has an SLD. A state cannot require a district to consider a severe discrepancy between intellectual development and achievement for determining whether a child has a specific learning disability (34 CFR 300.307(a)(1)). Moreover, the state must permit the use of a process based on the student's response to scientific, research-based intervention (34 CFR 300.307(a)(2)). A state may permit a district to use other alternative research-based procedures to determine whether a student has an SLD (34 CFR 300.307(a)(3)). The district, for its part, must use whatever criteria the state has adopted for determining whether a student has an SLD (34 CFR 300.307(b)).

This is a major change in the evaluation and identification of students with learning disabilities because there are now options in determining this disability category, and these options and procedures are determined at the state level. According to Zirkel and Thomas (2010), a survey of state laws indicated that 12 states were requiring RtI as the approach for determining SLD, most states were using a combination of RtI and severe discrepancy, and about 20 states were allowing the third research-based option. In a recent update, Zirkel (2011) reported that 14 states were requiring RtI for SLD eligibility and that Wisconsin would be joining this group effective 2013.

The *Analysis of Comments and Changes* in the Federal Register (2006) stated that "an RTI process does not replace the need for a comprehensive evaluation . . . and a child's eligibility for special education services cannot be changed solely on the basis of data from an RTI process" (p. 46648). Regardless of whether an LEA makes use of an RtI model mandatory or permissive, schools cannot base their eligibility determinations solely on the RtI process. School personnel must use a variety of assessment tools and strategies to evaluate students suspected of having SLDs, even if an RtI is part of the evaluation process.

Several cases illustrate the use of the various methods in determining SLD eligibility. In *E. M. by E. M. and E. M. v. Pajaro Valley Unified Sch. Dist.* (2009), a bilingual student was distracted in the classroom and failed to complete homework assignments. However, he responded well to interventions in the classroom. Thus, the student's positive response to interventions showed that he did not have an SLD and he did not require specialized instruction. In *Student bnf Parent v. Northwest ISD* (2009), a Texas hearing officer cited the various methods by which a student can be identified as SLD under IDEA. The district used the pattern of strengths and weaknesses in processing option to determine that the student was not exhibiting an SLD, whereas an independent evaluator used a discrepancy between IQ and achievement to determine the student did exhibit an SLD. The hearing officer noted that the "best practice" is the use of the processing model for SLD determination and found in favor of the school district. Because of various options allowed and reliance on state criteria, it is likely that the SLD determination will become an active area of litigation.

Determination of Eligibility

IDEA 2004 and its implementing regulations (34 CFR 300.306) contains the procedures for the determination of eligibility. The IDEA regulations also include a special rule for eligibility determination that details when a child must not be determined to be a child with a disability (34 CFR 300.306(b)). This rule is usually referred to exclusionary factors and indicates that children should

not be determined eligible if there has been a lack of appropriate instruction in reading and math or limited English proficiency, or if the student does not meet the eligibility criteria for a specific disability category.

Eligibility under IDEA is a two-prong approach. The student must meet the criteria for 1 of the 13 disability categories and also must demonstrate an educational need. The concept of the need for special education is not specifically outlined in IDEA. 34 CFR 300.306(a) provides that upon completion of the administration of assessments and other evaluation measures, a group of qualified professionals and the parent of the child determines whether the child is a child with a disability, as defined in 34 CFR 300.8, in accordance with 34 CFR 300.306(b) and the educational needs of the child. Parents must be included in the team making eligibility determinations (34 CFR 300.306(a)), and due process procedures generally govern instances when the parents disagree with other team members. When district team members disagree among themselves, resolution is a matter of state or local law or policy. Typically, team members strive for consensus.

Eligibility decisions are complex and require the review of all data when making decisions. IDEA also has regulations (34 CFR 300.305(a)) that certain procedures must be followed that include reviewing existing evaluation data, including information provided by parents. In addition to current classroom-based, local, or state assessments, classroom-based observations are reviewed along with observations by teachers and related service providers. On the basis of that review, and input from the child's parents, service providers identify what additional data, if any, are needed to determine whether the child is a child with a disability, as defined in 34 CFR 300.8, and determine the educational needs of the child.

Procedural Safeguards

The IDEA allows for many procedural safeguards in which the rights of parents are outlined and requires that school districts give parents a copy of the procedural safeguards document (usually downloaded from the state education agency) once per year. Those safeguards directly related to testing issues involve prior written notice for evaluations, informed consent, request for an independent educational evaluation (IEE), and the confidential review of educational records. Issues of consent have been addressed previously in this chapter. The information on FERPA in this chapter applies to IDEA as well. In addition, parents do have the right to be provided with a description of any evaluation, procedure, test, record, or report the district used as a basis for a proposed action (or inaction).

According to the OSEP (2009) Model Form-Procedural Safeguards Notice, the IEE is conducted by a qualified examiner who is not employed by the school district that is responsible for the education of the child. Parents have a right to obtain an IEE if they disagree with the evaluation of their child. If the parents request an IEE, the school district must provide them with information on where they may obtain an IEE and about the district's criteria that apply to IEEs, as the criteria under which the evaluation is obtained, including the location of the evaluation and the qualifications of the examiner, must be the same as the criteria that the district uses when it initiates an evaluation. If the parents obtain an IEE at public expense or share an evaluation of their child with the school district that was obtained at private expense, the district must consider the results of that evaluation in any decision made regarding the provision of FAPE to the child as long as it meets the district's criteria for an IEE. In addition, the evaluation may be presented as evidence at a due process hearing regarding the child.

A recent hearing in Texas addressed the issue of who could conduct the IEE (*Student bnf Parent v. Humble ISD*, 2010). In this case, the parent wanted to have the student evaluated by an individual who was qualified but did not have the same credential as that required by the district (the IEE provider did not have the same certification or licensure that would be required of district personnel). The hearing officer concluded that district could determine the criteria and the qualifications of the examiner, thus rejecting the parent's request for a specific evaluator. Thus, each school district can determine the criteria for IEEs.

Etscheidt (2003) conducted a qualitative content analysis of administrative decisions and cases related to IEEs to identify the criteria for judging the

appropriateness of a district's evaluation. Published decisions from administrative hearings, district courts, and appellate courts that addressed the issue of IEEs and the adequacy of district evaluations were included. The results of the analysis identified three criteria that administrative officers and judges utilized in determining the appropriateness of the evaluations. These criteria were as follows: (a) technical adequacy or compliance with the IDEA regulations for conducting the evaluation, (b) scope and ensuring that all areas of suspected disability and need for special and related services were included, and (c) utility or ability to use the evaluation to develop the IEP.

A case that illustrates some of these issues is *D. B. v. Bedford County Sch. Bd.* (2010). In this case, a Virginia school district evaluated a student for ADHD but failed to evaluate him for an SLD. The court ruled that a thorough evaluation of the student was not conducted. This relates to lack of scope as the child was not evaluated in all areas of suspected disability. After finding the student eligible as OHI, he was placed in inclusion classes for 4 consecutive years but did not achieve any reading goals. This case also noted that the IEP that was developed was not appropriate and that the student's program or services might have changed had he been fully evaluated. Thus, the evaluation also lacked appropriate utility in that it was insufficient to develop the IEP.

NO CHILD LEFT BEHIND ACT OF 2001

The NCLB is the 2001 reauthorized Elementary and Secondary Education Act (ESEA) of 1965. The major purpose of the NCLB is to "ensure that students in every public school achieve important learning goals while being educated in safe classrooms by well-prepared teachers" (Yell & Drasgow, 2005, p. 8). NCLB was passed in 2001, regulations were passed in 2003, and new regulations were published in 2007. NCLB was reauthorized in 2008 and as of 2011 was up for reauthorization; thus, the NCLB is dynamic and subject to ongoing changes in its implementation and requirements, but the basic tenets of the law remain intact. According to Thorndike and Thorndike-Christ (2010), four recurring themes or principles are found in the NCLB, and these include accountability, research-based instruction, control and flexibility in use of federal funds, and parental choice. Of particular relevance to this chapter is the first theme (principle of accountability) because this element of the NCLB involves measurement or testing.

States are required to develop content area standards (e.g., reading, math, science), develop tests to assess student proficiency regarding these standards, and assess all students in reading and math annually between Grades 3 and 8 and once between Grades 10 and 12. On the basis of NCLB, state test results are reported annually, and these data are used to evaluate the effectiveness of schools. In addition to the overall data set, data must be provided for various groups (e.g., racial and ethnic groups, students with disabilities, students who are economically disadvantaged, students with limited English proficiency) to ensure that schools are accountable for the academic improvement of all children.

Although all students are required to participate in this assessment program, there are acceptable accommodations and modifications based on student characteristics (e.g., disability), and for students with severe disabilities, the testing must be aligned with the student's IEP. There are also several allowable assessment options for students with disabilities, which have been outlined and discussed by Borreca and Borreca (2008, pp. 340–342). Allowable state assessment options range from participating in the general grade-level assessment without or with modifications, to alternate assessments judged against varying degrees of grade-level, modified, or alternate achievement standards. Thus, *all* students participate in this mandated assessment program. Given such a mandate, NCLB has ushered in the largest and most universal educational testing and assessment program in this century. As Thorndike and Thorndike-Christ (2010) pointed out, "every spring every third- through eighth-grade student enrolled in a public school in the United States is taking some kind of standardized test" (p. 225; note, however, that some states, e.g., Michigan, have fall testing).

One major issue in the accountability demands of NCLB is the requirement for adequate yearly progress (AYP). The expectation is that all students meet proficiency (grade-level mastery) in reading

and math. To accomplish such a goal, the state sets an annual proficiency target and measures student progress from baseline. Students and subgroups must meet these targets and show incremental increases in proficiency. The aim is to attain this goal by the year 2013–2014. At the time of the writing of this chapter, Congress was considering amending this AYP requirement.

Although the NCLB does not actually mandate a consequence to individual students for not meeting the expectations on state assessments, states have used state assessments to make decisions about individuals, especially as related to promotion, retention, and graduation. This is often referred to as high-stakes testing, and several legal issues are relevant in this context. One involves the basic question of whether state testing can be used for such decisions. In *Erik V. v. Causby* (1997), a case that occurred before the passage of NCLB, parents challenged the retention policy of Johnston County Schools, North Carolina, as related to failure of the state assessment in Grades 3 through 8. The court rejected the parents' due process and deferred to the school's ability to make its own policy.

Another issue is the degree to which the tests themselves have content-related validity evidence. In the landmark case of *Debra P. v. Turlington* (1979, 1981), also before NCLB, 10 African American students who failed a statewide test required for graduation in Hillsborough County, Florida, challenged the testing requirement as racially biased and administered without adequate notice. This case established two major requirements: (a) There must be adequate notice, meaning that students be told what a graduation test will cover several years before the test is implemented, and (b) there must be demonstrated instructional and curricular validity, meaning that the schools are teaching what is being tested. To demonstrate instructional validity for this case, the state conducted a study of its school districts, which involved self-report of instruction and activities related to the test, a teacher survey asking whether this instruction had been provided, and on-site visits to verify the districts' self-reports. A student survey was also part of the on-site visit.

In *Crump v. Gilmer Independent School District* (1992), high school seniors at Gilmer High School in Texas failed to successfully complete the Texas Assessment of Academic Skills (TAAS) and were denied a high school diploma and the right to participate in the graduation ceremony. The court allowed the students to participate in graduation although they had failed the TAAS test and held that the district had not yet made a showing that the material covered on the test was taught in the school. It was noted that the court could not assess whether the TAAS was "adequately linked to the school curriculum."

McClung (1979) identified two types of content validity evidence directly related to minimum competency testing: curricular (match between test and curriculum) and instructional (match between test and what is actually taught in the classroom). It appears that both of these types are necessary to establish the test's validity in decision making. The cases presented occurred before the passage of NCLB, and thus NCLB did not create high-stakes testing because these issues were present long before the passage of this law. Since NCLB, however, there have been no legal challenges that specifically relate to the content validity of the state assessment. Perhaps this is because the NCLB requires that the state achievement standards align with both what is being taught (curriculum) and how mastery is measured (the test). In addition, each state submitted individual peer-reviewed research in support of their own final assessment system, which was reviewed by the U.S. Department of Education (2007).

Another reason for the lack of lawsuits is that there is no private right of action under the NCLB. This means that individual rights are not conferred and only the U.S. secretary of education can enforce violations of NCLB. In *Newark Parents Association v. Newark Public Schools* (2008), the U.S. Court of Appeals for the Third Circuit ruled that there was no private right to sue regarding parental notice and tutoring provisions (under NCLB parents are notified of a school's low performance and that the district must provide tutoring services for students in schools in need of improvement). Connecticut was the first state to challenge NCLB as an unfunded mandate (argument that NCLB requires expensive standardized testing but the government does not

pay for this) and other states followed; however, the U.S. Supreme Court refused to hear this argument.

According to Thorndike and Thorndike-Christ (2010), "NCLB-mandated testing programs were developed to document the performance of *schools* in helping students master state objectives . . . it is unlikely that uses of those test scores for decision making about individual students was ever validated" (p. 242). The NCLB Act set forth laudable goals and is designed to improve instruction, ensure academic proficiency, and raise levels of achievement to high standards for all students. In its wake, however, it has crystallized and brought renewed national attention to educational testing concerns that have been with us for decades, including the validity of tests, the use of single measures to make high-stakes decisions, and the potential misuse of test results.

Many of these issues are addressed through professional standards such as the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999), which represents a consensus across professional groups regarding appropriate test use in education and psychology (see Volume 1, Chapter 13, this handbook). The AERA has a position statement on high-stakes testing in pre-K through 12th grade (AERA, 2000) which "sets forth a set of conditions essential to sound implementation of high-stakes educational testing programs" (p. 1). These conditions include issues such as not making decisions on the basis of test scores alone or on the basis of a single test, validating tests for specific uses, ensuring alignment between tests and curriculum, providing students with opportunities to learn and with opportunities for remediation, giving attention to students with special needs or language differences, and continuously monitoring not only the results of such high-stakes testing programs but also the effects (both positive and negative) of such testing.

It is unlikely that many of the issues generated by standardized tests and testing programs will ever be fully resolved, but acknowledgment that testing can carry both positive outcomes and serious consequences for students and educators is important. Concerns will and should continue to exist as test makers constantly strive to improve methods of measurement and generate data that will be most useful for the improvement of individuals and systems.

COMPARISON OF THE FIVE MAJOR LAWS

Table 25.1 is a chart listing the major issues for educational testing that have resulted in much litigation across the five major laws reviewed in this chapter. This table provides a comparison of how each issue is addressed relevant to each law. For example, evaluation procedures and timelines are not specifically identified in Section 504, but they are clearly delineated in IDEA; Section 504 only addresses accommodations, whereas the IDEA addresses accommodations, modifications, and alternate assessments; and disabilities have a much broader definition in Section 504 than the IDEA, which lists 13 specific categories. This table clearly shows that professionals cannot overgeneralize and must pay particular attention to the provisions of each law as violations of one law or one provision may or may not be addressed in another law.

Implications for Practice

This review of laws applied in educational settings, especially as related to testing and the use of assessment results, is certainly not exhaustive, but it does bring out some salient points that have been present for decades. The use of test results to make decisions regarding individuals is not only here to stay but also expands exponentially and affects almost all individuals. Numerous ethical and professional standards guide testing and testing practices (AERA et al., 1999; for a description of these standards, see Volume 1, Chapter 13, this handbook), and in general the tests themselves have withstood scrutiny. The application of high-stakes testing is sanctioned at a national level for public education, and many private schools and almost all postsecondary institutions rely on test results of some sort to make admission decisions. Therefore, because appropriate practice will improve the field of testing and lead to sound decision making for individuals, the chapter

TABLE 25.1

Overview of Federal Regulations That Apply to Testing

Major issues	FERPA	Section 504	ADA	IDEA Part B	NCLB
Purpose	Privacy protections for educational records held by federally funded educational institutions	Civil rights law that protects the rights of individuals with disabilities in programs and activities that receive federal financial assistance	National mandate for the elimination of discrimination against individuals with disabilities; extends 504 to private sector	Federal funding statute to provide financial assistance to states in their efforts to ensure a free appropriate public education for students with disabilities; ages 3–21	Mandated state assessment program to address accountability; data used to measure effectiveness of schools and student learning outcomes
Child find	Not applicable	Recipients must on an annual basis undertake to identify and locate every individual with a disability who is not receiving a public education and notify handicapped persons and their parents or guardians of the recipient's duty; colleges not required to engage in child find	Not applicable	Obligation to identify, locate, and evaluate all students suspected of a disability regardless of severity; includes children who are highly mobile, migrant, homeless, wards of the state, and in private schools	Not applicable
Consent	Signed and dated written consent required for disclosure of personally identifiable information; must indicate to whom disclosure is being made, what records, and for what purpose; there are 15 exemptions or limitations to disclosure; parents have a right to a copy of what was disclosed if they request this	Written, informed parental consent for initial Section 504 evaluation is required; the concept of *informed* consent under Section 504 is the same as described in the IDEA and if a parent refuses consent for an initial evaluation, Section 504 regulations provide that school districts *may*, at their discretion, use due process hearing procedures to override the parents' denial of consent	Not applicable	Informed parental consent required for initial evaluation; reevaluation can be done through review of existing data; if additional evaluation data is required, consent is required; due process allows for override of lack of consent through a hearing	Not applicable; all students must participate in the state assessment program

(*Continued*)

TABLE 25.1 (*Continued*)

Overview of Federal Regulations That Apply to Testing

Major issues	FERPA	Section 504	ADA	IDEA Part B	NCLB
Definition and determination of disability	Not applicable	Physical or mental impairment that substantially limits a major life activity, has a record of such an impairment, or is regarded as disabled by others; no specific list of diseases or conditions	Same as 504; "Substantially limits" is now interpreted as without regard to the ameliorative effects of mitigating measures; definition of disability under ADA has been broadened in scope	Student has one of 13 qualifying conditions and requires special education services as a result of the condition	Not applicable
Evaluation and reevaluation requirements	Not applicable	Evaluation conducted before taking any action with respect to initial placement and any subsequent significant change in placement; no time limit on initial evaluation; must establish procedures for periodic reevaluation (does not indicate timeframe like IDEA); evaluation procedures same as IDEA; admissions tests: (a) are selected and administered so as best to ensure that when a test is administered to an applicant who has a handicap that impairs sensory, manual, or speaking skills, the test results accurately reflect the applicant's aptitude or achievement level or whatever other factor the test purports to measure, rather than reflecting the applicant's impaired sensory, manual, or speaking skills (except where those skills are the factors that the test purports to measure); (b) are designed for persons with impaired sensory, manual, or speaking skills are offered as often and in as timely a manner as are other admissions tests; and (c) are administered in facilities that on the whole, are accessible to handicapped persons	Does not list specific evaluation procedures	Full and individual evaluation in all areas of suspected disability; multidisciplinary team conducts the evaluation; evaluation must be completed within 60 days; reevaluation required at least every 3 years; procedures: must use a variety of assessment tools and technically sound instruments; tests/evaluation materials have been validated for the specific purpose for which they are used, administered by trained personnel in conformance with the instructions provided by their producer; tests are selected and administered so as best to ensure that when a test is administered to a student with impaired sensory, manual, or speaking skills, the test results accurately reflect the student's aptitude or achievement level or whatever other factor the test purports to measure, rather than reflecting the student's impaired sensory, manual, or speaking skills (except where those skills are the factors that the test purports); tests are selected to be nondiscriminatory on racial or cultural basis, administered in native language	Not applicable as testing is universal for all students; testing under NCLB does allow for accommodations and alternate forms for students with disabilities

Access to and review of records and disclosure	Parent or eligible student must be given the opportunity to inspect and review the education record; must comply within a reasonable time but not more than 45 days; if circumstances prevent right to review, copy or other arrangements must be made; prior and written consent required to disclose information	Same as FERPA	Not applicable	Same as FERPA; also indicates the right to have a representative of the parent inspect and review records; must keep a record of parties obtaining access to education records including name of party, date access given, and purpose for which the party is authorized to use the records; parental consent required before disclosure	Same as FERPA; cannot use disaggregated data for subgroups if would reveal personally identifiable information
Nondiscriminatory Procedures	All of these laws have provisions for nondiscriminatory procedures ranging from procedures in initial evaluation, test selection, use of accommodations for testing, and issues regarding determination of disability conditions; disproportionality is a major concern for IDEA disability categories and placements				

Note. FERPA = Family Educational Rights and Privacy Act of 1974; Section 504 = Section 504 of the Rehabilitation Act of 1973; ADA = Americans with Disabilities Act of 1990 as amended in 2008; IDEA = Individuals with Disabilities Education Improvement Act of 2004; NCLB = No Child Left Behind Act of 2001.

closes with the following implications for practice regarding testing in educational settings:

1. Professionals who work in education at all levels must remain knowledgeable about laws and keep up with revisions done through amendments and reauthorizations. In some cases, previous practices can be overturned (e.g., Section 504 and mitigating circumstances) and reauthorizations can change the methods by which disabilities are tested and determined (e.g., SLD and ID). Best practices and legal practices continuously change and educators must keep abreast of these changes.
2. Educators must ensure that all evaluations are conducted in a manner consistent with federal laws. These evaluations need to be comprehensive, consisting of appropriately selected tests that are scored correctly and measure the student's ability and not disability. These evaluations need to be conducted with integrity and by highly trained professionals who know not only about testing but also about disabilities and the legal issues inherent in this field.
3. For individuals identified as having a disability, educational testing must measure the student's knowledge and skills rather than be adversely affected by their disability. Thus, the impact that the disability has on testing and how this impact will be accounted for needs to be identified to determine appropriate accommodations and modifications. It is likely that technology will improve testing based on the concept of universally designed assessment; thus, tests will be designed for universal access by all individuals.
4. The confidentiality of all testing records must be safeguarded and test protocols should not be destroyed. Decisions are based on tests and if such a decision is questioned, the data on which it rests must be available for review. The right to review documentation based on tests is fundamental and facilitates a system of checks and balances when questions arise.
5. Educational decisions must be based on multiple sources of data, not single test scores. Although this is obvious to all professionals who know about the strengths and limitations of tests, decision making based on single test scores occurs at all levels and can have devastating effects (e.g., graduation).
6. Tests themselves must undergo constant scrutiny to ensure that they have adequate content-related validity evidence and that they reflect academic standards delivered through a curriculum as well as what is being taught in that curriculum. If educational testing is to be useful, the tests themselves must be relevant and yield information that will improve the education and instruction for all students.

References

Allegheny (Pa.) Intermediate Unit, 20 IDELR 563 (OCR 1993).

American Educational Research Association. (2000, July). *AERA position statement on high- stakes testing in pre-K–12 education.* Retrieved from http://www.aera.net/policyandprograms/?id=378

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Americans with Disabilities Act of 1990, 42 U.S. C. *et seq.* §12101. Amendment Act of 2008, P.L. 110–325. ADA as Amended. Retrieved from http://www.ada.gov/pubs/adastatute08.htm

Bersoff, D. N. (1996, October). *Rights and responsibilities of test-takers: Legal issues raised by the Joint Committee on Testing Practices.* Invited paper presented at the Joint Committee on Testing Practices Invitational Forum on Test Taker Rights and Responsibilities, Rockville, MD.

Bevis ex rel. D. B. v. Jefferson County Bd. of Educ., 48 IDELR 100 (N. D. Ala. 2007).

Borreca, E. A., & Borreca, C. P. (2008). Assessing students with disabilities under the IDEA and NCLB. In K. E. Lane, M. A. Gooden, J. F. Mead, P. Pauken, & S. Eckes (Eds.), *The principal's legal handbook* (pp. 340–342). Dayton, OH: Education Law Association.

Brazosport ISD v. Student. (2007). Accessed at Texas Education Agency, Special Education Due Process Hearings, Docket No: 127-SE-0107.

Clark County Sch. Dist., 37 IDELR 169 (SEA NV 2002).

Crump v. Gilmer Independent School District, 797 F. Supp. 552 (1992).

D. B. v. Bedford County Sch. Bd., 54 IDELR 190 (W. D. Va. 2010).

Debra P. v. Turlington, 474 F. Supp 244 (M. D. Fla. 1979).

Debra P. v. Turlington, 644 F. 2d 397 (5th Cir. 1981). Retrieved from http://openjurist.org/644/f2d/397/debra-v-d-turlington

Department of Educ. v. Cari Rae S., 35 IDELR 90 (D. Haw. 2001).

Ellenberg v. New Mexico Military Institute, 572 F. 3d 815 (10th Cir. 2009).

E. M. by E. M. and E. M. v. Pajaro Valley Unified Sch. Dist., 53 IDELR 41 (N. D. Cal. 2009).

Erik V. v. Causby, 977 F. Supp. 384 (1997).

Etscheidt, S. (2003). Ascertaining the adequacy, scope, and utility of district evaluations. *Exceptional Children*, *69*, 227–247.

Family Educational Rights and Privacy Act of 1974, 20 U.S. C. §1232 *et seq.* Implementing regulations at 34 CFR 99. Retrieved from http://www2.ed.gov/policy/gen/guid/fpco/pdf/ferparegs.pdf

Federal Register. (2006). *Analysis of comments and changes to 2006 IDEA Part B regulations*, 71 Fed. Reg. 46632, 46648. U.S. Department of Education.

Garcia v. Northside Indep. Sch. Dist., 47 IDELR 6 (W. D. Tex. 2007).

G. J. v. Muscogee County School District, 54 IDELR 76 (M. D. Ga. 2010).

Individuals With Disabilities Education Improvement Act of 2004, P.L. 108–446, 20 U.S. C. 1400, *et seq.* 2006. Implementing regulations at 34 CFR Part 300. Retrieved from http://idea.ed.gov

J. H. v. Northfield Pub. Sch. Dist., No. 0659–01 52 IDELR 165 (Minn. Ct. App. 2009, unpublished).

Kubiszyn, T., & Borich, G. (2007). *Educational testing and measurement: Classroom application and practice* (8th ed.). New York, NY: Wiley.

Lake County (FL) Sch. Dist., 52 IDELR 139 (OCR 1/08).

Larry P. v. Riles, 343 F. Supp. 1306 (N. D. Cal 1972), *aff'd.*, 502 F. 2d 963 (9th Cir. 1974), *further proceedings*, 495 F. Supp. 926 (N. Cal. 1979), *aff'd.*, 793 F. 2d 969 (9th Cir. 1984), amended 1986.

Letter Re: Greater Clark County School District, 10 FAB 6 (FPCO 2006).

Letter to Fazio, 21 IDELR 572 (OSEP 1994).

Letter to Williams, 20 IDELR 1210 (OSEP 1993).

McClung, M. S. (1979). Competency testing programs: Legal and educational issues. *Fordham Law Review, 47*, 651–712.

National Board on Educational Testing and Public Policy. (2002). *About the Board*. Retrieved from http://www.bc.edu/research/nbetpp/about_general.html

N. B. and C. B. ex rel. C. B. v. Hellgate Elementary Sch. Dist., 108 LRP 51033 (9th Cir. 2008).

Newark Parents Association v. Newark Public Schools, 547 F. 3d 199 (C. A. 3, Nov. 20, 2008). Retrieved from http://www.ca3.uscourts.gov/opinarch/074002p.pdf

Newport-Mesa Unified School District v. State of California Department of Education, 43 IDELR 161 (C. D. Ca. 2005).

No Child Left Behind Act of 2001, 20 U.S. C. §6301 *et seq.* Implementing Regulations at 34 CFR Part 200. Retrieved from http://www2.ed.gov/nclb/landing.jhtml

Norlin, J. W. (2008). *Confidentiality in student testing: Access and disclosure requirements under FERPA and the IDEA*. Horsham, PA: LRP.

North Rockland (NY) Cent. Sch. Dist., 109 LRP 27208 (OCR 7/08).

Office of Special Education Programs. (2007, April). *Memorandum to state directors of special education*. Retrieved from http://www2.ed.gov/policy/speced/guid/idea/letters/2007-2/osep0709disproportionality2q2007.pdf

Office of Special Education Programs. (2009). *Model form: Procedural safeguards notice*. Retrieved from http://www2.ed.gov/policy/speced/guid/idea/modelform-safeguards.doc 09.disproportionality of racial and ethnic groups in special education.doc

Office of Special Education Programs. (2011, January). *Memorandum to state directors of special education*. Retrieved from http://www2.ed.gov/policy/speced/guid/idea/memosdcltrs/osep11-07rtimemo.doc

Pasatiempo v. Aizawa, 25 IDELR 64 (9th Cir. 1996).

Pearson Assessments. (2006). *Legal policies effective January 1, 2006*. Retrieved from http://www.pearsonassessments.com/haiweb/Cultures/en-US/Site/general/LegalPolicies.htm

Plainedge (NY) Union Free Sch. Dist., 46 IDELR 137 (OCR 2006).

Powell v. National Board of Medical Examiners, 364 F3d 79 (2nd Cir. 2004). Retrieved from http://openjurist.org/364/f3d/79/powell-v-national-board-of-medical-examiners

Regional (CT) Sch. Dist. No. 17, 47 IDELR 49 (OCR 5/06).

Regional Sch. Dist. No. 9 Bd. of Educ. v. Mr. and Mrs. M. ex rel. M. M., 53 IDELR 8 (D. Conn. 2009).

Richard S. v. Wissahickon Sch. Dist., 52 IDELR 245 (3rd Cir. 2009, unpublished).

Rooker, L. S. (1997, October 2). *FERPA memorandum: Access to test protocols and test answer sheets*. Retrieved from http://www.fetaweb.com/04/ferpa.rooker.ltr.protocols.htm

Rothberg v. Law School Admission Council, Inc. (2004). Retrieved from http://federal-circuits.vlex.com/vid/rothberg-law-school-admission-council-18497626

Rush v. National Board of Medical Examiners, 268 F. Supp. 2d 673 (N. D. Texas 2003).

Russo, C. J., & Osborne, A. G. (2009). *Section 504 and the ADA*. Thousand Oaks, CA: Corwin Press.

S. A. by L. A., & M. A. v. Tulare County Office of Education, 53 IDELR 111 & 143 (2009).

School District U-46, 45 IDELR 74 (SEA IL 2005).

Section 504, Rehabilitation Act of 1973, 29 U.S. C. §701 *et seq.* Implementing regulations at 34 CFR Part 104. Retrieved from http://www.ed.gov/policy/rights/reg/ocr/edlite-34cfr104.html

Slater, A. (2010). *The special education desk book*. Horsham, PA: LRP.

Springfield (MA) Pub. Schs., 53 IDELR 30 (OCR 2008).

Springfield Sch. Committee v. Doe, 53 IDELR 158 (D. Mass 2009).

St. Johnsbury Academy v. D. H., 240 F. 3d 163 (2nd Cir. 2001).

Student bnf Parent v. Humble ISD. (2010). Accessed at Texas Education Agency, Special Education Due Process Hearing Decisions, Docket No. 193-SE-0410.

Student bnf Parent v. Northwest ISD. (2009). Accessed at Texas Education Agency, Special Education Due Process Hearing Decisions, Docket No. 057-SE-1108.

Sutton v. United Airlines, Inc., 527 U.S. 471 (1999).

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Boston, MA: Pearson Education.

Toyota Motor Manufacturing, Kentucky, Inc. v. Williams, 534 U.S. 184 (2002).

U.S. Department of Education. (2007). *Decision letters on each state's final assessment system under NCLB*. Retrieved from http://www.ed.gov/adm.us/lead/account/nclbfinalassess/index.html

Washoe County School District, Nevada State Education Agency, 13 FAB 10, 109 LRP 78026 (2009).

W. H. by B. H. and K. H. v. Clovis Unified School District, 52 IDELR 258 (E. D. Cal. 2009).

Yell, M. L., & Drasgow, E. (2005). *No Child Left Behind: A guide for professionals*. Upper Saddle River, NJ: Prentice-Hall.

Yukon (OK) Pub. Schs., 50 IDELR 199 (OCR 2007).

Zirkel, P. A. (2011). What does the law say? *Teaching Exceptional Children*, *43*(3), 65–67.

Zirkel, P. A., & Thomas, L. (2010). State laws for RtI: An updated snapshot. *Teaching Exceptional Children*, *42*(3), 56–63.

PART III

FUTURE DIRECTIONS

CHAPTER 26

ADAPTING TESTS FOR USE IN OTHER LANGUAGES AND CULTURES

Kadriye Ercikan and Juliette Lyons-Thomas

The use of multiple-language versions of tests is not only desirable but necessary for many tests that involve individuals from different languages and cultures. This chapter describes and discusses issues related to adapting tests from one language and culture to another. It presents steps to adapt tests and to examine and establish comparability of source- and target-language versions of tests.

In North America, educational and psychological tests are administered to individuals from many cultures whose first language may not be English and who may not be fully proficient in English. These include intelligence tests, personality tests, diagnostic tests for determining special needs and assigning individuals to special programs, tests for employee selection, and screening tests administered to children or adults from different language and cultural backgrounds. When tests are intended to measure educational and psychological constructs independent of individuals' language proficiency, limited language proficiency of these individuals may interfere with valid measurement of the intended constructs. Therefore, tests are often adapted to many different languages to provide valid measurement and minimize bias. Adaptation of tests is also necessary for use in other cultures and countries and as part of cross-cultural research and international comparisons.

In this chapter, we make a distinction between test translation and test adaptation. A common term used for creating different language versions of tests is *test translation*. In practice, however, this test creation process involves more than linguistic translation of tests. It involves adapting tests to be appropriate for the culture for which the tests are intended. For example, adaptation may include changing the temperature metric from Fahrenheit to Centigrade, or changing names of places or people to those with which the examinees would be more familiar. In addition, test adaptation refers to the broader process of creating different language versions of tests. Hambleton (2005) described this distinction as follows:

> Test adaptation includes all the activities from deciding whether or not a test could measure the same construct in a different language and culture, to selecting translators, to deciding on appropriate accommodations to be made in preparing a test for use in a second language, to adapting the test and checking its equivalence in the adapted form. Test translation is only one of the steps in the process of test adaptation and even at this step, adaptation is often a more suitable term than translation to describe the actual process that takes place. (p. 4)

Throughout the chapter we refer to the original version of the test as the *source* version and the adapted test as the *target* version.

The validity of measurements and comparisons using adapted tests critically depend on the degree to which the adapted versions of tests indeed measure the intended constructs and provide comparable measurements. Proper adaptation of a test to a

DOI: 10.1037/14049-026
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

target language and culture requires many carefully implemented steps. In practice, however, psychologists, educators, and researchers do not necessarily follow these steps to create adapted versions of tests and may resort to different practices that do not include testing the participants in a language they understand and can perform in. López and Romero (1988) reported the following practices in testing Spanish-speaking participants:

> (a) administering the instrument in English and attempting to take language differences into account when interpreting the scores, (b) administering only the performance subtests, using either the English or Spanish instructions, (c) using an interpreter, or (d) referring the testing to a Spanish-speaking colleague or assistant who can translate instructions and test items during the test administration. (p. 264)

Any of these practices are problematic and may lead to significant adverse implications for examinees who take tests in a language in which they are not proficient. Some of these adverse effects include underestimation of competency of individuals and inappropriate labeling and diagnosis. Alderman (1981) found that Latino students' aptitudes were seriously underestimated on the SAT if the students were not proficient in English. In their study on science assessment, Solano-Flores, Ruiz-Primo, Baxter, and Shavelson (1992) found that the least English-proficient students had difficulty coping with the English-only version of the assessment and that students who used Spanish to respond to test items performed better than their linguistic counterparts who responded in English. August and Hakuta (1997) reported that low test scores received by bilinguals often were interpreted as evidence of deficits or even disorders. Other researchers have identified the language gap in testing as a major contributor to the disproportionate numbers of Latino bilinguals diagnosed as "mentally retarded" when intelligence test scores were used (Duran, 1988; Rueda & Mercer, 1985). Latino students in Riverside California, who constituted less than 10% of the school population at that time, accounted for 32% of the students identified as mentally retarded. For most of these students (62%), such decisions were based solely on low intelligence test scores (Rueda & Mercer, 1985).

During the past 2 decades, two sets of guidelines that identify the necessary development and verification steps for proper test adaptation have been developed. The *Standards for Educational and Psychological Testing* was developed jointly by the American Educational Research Association (AERA), American Psychological Association, and National Council on Measurement in Education (1999; for information on the test standards, see Volume 1, Chapter 13, this handbook). The second set of guidelines was developed by the International Test Commission (ITC; Hambleton, 2005; ITC, 2010). These standards and guidelines play significant roles in guiding the adaptation process as well as in evaluating quality of test adaptation processes. Another influence on the progress in test adaptation has been the research on comparability of adapted versions of tests. This research on international educational achievement tests, cross-cultural psychological tests, multilingual versions of licensure tests, and many others, demonstrated great degrees of incomparability between adapted versions of tests and the importance of the quality of the adaptation process on validity of measurement and comparability of scores (Ercikan, 1998; Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; Hambleton, 2005; Oliveri & Ercikan, 2011).

There is also evidence that developing tests for different language groups has not changed much during the past several decades. Merenda (2005) presented a historical perspective on cross-cultural adaptation in educational and psychological testing that draws attention to similarities between the procedures used 40 years ago and the present, after the publication of guidelines and tremendous research on test adaptations.

One practice that persists in the 21st century is the simple translation from one language into another without any attention given to verification of comparability of the two language versions and appropriateness of the test in the target language and culture. This approach can result in score incomparability between the two language versions

of an instrument and can jeopardize validity of interpretations of test scores.

Another common problem is negligence with regard to the appropriateness of the original test norms for the target culture. Often, those administering a test that has been adopted for another culture will interpret scores based on the original norms. Vital practices such as modifying items, restandardizing testing procedures, and investigating construct validity in the target culture are frequently neglected, even in the presence of guidelines that are meant to curb these oversights (Merenda, 2005; for another perspective on this task, see Volume 2, Chapter 11, this handbook).

The degree to which adaptations can deviate test language versions from each other are described by Maldonado and Geisinger (2005). These authors reviewed research that addresses comparability of the Wechsler Adult Intelligence Scale (WAIS) and Escala de Inteligencia Wechsler para Adultos (EIWA), which are English and Spanish versions of an intelligence test. When WAIS was adapted from English to Spanish, resulting in EIWA, one WAIS question was replaced by a completely different question. In the same adaptation, the Information subtest in English has 29 items on the WAIS, and the examiner stops testing after five consecutive errors, whereas the EIWA contains 32 Information items, and the examiner continues until seven consecutive errors are made. A third difference between the two language versions of these tests was identified by Melendez (1994), who documented that identically translated answers to test questions received one point on the EIWA and no points on the WAIS. Many researchers investigated the potential reasons for higher scores for Spanish students based on EIWA compared with the English speakers who took the WAIS, even though any one of the three differences created by these adaptation would have been sufficient to declare incomparability between WAIS and EIWA.

Interest is growing in cross-cultural research, international comparisons, and a greater level of globalization in the business world. In recent years, these developments have led to test adaptations becoming very important components of testing and research in these contexts. Cross-cultural research is now common in education as is evident by the dozens of countries participating in international assessments of educational achievement or learning outcomes such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment. In the business world, training and licensing of professionals in different countries requires licensure tests to be administered in multiple languages (Fitzgerald, 2005; Robin, Sireci, & Hambleton, 2003; Sireci, Yang, Harter, & Ehrlich, 2006). In psychological research, interests in understanding aptitude, personality, and other psychological tests are growing with the increased multicultural and globalization throughout the world. There is greater acknowledgment of multilingualism in different countries around the world, and in these countries, such as in Canada with English and French as official languages, the majority of tests are administered in the official languages. Proper test adaptation in these contexts plays a critical role in providing data that be used for meaningful interpretation of cross-cultural comparisons, setting similar standards for licensing professionals from different countries and for valid measurement in general.

TO ADAPT TESTS OR DEVELOP NEW TESTS

One of the key questions in cross-cultural research is whether to develop new tests or adapt existing ones for the target language and cultural group. Several factors affect which of these two options is preferable. One obvious context that would be preferable when adapting an existing instrument is when tests are intended for language groups within the same country, such as for testing French and English speakers in Canada and testing English language learners in the United States. In these contexts, adapted versions of tests are necessary, for example, for testing for the same learning outcomes to establish similar standards for different language groups, for establishing comparable pass–fail decisions in licensure examinations, or for establishing similar diagnostic criterion in clinical settings. Developing different assessments for different language groups may result in scores,

standards, cut scores, and diagnostic decisions that are not comparable for different language and cultural groups.

Several other factors may lead to adapting tests instead of developing new ones. Developing new tests requires tremendous amounts of resources, time, and expertise. Test development, in particular those for assessing psychological constructs, are based on years and decades worth of theoretical research to develop an understanding of the construct in the target culture. For example, if empirical evidence suggests that a construct such as motivation is understood and operates differently in a particular culture, then empirical research is needed to support these different conceptualizations of motivation. Even when a decision is made to base the test on the same theoretical conceptualization as the source version of the test, development of new test items, piloting, field-testing, and norming take years and a great deal of resources.

Another factor that might affect the "adapt or develop a new test" decision is whether data from tests are going to be used for cross-cultural research. Comparisons of research findings in different cultural settings critically depend on the degree to which data in such research are based on comparable test scores. In most cases, adapted versions of tests would be expected to be more similar to each other and lead to more comparable test scores, which make them necessary aspects of cross-cultural research.

An additional reason for using adapted versions of tests instead of developing new tests is the existing validity and reliability evidence for the original version of the test. This evidence provides the users of the adapted versions of the tests some information about the kinds of psychometric properties they can expect from these tests. This evidence, however, may create a false sense of assurance about adapted tests because validity and reliability evidence for original versions of tests may not hold for the adapted versions and for the cultural groups for which the tests are intended.

There are some disadvantages to developing adapted versions of existing tests instead of developing new ones tailored to the cultural and language groups. The key disadvantage is that the adapted versions of tests may not provide valid measurement of the construct in the target language and culture. In particular, this finding can happen when the original test is based on older notions and conceptual understanding of the construct. For example, as Jackson (1991, as cited by Hambleton & Patsula, 1998) described, "many of the most popular measures of personality were developed in an earlier era when our understanding of personality measures was in its infancy and conceptual, quantitative and technological support for test construction was relatively primitive" (p. 156). Therefore, neither the source nor the target versions of such personality tests can be expected to provide valid measurement of personality. In addition, the operational definitions of the construct should not be assumed to be valid in different cultures and languages. Spielberger, Moscoso, and Brunner (2005) demonstrated large degrees of cultural differences in such tests. These researchers emphasize the nonequivalence of psychological constructs in different cultures and a need to develop new tests tailored to the target language and culture.

Chapter Outline

The remainder of the chapter consists of three main sections. The first section reviews and discusses (a) standards for educational and psychological testing, (b) ITC test adaptation guidelines, (c) test development and adaptation processes, (d) test adaptation errors, (e) measurement equivalence, and (f) score comparability. The following section covers empirical evidence for measurement equivalence, including psychometric evidence for test equivalence, sources of differential item functioning, and measurement units and scalar equivalence. The chapter concludes with a summary of steps and recommendations for developing adapted tests.

DEVELOPING ADAPTED VERSIONS OF TESTS

One of the basic challenges in adapting instruments into different languages is the intrinsic differences between languages. Previous research identified several differences between languages that cause problems in test adaptation, including the following:

(a) variations in the frequency of word use and in word difficulty, (b) words that may be commonplace and "easy" in one language may not be equally so in another language, (c) grammatical forms either do not have equivalents or else have many of them in another language, and (d) syntactical style is one of the most difficult features to carry over from one language to another (Ercikan, 1998). Additional reasons for linguistic incomparability can be derived from language philosophy. According to language philosophy, the sense of a word is a function of language as a whole rather than of a single definition, and language and social life are integrally interconnected (Derrida, 1998; Heidegger, 1996; Wittgenstein, 1958). As a consequence of this interconnection, creating equivalent meaning and function in life may not be possible in the translation process (Benjamin, 1972; Derrida, 1986; Ricœur, 2004). In different languages, simple words can elicit different semantic relations that affect the trajectory of thought processes of examinees reading test questions (e.g., Ercikan et al., 2010; Ercikan & Roth, 2006; Roth, 2009).

Even with these challenges, there are many benefits to developing adapted versions of tests instead of developing new ones in the target language. This section discusses four issues that are essential aspects of adapting educational and psychological tests from one language to another. These are (a) test adaptation processes, (b) translation errors, (c) measurement equivalence, and (d) score comparability. In considering these issues, it is important to draw attention to two sets of testing and measurement standards and guidelines that address test adaptation. We discuss the *Standards for Educational and Psychological Testing* (AERA et al., 1999) and the *ITC Guidelines for Test Adaptation* (ITC, 2001), both of which are essential for test developers and clinicians adapting instruments from one language to another to be familiar with.

Standards for Educational and Psychological Testing

The *Standards for Educational and Psychological Testing* (the *Standards*; AERA et al. 1999) are intended to guide and set standards for designing, developing, and using a variety of tests, including educational, psychological, and professional licensure tests. Four of the *Standards*[1] are particularly relevant to test adaptation:

- Standard 9.4 highlights the need for test publishers to explain and provide justification for linguistic modifications that they deem to be appropriate in specific situations. These modifications should be taken into account in score interpretations.
- Standard 9.5 recommends that if there is evidence that scores are not comparable across multiple versions of tests, additional information should be provided to help test users in correctly interpreting test scores.
- Standard 9.7 calls attention to the need to describe the approaches used in establishing the adequacy of translation, and empirical and logical evidence should be provided for score reliability and the validity of the inferences based on the target test for all linguistic groups.

Consider the case of adapting an instrument developed in English into Spanish for various cultural groups. If the instrument is meant to be used with Mexican, Cuban, Spanish, and other Spanish-speaking subgroups, it is the responsibility of the test developer to provide independent reliability and validity evidence for each of those subgroups.

Standard 9.9 recommends that test developers provide evidence of the comparability of different language versions of a test. For example, the test developer should present evidence that the same construct is being measured in both tests.

In addition, the *Standards* (AERA et al., 1999) recommends against using back translation, which involves comparisons of the source version with the target-to-source translation, as a sole method for verifying linguistic comparability. The comparability of source and the back-translated target versions is not sufficient evidence that the two language versions have the same meaning and provide similar information to examinees. The *Standards* cautions against using interpreters to administer tests that

[1]New *Standards* are forthcoming as this handbook is being published.

have not been properly adapted to the target language. An interpreter or translator who may not be familiar with proper testing procedures or purposes of testing may lead to inadequate translation and adaptation of the test and inappropriate test administration. The *Standards* also point out linguistic or cultural differences that may cause different response patterns among test takers that need to be taken into account. For example, individuals from some cultures may be reluctant to provide lengthy or elaborate answers to interviewers who are considered to be of a higher status or maturity or may be more timid about indicating confidence and success level or disclose personal information. Interpretations of scores across different cultural or linguistic versions of tests should account for these differences.

ITC Guidelines for Test Adaptation

In 1992, the ITC initiated a project to develop guidelines for translating and adapting educational and psychological tests. During a period of 7 years, the guidelines were developed by a group of 12 psychologists and were first published in draft form in 1994 (Hambleton, 1994). In 2001, they were further elaborated on and published in an ITC report (ITC, 2001). Presently, the current guidelines can be found online (ITC, 2010). These guidelines address (a) *Context* (C1–C2), (b) *Test Development and Adaptation* (D1–D10), (c) *Administration* (A1–A6), and (d) *Documentation/Score Interpretations* (I1–I4). The guidelines in each of these categories are discussed in the following paragraphs.

The two guidelines with respect to context highlight the effects of cultural differences on equivalence of measurements in language and cultural groups. Possible cultural differences in C1 may include differences in motivation, experience with psychological tests, and speededness. C2 on the other hand addresses similarity of definition and operationalization of the construct for the language groups. It is important to distinguish between the conceptual and operational aspects of the construct. The evidence of similarity of conceptual definition of the construct needs to be based on theoretical grounds, such as whether a construct like self-esteem is conceptualized the same way in the language and culture groups. The equivalence of the operationalization of the construct by the test items in the source and target languages can be examined using statistical analyses, such as structural equation modeling, confirmatory or exploratory factor analysis, discussed in the next section.

The test development and adaptation guidelines emphasize both the *appropriateness* of adapted versions of tests for the intended cultures and populations as well as the *equivalence* of the adapted versions of tests. The first guideline (D1) for adaptation and development highlights the importance of using translators who are not only proficient in the translation languages but also knowledgeable about the culture for which the target test version is intended. Guidelines D2 to D4 emphasize the need for test developers to provide evidence of linguistic equivalence in all materials related to testing (test questions, instructions, scoring rubrics, etc.), familiarity with testing method, format, content, and stimulus materials to the target population. The last set of guidelines, D5 to D10, requires test developers to provide evidence of equivalence between the different language versions using judgmental reviews of linguistic and construct equivalence, statistical and psychometric evidence of item and test equivalence, and validity evidence for the target populations.

Test administration–related guidelines draw attention to the fact that tests are only one component of testing, and administration of tests and conditions of test administration are critical components of testing that may affect validity of interpretations and comparability of scores from different language versions of tests. As a set, these guidelines underscore that valid measurement requires administration procedures to minimize interference of factors that may affect examinee's ability to respond to the test questions in a way that accurately reflects their abilities, opinions, and psychological state. Even when tests are properly adapted to the intended languages and cultures, problems in administration procedures can affect how examinees respond to test questions and jeopardize the validity of measurement as well as comparability of measurement across the adapted versions of tests.

The documentation and score interpretations guidelines are at the core of the validity of comparisons

between scores from the different language versions of tests. First, any changes between language versions need to be documented so that users of tests can take these changes into account in interpreting scores. Second, like any test score, differences in test scores between adapted versions of tests need to be validated with empirical evidence. The meaningfulness of interpretations of score differences depends on the degree of comparability and measurement invariance between the adapted versions of tests. Finally, test developers need to provide information about the sociocultural and ecological context (such as examinee motivation and importance and use of test results) to provide a context for interpretation of results. For examples of how each of these guidelines have affected development of adapted versions of tests and descriptions of how each of these issues have been addressed in practice the reader, see Hambleton (2001).

Test Development and Adaptation Processes

Several different test adaptation processes have been investigated and practiced. These include *parallel*, *successive*, *simultaneous*, and *concurrent* development of different language versions of tests. Appropriateness and effectiveness of these methods varies depending on the purpose and degree of comparability required. *Parallel* development involves having different language versions developed by experts from each language group based on a common test blueprint with sections from each language version adapted to the other language. This approach results in two tests designed to be assessing the same construct but each originates in the language for which it is targeted and is developed by content experts from the target cultural group, except for a small portion of the test which is adapted from another language. This process is similar to developing parallel tests in a single language. The adapted portion of the test needs to be reviewed for linguistic equivalence in the two languages. In addition, the comparability of different language versions of such tests need to be verified by examining content and psychometric comparability similar to the way single language parallel tests are examined (Grisay et al., 2007; Organisation for Economic Co-operation and Development, 2005).

Successive test adaptation is the most commonly used method wherein the test is developed in a source language and one or more bilingual translators adapt the test to the target language and culture using the translation method, that is translation from source to the target language. Bilingual experts review the test to improve the match of the test in the target language to the one in the source language (Rogers, Gierl, Tardif, Lin, & Rinaldi, 2003; Tanzer, 2005).

In successive test adaptation, tests are developed for one culture and are adapted to other cultures later. Therefore, the conceptualization of the construct being assessed is based on one culture, the wording of test items, the actual items included in the test, how they should be evaluated, and how they relate to the construct. These items are all based on the culture for which the test is originally developed. To decenter this language and cultural bias, Tanzer (2005) described and discussed a *simultaneous* multilingual test development. In simultaneous test construction, the emphasis is on the use of a multidisciplinary committee of experts in the targeted languages, in psychometrics, and in the content domain for developing test items (Lin & Rogers, 2005; Tanzer & Sim, 1999). Items are developed by bilingual item writers in one language and are immediately adapted into the other language. Tanzer argued that in simultaneous development, errors resulting from (a) measurement artifacts such as poor item translation or ambiguities in the original item content or (b) genuine "culture specifics" such as low familiarity or appropriateness of the item content in certain cultures may be reduced. In simultaneous test development, culturally incompatible test designs, test instructions and administration procedures, culture-specific images, and linguistic subtleties in the meaning of distractors may be detected more readily than after a source test has been developed.

Similarly Solano-Flores, Trumbull, and Nelson-Barber (2002) proposed a concurrent test development model for multiple language versions of tests to promote the development of equitable tests. Both the simultaneous and concurrent test development models are offered as an alternative to the traditional approach of translating tests originally created in a

single language, typically English, in the North American context. Problems associated with the traditional model of single source-language versions of adapted tests are described as follows:

> Serious theoretical, methodological, and practical limitations of test translation result from two facts. First, simple translation procedures are not entirely sensitive to the fact that culture is a phenomenon that cannot be dissociated from language. Second, student performance is extremely sensitive to wording, and the wording used in the translated version of an assessment does not undergo the same process of refinement as the wording used in the assessment written in the original language. (p. 107)

Concurrent test development utilizes *shells* or *templates* that define item structure and cognitive demands of each item. Using these templates for item development, the two linguistic groups work jointly in all stages of the test development process. This approach of test development is most applicable in the development of extended-response items in cognitive tests.

Decisions about the four development processes—parallel, successive, simultaneous, and parallel—should be made on the basis of the prioritization of comparability and cultural authenticity. The four development processes have trade-offs between *comparability* and *cultural authenticity* of adapted tests. Although the concurrent development prioritizes cultural authenticity, successive development prioritizes comparability and simultaneous and parallel development target a compromise between comparability and cultural authenticity. In a test development context, which of these two aspects is prioritized should depend on the cultural differences expected in the conceptual and operational definition of a construct and on the purposes of and stakes associated with the adapted tests. If large degrees of differences are expected between the construct in different cultures, such as in personality and emotional state assessments (Spielberger, Moscoso, & Brunner, 2005), cultural authenticity should be prioritized. In high-stakes testing contexts, such as in licensure examinations, comparability of pass–fail decisions and cut scores across different language versions will require tests to be parallel to each other, and comparability needs to be prioritized. If test scores do not need to be directly compared, such as in development of research data collection tools (e.g., surveys of classroom climate or student attitudes) for independent research, it is preferable to prioritize cultural authenticity in the test development process.

Many aspects of development of adapted versions of tests affect the equivalence of tests in different languages. Some decisions and choices made with regards to test development, adaptation, and adaptation verification processes can make a significant difference in the equivalence of adapted versions of tests. Three key factors are described in the following sections: (a) developing translatable tests, (b) selection and training of translators, and (c) adaptation verification procedures.

Developing translatable tests. The first factor is taking special precautions at the outset to maximize the suitability of the test for adapting to different languages. Otherwise, problems in the selection of content, format, and other aspects of tests may need to be overcome during the test adaptation process to make the test suitable to the target language or culture. Brislin, Lonner, and Thorndike (1973, as cited by Ercikan, 1998) presented the following guidelines to help others write translatable English that can be useful to consider when developing tests that are intended to be translated to other languages:

(1) Use short, simple sentences of fewer than 16 words.
(2) Employ the active rather than the passive voice.
(3) Repeat nouns instead of using pronouns.
(4) Avoid metaphors or colloquialisms. Such phrases are the least likely to have equivalents in the target language.
(5) Avoid the subjunctive mode, for example, verb forms with "could" or "would."

(6) Avoid adverbs and prepositions telling "where" or "when" (e.g., "frequent," "beyond," "upper").
(7) Avoid possessive forms where possible.
(8) Use specific rather than general terms (e.g., the specific animal, such as cows, chickens, pigs, rather than the general term "livestock").
(9) Avoid words that indicate vagueness regarding some event or thing (e.g., "probably" and "frequently").
(10) Avoid sentences with two different verbs if the verbs suggest different actions. (Ercikan, 1998, p. 544)

Another key consideration should be the appropriateness of the testing format for both the source and the target culture. Although multiple-choice, true-false, or Likert-scale formats are familiar to most populations in North America, this may not be the case in other cultures and countries where testing and surveying are not part of schooling or everyday experiences of individuals. Even though open-ended questions may seem to get around this potential format familiarity difference, examining and establishing the equivalence of interpretations of open-ended responses is challenging.

Selection and training of translators. The formation of the review panel is critical to the effectiveness of the review process. Review panels should consist of between four to eight individuals who (a) have their first language in the target or source language to attend to the subtleties and nuances in the target language in the translation process, (b) are proficient in both languages, (c) are familiar with the target language, (d) have relevant content expertise (e.g., teaching and learning in the subject for educational achievement tests), and (e) understand the basic principles of test construction (Ercikan, Simon, & Oliveri, 2012).

Evaluation of equivalence by expert reviewers. Translations can affect the meaning and functions of single words, sentences, and passages, the content of the items, and the skills measured by the items. The degree and manner in which item features are changed during translation will determine whether the equivalence of items is maintained. Changes in any of these item features may alter its difficulty or even what is being measured. For verification of equivalence of different language versions of tests, expert reviews are conducted. These involve two key types of reviews: (a) based on content and (b) cultural and linguistic reviews. Once the multiple language test versions are developed, content reviews are conducted to establish content and construct-related validity evidence (Bowles & Stansfield, 2008). Content reviews include appropriateness of test items for the language groups that may be exposed to different curricula and instruction in cognitive tests and appropriateness of item content to capture construct-related responses more broadly. Cultural and linguistic reviews are conducted to determine cultural relevance and equivalence of meaning, cognitive requirements, difficulty of vocabulary and expressions, and cues given to help examinees solve the problem in cognitive tests. Reviews of items include (a) word difficulty; (b) semantic differences; (c) item format; (d) item content; (e) idiomatic relationship; (f) grammatical form or syntactic differences; (g) reading level; (h) sentence structure; (i) familiarity with vocabulary; (j) omissions or additions that may affect meaning; and (k) overall format—punctuation, capitalization, typeface, and structure (Allalouf, 2003; Ercikan 1998, 2002; Gierl & Khaliq, 2001; Oliveri & Ercikan, 2011).

Expert reviews involve bilingual experts reviewing both language versions of items and evaluating equivalence. The following linguistic review steps were identified by Ercikan et al. (2004) to ensure comparability of items in the source and target languages: (a) group review of sample items to discuss review criteria, (b) independent review of each item by individual reviewers, and (c) group discussion and consensus for rating adaptation differences between two language versions. The equivalence can be systematically evaluated by expert reviewers by using rating sheets or equivalence evaluation criteria. An example of an equivalence checklist is

included in Appendix 26.1. This checklist includes reviews of the following:

1. *Differences in cultural relevance*: Is the item content more relevant to one group than the other? Example: The passage of a problem-solving item contains food or cultural events that are more relevant or familiar to one group of examinees than the other.
2. *Differences in the actual meaning of an item:* Was the meaning of an item changed in the adaptation process? Example: In English, an item asked to compute the kilograms of apples in *each of two boxes*. When translated in French, the item asked to compute the kilograms of apples in *each one of the two boxes*, which might have caused some confusion (e.g., about whether or not to add the amounts for each box).
3. *Differences in the item format*. Are there differences in punctuation, capitalization, item structure, typeface, and other formatting usages that are likely to influence the performance for one group of examinees? Example: In an English version, a problem-solving item provided a tree diagram at the left edge of the page, whereas in Korean, it was located in the center of the page. The location of the diagram at the left edge of the page may have clued the English-speaking examinees in on the answer, which was to expand the diagram to the right.
4. *Omissions or additions that affect meaning*: Are there omissions or additions of words, phrases, or expressions that may influence the meaning of an item or the performance of one group of examinees? Example: The English form of an item contained the expression "this number written in standard form" whereas the Korean version only mentioned "이숫자는" (i.e., "this number is"), omitting the idea of *standard form*.
5. *Differences in verb tense*: Is the verb tense different in one language version from the other? Example: In the English version, the verb *read* was presented in the present tense, whereas in the Korean version, it was presented in past tense.
6. *Differences in the difficulty or commonness of vocabulary*: Is a certain vocabulary word more difficult or less common in one group than the other? Example: In the English version, the word *burn* was used, whereas in the French version, it was presented as *combustion*, which is more difficult and less frequently used than the word *burn*.
7. *Exclusion or inappropriate translation of key words*: Is any key word that provides clues to guide examinees' thinking processes excluded or inappropriately translated for one group of examinees? Example: In the English version, the stem of a question stated, "whenever scientists carefully measure *any* quantity many times, they expect that . . ." The correct answer was "most of measurements will be close but not exactly the same." However, the French version asked, "when scientists measure the *same* quantity many times, they expect that . . ." The word *same* in the French version could have led the examinees to think that this quantity was known and the answer should be that the scientists should get the same amount each time.
8. *Differences in additional information provided to guide examinees' thinking process*: Are there differences in additional information given to guide examinees' thinking processes? Example: In the English version, a question asked, "At what point will the reflection of the candle *appear to be*?" The French version asked, "At what point will the image of the candle *seem to appear to be*?" The French version is more informative suggesting that the reflection in the mirror may seem different than the actual object. This additional information could make the item easier for French-speaking examinees.
9. *Differences in length or complexity of sentences*: Are there differences in the length or complexity of sentences between the two language versions? Example: Although the translation was accurate, the sentence became shorter or longer or the sentence content became easier or harder to understand.
10. *Differences in words, expressions, or sentence structure inherent to language and or culture:* Are there differences between the two language versions in words, expressions, or sentence structure inherent to language or culture? Example: The English sentence "Most

> rollerbladers do not favor a helmet bylaw" was translated into French with expressions "*personnes qui ne font pas de patin à roulettes*" for "rollerbladers" and "*un règlement municipal du casque protecteur*" for "helmet bylaw." These are drastically different from English forms because there are no French words that are directly parallel to the English words.

Instead of reviewing equivalence in two language versions, another commonly used review method is backward translation (also known as back-translation). Hambleton and Patsula (1999) compared translation and back-translation:

> Backward translation designs are popular but forward translation designs provide stronger evidence of test equivalence because both the source and target language versions of the test are scrutinized. That a test can be back-translated correctly (backward translation design) is not a guarantee of the validity of the target language version of the test. Unfortunately, backward translation designs are popular and yet fundamental errors are associated with this approach. (p. 160)

Therefore, the effectiveness of back-translation for evaluating equivalence is limited, and it is not recommended as a single equivalence verification process in adapted tests.

Test adaptation errors. Errors created in test adaptation process are one of the key sources of incomparability between language versions of tests. Hambleton and Patsula (1999) provided an example of poor adaptation and how it may lead to equivalence error:

> In a recent international comparative study of reading, American students were asked to consider pairs of words and identify them as similar or different in meaning. "Pessimistic–sanguine" was one of the pairs of words where American student performance was only slightly above chance. Only 54% of the American students answered the question correctly. In the country ranked first in performance, about 98% of the students answered the question correctly! In the process of attempting to better understand the reason for the huge difference in performance it was discovered that the word "sanguine" had no equivalent word in the language of this top performing country and so the foreign language equivalent of the English word "optimistic" was chosen. This substitution made the question considerably easier. In fact, pessimistic and optimistic are clearly words with opposite meaning, and would have been answered as such by a high percentage of the American students had they been presented with the pair of words "pessimistic–optimistic." (p. 158)

The authors highlighted the importance of evidence of equivalence in multiple language versions of assessments for appropriate comparative interpretations of test scores. Such adaptation errors would be discovered in a careful review of the two language versions of test items and possibly by conducting psychometric analyses investigating comparability of the two language versions.

Solano-Flores and his colleagues proposed a theory of test translation (Solano-Flores, Backhoff, & Contreras-Nino, 2009) that focused on multiple sources of translation error that may affect equivalence of tests across languages. Two notions are key to this theory: translation error dimension and translation error multidimensionality. In this theory, test translation error is defined as follows:

> The lack of equivalence between the source language version and the target language version of test items. This equivalence refers to a wide variety of properties of test items, including format and visual layout, content, and the cognitive and linguistic demands posed to test takers, among others. (p. 80)

Such errors need to be considered not just the result of poor translation but possibly also due to

translation review procedures. Review procedures include revisions and iterations in the translation process, types of reviews of translations, and piloting. Additional sources of translation error include differences integral to the languages—for example, how meaning is encoded in the languages (Nettle & Romaine, 2002).

Solano-Flores et al. (2009) categorized translation errors into translation error dimensions that can result from not addressing or complying adequately with the following:

- Criteria for assessing translation quality (e.g., American Translators Association, 2003)
- Guidelines for test adaptation (e.g., Hambleton, 1994; see also Hambleton, 2005; Mullis, Kelly, & Haley, 1996)
- Norms for the cultural appropriateness of translated items (e.g., Grisay, 2002; Maxwell, 1996)
- Knowledge on the relevance of syntactic and semantic structure of items (e.g., De Corte, Verschaffel, & De Win, 1991; Solano-Flores, Trumbull, & Kwon, 2003)
- Knowledge on the epistemology of students' interpretations of items (e.g., Solano- Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003)
- Knowledge on the sources that affect the differential functioning of translated items (e.g., Allalouf & Sireci, 1998; Allalouf, Hambleton, & Sireci, 1999; Ercikan, 2002)
- Language usage in the enacted curriculum (e.g., register)
- General writing (e.g., spelling) and item writing conventions in the target language

The notion of translation error dimensions clearly demonstrate the complexity of developing quality test adaptations and multiple language versions of tests. As an example, 10 translation error dimensions were identified in TIMSS 1995 adaptations from English to Spanish: (a) style, (b) format, (c) conventions, (d) grammar and syntax, (e) semantics, (f) register, (g) information, (h) construct, (i) curriculum, and (j) origin. Each of these translation error dimensions may include different error types. Error types in the style dimension may include incorrect use of accents, uppercase and lowercase letters, and punctuation as well as subject–verb inconsistency and spelling mistakes. Both the dimensions and the types of errors within these dimensions may be different in different test translation contexts, and translation error dimensions may be interrelated with the notion of translation error multidimensionality. This multidimensionality emphasizes the interrelatedness of linguistic features of items, including for example, the following:

> Improper insertion of a comma in a sentence may violate some writing conventions in the target language (Grammar and Syntax) but it also may change the intended meaning of an idea (Semantics). As a consequence of this multidimensionality, translation actions intended to address error on one dimension may also produce error on other dimensions. (Solano-Flores et al., 2009, p. 83)

Measurement Equivalence

To the extent that the translation errors lead to lack of equivalence between adapted versions of tests, these differences may constitute "equivalence error." Lack of equivalence between adapted versions of tests may be due to many factors other than adaptation errors. Equivalence of measurement includes (a) equivalence of constructs, (b) equivalence of tests, and (c) equivalence of testing conditions. *Construct equivalence* refers to the similarity of meaning of the construct in the cultures of the adapted versions of tests. This equivalence addresses whether theoretical and empirical evidence support similar development and equivalent definitions of the construct in the comparison groups. Construct equivalence is essential for making comparative interpretations of scores from adapted versions of tests. *Test equivalence* includes content, linguistic, and cultural equivalence between language versions. Construct equivalence is necessary, and without it, test equivalence is not meaningful to consider; however, test equivalence evidence is necessary to make claims about adapted versions of tests measuring equivalent constructs. Thus, even though construct equivalence is based on theoretical and conceptual grounds, the evidence of test equivalence is integral

to examining comparability of constructs measured by adapted versions of tests. *Equivalence of testing conditions* refers to whether (a) different language versions of tests were administered in an identical fashion; (b) the test format was equally appropriate in each language version; (c) speed of response was not more of a factor in one language version of the test than the other; and (d) other response styles such as acquiescence, tendency to guess, and social desirability (Hambleton, 2005; Hambleton & Patsula, 1999). Investigations of measurement equivalence need to take construct, test, and testing conditions into account, in addition to linguistic equivalence.

Score Comparability

The ultimate effect of adaptation-related measurement inequivalence is in the interpretation of test scores from different language versions of tests. First, comparisons of scores from different language versions of tests cannot be made without strong evidence to support score comparability. Two necessary requirements for score comparability are measurement equivalence and measurement unit or scalar equivalence. Measurement equivalence requires construct, test, and testing condition equivalence, as discussed. Measurement unit equivalence refers to whether units on the score scales based on different language versions of tests have equivalent units. When separate score scales are created for tests, these scores cannot be assumed to have equivalent scale units. That is, a score difference of 10 score points on one scale based on the source-language version of the test cannot be considered equivalent to 10 score points from a scale based on the target-language version of the test. Measurement unit equivalence requires evidence based on psychometric research to support such equivalence. Scores with scalar equivalence or full score equivalence have measurements with the same origin and measurement units. To make direct comparisons between score scales from different language versions of tests, scalar equivalence is required.

Even when score scale comparability is established between scores obtained from different language versions of tests, the same set of norms may not be used for the language and culture groups. A norm-referenced score indicates how an examinee's performance or responses to a measurement tool compare with those from a particular population. If the "population" is defined the same way for all language groups, for example, fourth graders in Canada, the use of the same set of norms may be appropriate to use for both fourth-grade English-speaking and French-speaking Canadian students, once the scalar equivalence has been established. If on the other hand, the intent is to make comparisons with different "populations" based on each language version, for example, intelligence quotient (IQ) scores for Canadian fourth graders who take the test in French versus U.S. students, it is not appropriate to use the same set of norms.

Often, adapted versions of tests are used primarily to make comparisons between language and cultural groups. Differences in scores cannot solely be attributed to be due to language or culture group membership. For example, in international assessments of educational achievement, "curricula, educational policies and standards, wealth, standard of living, cultural values, motivation to take the test, etc., may be essential for properly interpreting scores across cultural/language and/or national groups" (Hambleton & Patsula, 1999, p. 162). Language and culture group identification is only one aspect of what defines groups. Other distinctions between groups that may overlap or coincide with the language group membership may be the real "causes" of differences. For example, differences in intelligence test scores between linguistic or other ethnic groups—such as English language learners, who may be taking the test in another language, and others—may be due to such factors as differences in motivation, test-taking skills, and familiarity with contexts used in the test items rather than language group membership.

EMPIRICAL EVIDENCE FOR MEASUREMENT EQUIVALENCE

Both the *Standards* (AERA et al., 1999) and ITC Guidelines (ITC, 2001) have emphasized the need for test developers and users of adapted versions of tests to provide validity evidence for all interpretations based on such tests. The validity evidence

includes evidence of meaningfulness and appropriateness of test score interpretation for the target culture. Types of validity evidence and methods for gathering and examining such evidence for the target culture are identical to other validity investigations that do not involve adapted versions of tests. Therefore, for a more complete discussion of validity broadly, see Volume 1, Chapter 4, this handbook. An additional validity evidence requirement for adapted tests is evidence of measurement equivalence for the source and target language and culture groups. This section focuses on evidence of measurement and measurement unit equivalence. In particular, what counts as evidence and how to obtain such evidence are described. The subsections illustrate and discuss methods and evidence used to examine measurement equivalence, identifying sources of inequivalence at the item level, and establishing measurement unit and scalar equivalence.

Measurement equivalence requires evidence that (a) tests are capturing equivalent constructs, (b) tests have similar measurement characteristics and properties, and (c) tests are administered in equivalent testing conditions. In relation to *construct equivalence*, researchers need to ask, "Is it sensible to compare these two cultures on this construct? Does the construct that is being measured have similar meaning in all cultures being compared? Is the construct being operationalized the same way in all cultures being studied?" (Hambleton, 2005, p. 7). Construct equivalence evidence includes evidence of similar construct meaning for the comparison groups. van de Vijver and Tanzer (2004) described a useful method that involves using a survey to discover what behaviors and characteristics are associated with a construct in a specific culture and time. If respondents from different cultures list the same behaviors and characteristics as associated with the construct, such data can provide supporting evidence of construct equivalence. Other types of research for construct equivalence evidence include using bilingual expert reviewers to rate the equivalence of the constructs captured by pairs of items in the comparison languages.

Test equivalence refers to evidence of linguistic, content, and cultural equivalence. This requires evidence that items have similar meaning, capture similar constructs, have the same test length, have similar format, and do not include cultural references that may disadvantage one or the other language group. Test equivalence evidence is based on three key sources (a) bilingual expert reviews, (b) psychometric comparability, and (c) cognitive analysis. Examining equivalence of test items with respect to equivalence of meaning, construct, test length, format, and cultural references can be examined using expert reviews. Psychometric and cognitive aspects of equivalence are described and discussed in the following subsections.

Establishing *testing conditions equivalence* requires evidence that (a) different language versions of tests were administered in an identical fashion; (b) test format was equally appropriate in each language version; (c) test speededness was not more of a factor in one language version of the test than another; and (d) other response styles are accounted for, such as acquiescence, tendency to guess, and social desirability (Hambleton & Patsula, 1999). Each of these types of evidence requires special data collection efforts. Equivalence of administration requires data on the procedures used in test administration and documentation of the potential special circumstances of testing in the cultural context. Test format appropriateness should be documented before test administration and preferably at the test development stage.

Speededness has been documented to affect overall performance of examinees and validity of score interpretations. Data regarding speededness may include psychometric analyses of test response data to determine whether different proportions of examinees were reaching the end of the test. It can also be based on test administration documentation regarding the amount of testing time examinees took to complete the test. Perhaps the type of evidence that would be most difficult to gather would be regarding response styles. Whether examinees or respondents from different cultures have different acquiescence, tendency to guess, or social desirability may be based on broader research on cultural differences. It also may be possible to measure factors that may affect response styles, such as social desirability, by tests or measures designed to capture these constructs.

Psychometric Evidence for Test Equivalence

Commonly used psychometric analyses to investigate the equivalence of adapted versions of educational or psychological tests can be grouped into three main categories: (a) classical test theory–based analyses, (b) dimensionality analyses, and (c) identification of differential item functioning (DIF) and its sources. These approaches are described and discussed in the following subsections. Some elaborations of these approaches may be found in other chapters in this handbook, including reliability and item analysis (Volume 1, Chapters 2 and 7), factor analysis (Volume 1, Chapter 5), and bias (Volume 1, Chapter 8).

Classical test theory–based approach. Data available for investigating equivalence of different language test versions often are based on small sample sizes because of relatively small populations—for example, Francophone minorities outside of Quebec in Canada—or because of small-scale field-test studies typical in many investigations. Small sample sizes limit the types of psychometric analyses that can be used to investigate comparability. In such contexts, classical test theory–based analyses, which do not require large sample sizes, may be the only types of psychometric analyses that can be conducted. These types of comparability investigations may be sufficient, in particular when the stakes associated with the tests are not high. For example, the stakes associated with licensure examinations in multiple languages are expected to be much higher than if the testing is used as one of multiple sources of information for decision making, such as in the case of diagnostic information in educational and psychological testing. As a first step, comparisons of classical descriptive item statistics, such as item difficulty and item discrimination indices, can be made. When the difficulty parameters of test items in each language are highly correlated, this finding may be considered evidence that item difficulties are ordered similarly for the two language groups and contribute to comparability evidence. Similarly, high correlation of item discrimination parameters for different language versions of tests provides supporting evidence for comparability. A second comparison can be made between internal consistency coefficients for the two language versions of tests. Even though a similar internal consistency coefficient is not a sufficient indicator of measurement equivalence, differences in internal consistency coefficients would indicate differences in measurement accuracy between the two versions, and therefore, a lack of measurement equivalence.

Dimensionality analyses. Intercorrelations among test items provide information about how the items are related to each other and to the overall construct being measured. Such intercorrelations are typically examined using dimensionality analyses. Comparisons of dimensionality for different language and culture groups provide a measure of the equivalence of factor structures at the test level (Ercikan & Koh, 2005). These analyses may be based on exploratory (Oliveri & Ercikan, 2011) or multigroup confirmatory factor analyses (Ercikan & Koh, 2005), nonlinear exploratory procedures such as weighted multidimensional scaling (Robin et al., 2003), or confirmatory factor analyses using structural equation modeling (Wang, Wang, & Hoadley, 2007).

An assessment often is intended to measure a number of factors (e.g., construct components) within a construct, and test questions are targeted to measure each of those factors. For instance, if an assessment is designed to measure school experience, it may focus on the student's relationship with his or her teachers, relationships with other students, and interest in classroom materials. The questions that are included in the assessment will be targeted to address one of those three factors. A factor analysis presents a loading matrix that provides a statistical correlation for each item and factor. If the item is an accurate measure of the factor that it is trying to capture, it will load high on that factor and low on others.

Similarity of factor structures between language versions of tests can be examined using an exploratory or confirmatory factor analysis. An exploratory factor analysis is performed separately for each of the comparison groups. The loading matrices for each group are then compared and differences and similarities are reviewed. Clear criteria are lacking, however, to differentiate what constitutes comparable factor structures based on exploratory factor

analysis (Sireci, Patsula, & Hambleton, 2005), which creates difficulty in interpreting factor analyses results with regards to equivalence.

Alternatively, a confirmatory factor analysis allows researchers to hypothesize the number of factors present in an assessment before the analysis. The dimensionality hypotheses should be based on theoretical and empirical research on the construct and the dimensionality of tests expected to measure the construct. The loadings are forced into the hypothesized factor structure, and the fit of the data with the model is statistically tested. Inconsistencies between the loadings can be used to flag items for further analyses to determine whether they are tapping a somewhat-different construct for the language comparison groups.

Differential item functioning. DIF refers to a group difference in performance on a single item of an assessment, despite members of those groups having equivalent ability levels. Although there are many possible explanations for why one group may perform at a higher level on a single item of an assessment, linguistic or cultural differences often play a role. DIF indicates differences in psychometric properties of the compared test items and existence of DIF points to lack of item equivalence. Some of the most commonly used DIF detection methods include Mantel–Haenszel (MH), delta plot, standardization index, and item response theory (IRT) methods (see Volume 1, Chapters 7 and 8, this handbook).

The *MH method* was modeled after the chi-square test and developed by Holland and Thayer (1988) to compare two groups matched on ability based on their likelihood of answering an item correctly. A table is constructed for each item that charts proportions of examinees from each of the two groups matched on ability who answered the question correctly and incorrectly. The probability of answering the item correctly for members of each group is determined and compared.

Another DIF detection method for binary-scored test items is the *delta plot method*. This method involves plotting the *p* values for one cultural group on one axis and the *p* values for the other group on the other axis. If the difficulties of the items are consistent across the two groups, they will fall on a line with a 45-degree angle. Items with *p* values deviating from the 45-degree line can be marked for potential bias and followed up with further analyses to identify sources of DIF (Muñiz, Hambleton, & Xing, 2001).

The *standardization index* involves computing "conditional *p* value," that is, separate *p* values for each item conditional on total test score. Test score intervals are computed to match examinees, and the proportion of examinees who answered the item correctly at each interval is computed and compared for each group. For items functioning similarly, the two language groups should perform similarly. Items are flagged as DIF based on an effect size of conditional *p* value differences (Dorans & Kulick, 1986).

IRT-based DIF detection methods make use of item characteristic curves (ICCs) to examine differences in parameters that an item may have for the comparison groups. Item difficulty and discrimination parameters are compared for the comparison groups to determine whether there is a difference in the functioning of an item (Lord, 1980). Similarly, Raju's method compares ICCs for paired groups on the same item and determines whether there is a significant difference in the area between the two ICCs (Raju, 1988). Sample research that utilized these and some other DIF detection methods for investigating test item equivalence in adapted tests are presented in Table 26.1.

Adapted tests are often developed for small numbers of individuals (e.g., ethnic and language minority groups) and research involving equivalence investigations involves small sample sizes from such language and culture groups. Therefore, it is important to identify DIF detection methods that lend themselves well to conducting analyses using small sample sizes. In particular, some DIF methodologies based on Bayesian estimation approaches have been identified to work better with small sample sizes (Zwick, Thayer, & Lewis, 1999, 2000). The empirical Bayes DIF approach is based on the MH DIF detection approach. In this approach, DIF estimates are obtained by combining the observed values of the MH-index of DIF with an assumed prior distribution for the DIF parameters. This approach has been demonstrated to provide more stable

TABLE 26.1

Differential Item Functioning Detection Methods

Method	Applications in research on test adaptation
Mantel–Haenszel (Holland & Thayer, 1988)	Allalouf et al. (1999); Budgell, Raju, and Quartetti (1995); Dorans and Kulick (2006); Ercikan (1998); Muñiz et al. (2001)
Delta plot (Angoff, 1972, 1993)	Angoff & Modu (1973); Cook (1996); Facon & Nuchadee (2010); Muñiz et al. (2001); Robin, Sireci, & Hambleton (2003)
Standardization (Dorans & Holland, 1993; Dorans & Kulick, 1986)	Dorans and Kulick (2006); Sireci, Fitzgerald, & Xing (1998)
Logistic regression (Swaminathan & Rogers, 1990)	Allalouf et al., (1999); Ercikan (2003); Ercikan & Koh (2005); Oliveri & Ercikan (2011)
IRT based (Raju, 1988; Linn & Harnisch, 1981; Thissen, Steinberg, & Wainer, 1988)	Budgell et al., (1995); Ercikan et al., (2004); Ercikan & Koh (2005); Ercikan & McCreith (2002)
Simultaneous Item Bias Test (SIBTEST; Douglas, Roussos, & Stout, 1996)	Ercikan et al. (2004); Mendes-Barnett & Ercikan (2006)

estimates of DIF than the original MH statistic, with small sample sizes (as low as 25; Zwick et al., 2000). Two other methods that have been shown to work well with small sample sizes include the delta plot methods and the standardization index.

One of the limitations of the DIF methods and the methods used to examine test equivalence is the lack of an explicit demonstration of the effect of differences on the interpretation of results. The overall effect of DIF at the test level can be examined by comparing IRT-based test characteristic curves (TCCs). TCC of a test is the sum of ICCs for all the items in the test and summarizes the probability of getting the maximum test score for different latent ability levels. Statistically significant differences between TCCs are identified as differential test functioning (Drasgow & Probst, 2005; Ercikan & Gonzalez, 2008).

Examining Sources of DIF

Identifying sources of DIF is essential to make meaningful interpretations of DIF and to examine the validity of comparability of scores from adapted versions of tests. We discuss two key methodologies, expert reviews and cognitive analyses, to identify sources of DIF in the following subsections.

Expert reviews. Reviews of test items by bilingual experts have been the primary method for identifying sources of DIF between language groups (Ercikan, 2002; Gierl & Khaliq, 2001; Hambleton, 2005; Wu & Ercikan, 2006). Items are reviewed by content or bilingual experts to identify differences between language versions of items to identify linguistic, format, or cultural references that may affect examinees' performances or responses differentially. In the section Evaluation of Equivalence by Expert Reviewers, the expert review process and requirements for effective reviews of items were considered for evaluating equivalence in different languages. The purpose of the review process for identifying sources of DIF is the same as the review process during the adaptation process for evaluating equivalence, that is, to identify differences between source- and target-language versions of test items that may lead to differential performance or response patterns between the language groups. To examine sources of DIF for linguistic groups, typically, bilingual experts review items in the two comparison languages for potential linguistic differences. Experts evaluate and rate the equivalence of items with regards to equivalence of meaning, difficulty of vocabulary and sentence structure, nonlinguistic presentation differences, and key words that may guide student thinking. The review checklist presented in Appendix 26.1 can be adapted to identify differences between language versions of items to examine whether a pattern exists between these differences and DIF identification of items.

Cognitive analyses. Expert reviews of items to identify differences between different language

versions are based on surface characteristics of items and whether these differences identified by experts lead to differential performance or response patterns for different language groups needs to be investigated empirically. For example, if the review process identifies that the vocabulary used in the English test version has a different meaning when adapted to French, the review process does not tell us whether these differences do indeed lead to differential performance or response patterns.

Cognitive analyses include examinees' think-aloud verbalizations while they take tests. These verbalizations can be compared to examine differential thought processes for examinees from different language backgrounds. Recently, Ercikan et al. (2010) proposed think-aloud protocols (TAPs) as a new approach to compare examinee thought processes while they take different language versions of tests and to identify and confirm sources of DIF. TAPs involve having examinees verbally report their thought processes as they attempt to solve problems presented by an assessment.

In the case of assessments that have been adapted to be used with other cultural or linguistic groups, samples of examinees from those different groups are used for the cognitive analysis and differences among the groups are observed. Verbalizations from TAPs help identify inequalities in the meaning of test language, format, and other presentation-related aspects of test items as well as the impact of these inequalities on examinee cognitive processes and performance for examinees from different language backgrounds. The verbal data provided by the test takers are then collected by researchers and analyzed to investigate how examinees solve test problems and to identify whether differences exist between cognitive processes of examinees from different language groups. This approach can provide additional information to expert reviews or can confirm sources of DIF identified by expert reviews (Ercikan et al., 2010).

In Ercikan et al. (2010), test administrators asked students to verbalize their thoughts as they worked on responding to a question, and noted the students' understanding and perceived difficulty of the question as well as what parts of the question were helpful or unhelpful in solving the problem. Both concurrent and retrospective TAPs were used. For example, before solving a problem, students were given the following instructions:

> I would like you to start reading the questions aloud and tell me what you are thinking as you read the questions. After you have read the question, interpret the question in your own words. Think aloud and tell me what you are doing. What is the question asking you do to? What do you have to do to answer the question? How did you come up with your solution? Tell me everything you are thinking while you are doing the question. (Ercikan et al., 2010, p. 27)

If more information was believed to be necessary by the administrator, such questions could then be asked after the student completed the question:

> What is the question asking you to do? Can you tell me what steps you took to actually solve the problem? Why did you pick that answer? What helped you figure out the answer? What did you find difficult about this answer? (Ercikan et al., 2010, p. 27)

Students' verbalizations were then transcribed and coded for analysis.

On the basis of this the research, Ercikan et al. (2010) provided six recommendations for using TAPs for investigating equivalence of cognitive processes during test taking for language groups. First, TAP data should be combined with expert reviews. Second, three characteristics of student comprehension should be taken into account when determining how to execute TAPs: students' understanding of the data, what factors contributed to solving the question, and what elements were not helpful for students. Third, test administrator observations are useful in determining students' understanding and perceived difficulty of the questions, rather than merely relying on student reports. Fourth, both concurrent and retrospective responses are vital in gathering information. The former provides insights on cognitive processes that occur as the student solves a problem, and the latter can describe a student's

overall reaction to a question. A fifth suggestion is to include all test questions, and at the very least, all DIF items, to identify and confirm DIF sources using TAPs. Sixth, a final recommendation on gathering verbal responses of this type is to use a sample that is as close as possible to the original group on which the DIF analyses are based.

Establishing Measurement Unit and Scalar Equivalence

Even when measurement equivalence has been established, scores from different language versions of tests cannot be considered comparable unless a score scale linking has been conducted. Linking methods consist of a variety of psychometric and statistical procedures to establish comparability between score scales from different test versions. Depending on the degree of equivalence between test versions and the purposes of linking, different linking designs may be appropriate. Three main linking research designs have been used in linking adapted versions of tests (Sireci, 1997). First, *separate monolingual group designs* employ the use of a source-language group for the original test and a target-language group for the adapted test. A set of items determined to be equivalent in the two languages is used to anchor items in an IRT-based calibration. *Bilingual group designs* rely on the assumption that test takers are fully bilingual in each of the source and the target languages. Almost always, however, bilinguals are stronger in one language compared with the other. Differences in proficiency levels can affect the degree of comparability of performance on one language test version compared with the other. Therefore, bilinguals need to have similar levels of proficiency in the two languages. Even though this is not possible in absolute terms, language proficiencies of bilinguals need to be assessed before they can be included in linking studies. One type of bilingual design is to have one group of bilinguals take both language versions of tests. This design eliminates error due to possible group differences but allows error due to learning effect from one test to the other by gaining familiarity with test content. An alternative to the single bilingual group design is to have two randomly equivalent groups of bilinguals take either the source or adapted test. A third option is to have two groups of bilingual examinees answer a combination of source- and target-language questions. Finally, the *matched monolingual group design* attempts to match different examinee language groups on aptitude according to the ability level that is being measured. The matching of ability levels is utilized in lieu of using anchor items (Sireci, 1997).

Data from these different designs can be used to establish statistical linking. Depending on the degree of measurement comparability required between the two language versions of tests, equating, calibration, or prediction methods may be used to link score scales for the source- and target-language versions of tests (Cook & Schmitt-Cascallar, 2005). *Equating* is typically viewed as the most stringent form of linking, and it is used in situations in which the test results are intended to be completely interchangeable. Equating requires tests to be equally reliable and the construct being measured to be the same as well as meet other stringent requirements. See Kolen and Brennan (1995) for specific statistical techniques used in equating.

The adapted versions of tests hardly ever meet the requirements for equating. C*alibration* does not require equal reliability, and therefore it is typically more appropriate for linking adapted versions of tests. Data from both separate monolingual group designs as well as the bilingual group designs can be analyzed using IRT-based calibration analyses.

Finally, *prediction* is used when scores from one test are used to predict scores on another test or task. This procedure is particularly group dependent and is considered to be the weakest form of linking. Prediction does not produce equivalent scores, rather it produces scores that can have limited comparability (Cook & Schmitt-Cascallar, 2005). Prediction would be the appropriate linking method for the matched monolingual group designs.

CONCLUSION

This chapter has presented several issues regarding test adaptation and adapted tests. This section summarizes steps and recommendations for adapting

tests for use in different languages and cultures. These steps follow:

1. *Examine construct equivalence*. Examine the construct definitions of both the source test as well as that of the target version in the respective language and culture. For example, a review panel may determine the extent to which the construct is appropriate in the target culture and identify aspects of the construct that may be different for the two language and cultural groups.
2. *Select a test adaptation and development method.* Choose which type of test development is most appropriate for the purposes of your adaptation. For example, if test developers are able to build a test simultaneously with other language versions, parallel or simultaneous development may be employed. If a source-language version of an assessment already exists, and developers wish to create a target version, successive test adaptation may need to be used.
3. *Perform the adaptation of the test or measure*. Translating a measure or a test requires not only that translators be fluent in both languages but also that they are knowledgeable about both the source and target culture and that they understand the construct being studied and use of the tests (Geisinger, 1994). Other suggestions noted in this chapter include using short sentences, repeating nouns instead of pronouns, and avoiding the use of metaphors and passive voice in developing tests.
4. *Evaluate the linguistic equivalence between source and target versions*. Bilingual expert reviews should be conducted to evaluate and determine differences in language, content, format, and other appearance-related aspects of items in the two comparison languages. Feedback from reviews can be used to revise the adapted versions of tests, which can be reevaluated by bilingual experts.
5. *Document changes made in the adaptation process*. Document changes and the rationale for these changes between the two language versions of tests for future test users.
6. *Conduct a field-test study to examine measurement equivalence.* These data are used to examine reliability and validity of both language versions of tests as well as measurement equivalence using classical test theory–based analyses, factor analyses, DIF analyses, and comparisons of TCC curves. A second round of expert reviews and cognitive analyses can be used to provide further support for comparability of the language versions of tests.
7. *Conduct linking studies.* Once measurement equivalence has been established, conduct a linking study to create measurement unit equivalence.

Responsibilities of test developers do not end once tests have been developed, and measurement and measurement unit equivalence have been established. The validity and measurement equivalence evidence needs to be updated periodically given potential changes in society, education systems, and sociocultural contexts of assessments over the years.

APPENDIX 26.1: CHECKLIST OF POSSIBLE TRANSLATION DIFFERENCES OR ERRORS

No.	Topics	Review questions and examples	Please check if applicable
1	Differences in cultural relevance	**Is the item content more relevant to one group than the other?** Example: The passage of a problem-solving item contains food or cultural events that are more relevant or familiar to one group of examinees than the other.	
2	Differences in the actual meaning of an item	**Was the meaning of an item changed in the translation process?** Example: A word that has a single meaning was translated into a word with more than one meaning.Example: In English, an item asked to compute the kilograms of apples in *each of two boxes*. When translated in French, the item asked to compute the kilograms of apples in *each one of the two boxes*, which might have caused some confusion (e.g., about whether or not to add the amounts for each box).	
3	Differences in the item format	**Are there differences in punctuation, capitalization, item structure, typeface, and other formatting usages that are likely to influence the performance for one group of examinees?** Example: In the English version, a problem-solving item provided a tree diagram at the left edge of the page, whereas in Korean, it was located in the center of the page. The location of the diagram at the left edge of the page may have clued the English-speaking examinees to the answer, which was to expand the diagram to the right.	
4	Omissions or additions that affect meaning	**Are there omissions or additions of words, phrases, or expressions that may influence the meaning of an item or the performance of one group of examinees?** Example: The English form of an item contained the expression "this number written in standard form," whereas the Korean version only mentioned "이숫자는" (i.e., "this number is"), omitting the idea of *standard form*.	
5	Differences in verb tense	**Is the verb tense different in one language version from the other?** Example: In the English version, the verb "read" was presented in the present tense, whereas in the Korean version, it was presented in past tense.	
6	Differences in the difficulty or commonness of vocabulary	**Is a certain vocabulary word more difficult or less common in one group than the other?** Example: In the English version, the word "burn" was used, whereas in the French version, it was presented as "combustion," which is more difficult and less frequently used than the word "burn."	
7	Exclusion or inappropriate translation of key words	**Is any key word that provides clues to guide examinees' thinking processes excluded or inappropriately translated for one group of examinees?** Example: In the English version, the stem of a question stated, "whenever scientists carefully measure *any* quantity many times, they expect that . . . " The correct answer was "most of measurements will be close but not exactly the same." However, French version asked, "when scientists measure the *same* quantity many times, they expect that . . . " The word *same* in the French version could have led the examinees to think that this quantity was known and the answer should be that the scientists should get the same amount each time.	
8	Differences in additional information provided to guide examinees' thinking process	**Are there differences in additional information given to guide examinees' thinking processes?** Example: In the English version a question asked, "At what point will the reflection of the candle *appear to be*?" In the French version, it asked, "At what point will the image of the candle *seem to appear to be*?" The French version is more informative suggesting that the reflection in the mirror may seem different than the actual object. This additional information could make the item easier for French-speaking examinees.	
9	Differences in length or complexity of sentences	**Are there differences in the length or complexity of sentences?** Example: Although the translation was accurate, the sentence length became shorter or longer or the sentence content became easier or harder to understand.	
10	Differences in words, expressions, or sentence structure inherent to language or culture	**Are there differences in words, expressions, or sentence structure inherent to language or culture?** Example: The English sentence "Most rollerbladers do not favor a helmet bylaw" was translated into French with expressions "*personnes qui ne font pas de patin à roulettes*" for "rollerbladers" and "*un règlement municipal du casque protecteur*" for "helmet bylaw." These are drastically different from the English forms because there are no French words that are directly parallel to the English words.	

Note. The purpose of this checklist is to guide the review of test items by providing a list of possible translation differences or errors. Please consider the list of topics for each item and use the last column to keep records. From *Measurement Equivalence of Korean and English Versions of PISA Mathematics Item*, by K. Y. Jang, 2010. Used with the permission of the author.

References

Alderman, D. L. (1981). *Language proficiency as a moderator variable in testing academic aptitude.* Princeton, NJ: Educational Testing Service.

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education*, *16*, 55–73. doi:10.1207/S15324818AME1601_3

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*, 185–198. doi:10.1111/j.1745-3984.1999.tb00553.x

Allalouf, A., & Sireci, S. G. (1998, April). *Detecting sources of DIF in translated verbal items.* Annual Meeting of the American Educational Researchers Association, San Diego, CA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Translators Association. (2003). *Framework for standard error marking and explanation of error categories.* Retrieved from http://www.atanet.org/certification/aboutexams_error.php

Angoff, W. H. (1972). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96–116). Baltimore, MD: Johns Hopkins University Press.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 96–116). Baltimore, MD: Johns Hopkins University Press.

Angoff, W. H., & Modu, C. C. (1973). *Equating the scores of the Prueba de Aptitud Academia and the Scholastic Aptitude Test* (Research Rep. No. 3). New York, NY: College Entrance Examination Board.

August, D., & Hakuta, K. (1997). *Improving schooling for language-minority children: A research agenda.* Washington, DC: National Academies Press.

Benjamin, W. (1972). Die Aufgabe des Übersetzers [The task of the translator]. In *Gesammelte Schriften* (Vol. 4). Frankfurt, Germany: Suhrkamp-Verlag.

Bowles, M., & Stansfield, C. W. (2008). *A practical guide to standards-based assessment in the native language.* Retrieved from http://www.ncela.gwu.edu/files/uploads/11/bowles_stansfield.pdf

Brislin, R. W., Lonner, W. J., & Thorndike, R. M. (1973). *Cross-cultural research methods.* New York, NY: Wiley.

Budgell, G., Raju, N., & Quartetti, D. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19*, 309–321. doi:10.1177/014662169501900401

Cook, L. I. (1996, August). *Establishing score comparability for tests given in different languages.* Paper presented at the 104th Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

Cook, L. I., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 139–169). Mahwah, NJ: Erlbaum.

De Corte, E., Verschaffel, L., & De Win, L. (1991). Some factors influencing the solution of addition and subtraction word problems. In K. Durkin & B. Shire (Eds.), *Language in mathematical education: Research and practice* (pp. 17–30). Buckingham, England: Open University Press.

Derrida, J. (1986). *Parages.* Paris, France: Galilée.

Derrida, J. (1998). *Monolingualism of the other; or the prosthesis of origin.* Stanford, CA: Stanford University Press.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, *23*, 355–368. doi:10.1111/j.1745-3984.1986.tb00255.x

Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the Mini-Mental State Examination: An application of the Mantel-Haenszel and standardization procedures. *Medical Care*, *44*(11, Suppl. 3), S107–S114. doi:10.1097/01.mlr.0000245182.36914.4a

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*, 465–484. doi:10.1111/j.1745-3984.1996.tb00502.x

Duran, R. P. (1988). Validity and language skills assessment: Non-English background students. In H. Wainer & H. I. Braun (Eds.), *Test validity* (p.105–127). Hillsdale, NJ: Erlbaum.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, *29*, 543–553. doi:10.1016/S0883-0355(98)00047-0

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, *2*, 199–215. doi:10.1207/S15327574IJT023&4_2

Ercikan, K. (2003). Are the English and French versions of the third international mathematics and science study administered in Canada comparable? Effects of adaptations. *International Journal of Educational Policy, Research, and Practice, 4*, 55–75.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, *29*(2), 24–35. doi:10.1111/j.1745-3992.2010.00173.x

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, *17*, 301–321. doi:10.1207/s15324818ame1703_4

Ercikan, K., & Gonzalez, E. (2008, March). *Score scale comparability in international assessments*. Paper presented at the National Council on Measurement in Education, New York, NY.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*, 23–35. doi:10.1207/s15327574ijt0501_3

Ercikan, K., & McCreith, T. (2002). Effects of adaptations on comparability of test items and test scores. In D. Robitaille & A. Beaton (Eds.), *Secondary analysis of the TIMSS results: A synthesis of current research* (pp. 391–405). Dordrecht, the Netherlands: Kluwer. doi:10.1007/0-306-47642-8_24

Ercikan, K., & Roth, W. M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher*, *35*(5), 14–23. doi:10.3102/0013189X035005014

Ercikan, K., Simon, M., & Oliveri, M. E. (2012). Score comparability of multiple language versions of assessments within jurisdictions. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large scale assessment in education: Theory, issues, and practice* (pp. 110–124). New York, NY: Routledge.

Facon, B., & Nuchadee, M-L. (2010). An item analysis of Raven's colored progressive matrices among participants with Down syndrome. *Research in Developmental Disabilities*, *31*, 243–249. doi:10.1016/j.ridd.2009.09.011

Fitzgerald, C. T. (2005). Test adaptation in certification program. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 195–212). Mahwah, NJ: Erlbaum.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*, 304–312. doi:10.1037/1040-3590.6.4.304

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, *38*, 164–187. doi:10.1111/j.1745-3984.2001.tb01121.x

Grisay, A. (2002). *Translation and cultural appropriateness of the test and survey material.* Paris, France: Organisation for Economic Co-operation and Development.

Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation Equivalence across PISA Countries. *Journal of Applied Measurement*, *8*, 249–266.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*, 229–244.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, *17*, 164–172. doi:10.1027//1015-5759.17.3.164

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah, NJ: Erlbaum.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research*, *45*, 153–171. doi:10.1023/A:1006941729637

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, *1*, 1–30.

Heidegger, M. (1996). *Being and time* (J. Stambaugh, Trans.). Albany: State University of New York Press.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

International Test Commission. (2001). *International Test Commission guidelines for test adaptation.* London, England: Author.

International Test Commission. (2010). *International Test Commission guidelines for translating and adapting tests.* Retrieved from http://www.intestcom.org/upload/sitefiles/40.pdf

Jackson, D. (1991). Problems in preparing personality tests and interest inventories for use in multiple cultures. *Bulletin of the International Test Commission*, *18*, 88–93.

Jang, K. Y. (2010). *Measurement equivalence of Korean and English versions of PISA mathematics item.*

Unpublished report, University of British Columbia, Vancouver, British Columbia, Canada.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York, NY: Springer.

Lin, J., & Rogers, W. T. (2005). *Validity of the simultaneous approach to the development of equivalent achievement tests in English and French (Stage II)*. Poster presentation at the annual conference of the Nation Council for Measurement in Education.

Linn, R. L., & Harnisch, D. (1981). Interactions between group membership in achievement test items. *Journal of Educational Measurement, 18*, 109–118. doi:10.1111/j.1745-3984.1981.tb00846.x

López, S., & Romero, A. (1988). Assessing the intellectual functioning of Spanish-speaking adults: Comparison of the EIWA and the WAIS. *Professional Psychology: Research and Practice*, *19*, 263–270. doi:10.1037/0735-7028.19.3.263

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Maldonado, C. Y., & Geisinger, K. F. (2005). Conversion of the Wechsler adult intelligence scale into Spanish: An early test adaptation effort of considerable consequence. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 213–234). Mahwah, NJ: Erlbaum.

Maxwell, B. (1996). *Translation and cultural adaptation of the survey instruments*. Chestnut Hill, MA: Boston College.

Melendez, F. (1994). The Spanish version of the WAIS: Some ethical considerations. *Clinical Neuropsychologist*, *8*, 388–393. doi:10.1080/13854049408402041

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19*, 289–304. doi:10.1207/s15324818ame1904_4

Merenda, P. F. (2005). Cross-cultural adaptation of educational and psychological testing. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 321–341). Mahwah, NJ: Erlbaum.

Mullis, I. V. S., Kelly, D. L., & Haley, K. (1996). Translation verification procedures. In M. O. Martin & I. V. S. Mullis (Eds.), *Third international mathematics and science study: Quality assurance in data collection* (pp. 1–14). Chestnut Hill, MA: Boston College.

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, *1*, 115–135. doi:10.1207/S15327574IJT0102_2

Nettle, D., & Romaine, S. (2002). *Vanishing voices: The extinction of the world's languages*. Oxford, England: Oxford University Press.

Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability lead to similar conclusions? *Applied Measurement in Education*.

Organisation for Economic Co-operation and Development. (2005). *PISA 2003 technical manual*. Retrieved from http://www.pisa.oecd.org/dataoecd/49/60/35188570.pdf

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502. doi:10.1007/BF02294403

Ricœur, P. (2004). *Sur la traduction* [On translation]. Paris, France: Bayard.

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, *3*, 1–20. doi:10.1207/S15327574IJT0301_1

Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. M. (2003). Differential validity and utility of successive and simultaneous approaches to the development of equivalent achievement tests in French and English. *Alberta Journal of Educational Research*, *49*, 290–304.

Roth, W. M. (2009). Phenomenological and dialectical perspectives on the relation between the general and the particular. In K. Ercikan & W. M. Roth (Eds.), *Generalization in educational research* (pp. 235–260). New York, NY: Routledge.

Rueda, R., & Mercer, J. (1985). *A predictive analysis of decision-making practices with limited English proficient handicapped students*. Los Alamitos, CA: Southwest Regional Laboratory for Educational Research and Development.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice*, *16*(1), 12–19. doi:10.1111/j.1745-3992.1997.tb00581.x

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998, April). *Adapting credentialing examinations for international uses*. Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–115). Mahwah, NJ: Erlbaum.

Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. (2006). Evaluating guidelines for test adaptations: An empirical analysis of translation quality. *Journal of Cross-Cultural Psychology, 37*, 557–567. doi:10.1177/0022022106290478

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*, 553–573. doi:10.1002/tea.1018

Solano-Flores, G., Backhoff, E., & Contreras-Nino, L. A. (2009). Theory of test translation error. *International Journal of Testing*, *9*, 78–91. doi:10.1080/15305050902880835

Solano-Flores, G., Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1992). *Science performance assessments: Use with language minority students.* Unpublished manuscript.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, *32*(2), 3–13. doi:10.3102/0013189X032002003

Solano-Flores, G., Trumbull, E., & Kwon, M. (2003, April). *The metrics of linguistic complexity and the metrics of student performance in the testing of English language learners.* Paper presented at the Annual Meeting of the American Evaluation Research Association, Chicago, IL.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, *2*, 107–129. doi:10.1207/S15327574IJT0202_2

Spielberger, C. D., Moscoso, M. S., & Brunner, T. M. (2005). Cross-cultural assessment of emotional states and personality traits. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 343–367). Mahwah, NJ: Erlbaum.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement,* *27*, 361–370. doi:10.1111/j.1745-3984.1990.tb00754.x

Tanzer, N. K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 235–263). Mahwah, NJ: Erlbaum.

Tanzer, N. K., & Sim, C. O. E. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC guidelines for test adaptations. *European Journal of Psychological Assessment, 15*, 258–269. doi:10.1027//1015-5759.15.3.258

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–172). Hillsdale, NJ: Erlbaum.

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54*, 119–135. doi:10.1016/j.erap.2003.12.004

Wang, S., Wang, N., & Hoadley, D. (2007). Construct equivalence of a national certification examination that uses dual languages and audio assistant. *International Journal of Testing*, *7*, 255–268. doi:10.1080/15305050701437969

Wittgenstein, L. (1958). *Philosophical investigations* (3rd ed.). New York, NY: Macmillan.

Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing*, *6*, 287–300. doi:10.1207/s15327574ijt0603_5

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*, 1–28. doi:10.1111/j.1745-3984.1999.tb00543.x

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics*, *25*, 225–247.

CHAPTER 27

PSYCHOMETRIC PERSPECTIVES ON TEST FAIRNESS: SHRINKAGE ESTIMATION

Gregory Camilli, Derek C. Briggs, Finbarr C. Sloane, and Ting-Wei Chiu

Test fairness has many dimensions, some of which concern the consequences of test use, and some the validity of inferences based on test results. Still others have a more mathematical nature and can be demonstrated through proof, derivation, or simulation studies. In this latter case, conceptual and empirical results may have important implications for procedural choices in how test scores are constructed. Furthermore, they can be used to link technical considerations to decision-making processes and outcomes based on test scores. In this chapter, three psychometric issues are considered and related through the idea of shrinkage—a statistical procedure for purportedly obtaining more "stable" estimates of test scores. As will be argued, the term *stability* carries a heavy burden of assumptions.

Although this chapter focuses on psychometric and related statistical issues in this paper, the topic of test fairness is considerably broader. The *Standards for Educational and Psychological Testing* (hereafter, the *Standards*; American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999) reflects a broad representation of test fairness:

> A full consideration of fairness would explore the many functions of testing in relation to its many goals, including the broad goal of achieving equality of opportunity in our society. It would consider the technical properties of tests, the ways test results are reported, and the factors that are validly or erroneously thought to account for patterns of test performance for groups and individuals. (p. 73)

Consistent with this comprehensive perspective, Camilli (2006) described fairness in terms of inferences drawn from test scores or test items in evaluating or selecting tests or test takers. Investigating test fairness in testing is thus parallel to investigating test validity, and thereby involves the use of an array of evidence that includes empirical data as well as legal, ethical, political, philosophical, and economic reasoning. Fairness and validity, however, are distinctly different properties of test score interpretation.

As noted in the *Standards*, fairness "is subject to different definitions and interpretations in different social and political circumstances" (AERA et al., 1999, p. 80). The analysis of fairness, however, usually begins with determining whether an item or test score functions the same for different groups of individuals (such as men and women). The phrase *functions the same* in psychometrics is often shortened to *invariant* for convenient reference, and this lexicon is adapted herein. For test items, invariant means that for comparable individuals from the two groups, average item scores are about the same. That is, two groups of comparable individuals should each have about the same proportion of correct answers on a multiple-choice question. Statistical methods of *differential item functioning* (DIF) are used to find items that are not invariant in this sense. In the case of test scores, invariant means that the degree of success predicted on a criterion is

DOI: 10.1037/14049-027
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

about the same for individuals from different groups who have the same predictor test score, and statistical methods of *differential prediction* are used to determine whether test scores are invariant in this sense. For example, two individuals with the same college admission score, one male and one female, should have about the same degree of success in their first year of college as measured by grade point average. If DIF is observed, then further investigation is required to determine whether the item score is a fair measure. When differential prediction is established, and the criterion is a valid measure of success, then additional investigation of the selection test is also required.

In the case of DIF, the lack of invariance may lead to unfairness in the sense that an item might be more difficult for a group of examinees for reasons that are irrelevant to the ability or capacity (i.e., the construct) intended to be measured. With differential prediction, the lack of invariance leads to different selection rates for comparable candidates. In both cases, further investigation is warranted to determine the probable cause of differential functioning or prediction as well as the probable direction of the statistical bias. At that point, a strategy would be adopted to ameliorate the bias.

Consider an analogy for understanding the basics of determining invariance in terms of the difference in average running speeds over a fixed distance for two groups. Suppose that two groups (R and F) in reality have highly similar average running times but run in different geographic regions. Also suppose a single particular stopwatch available for measuring running times in both regions. Then statistical bias arises when the stopwatch for group R is accurate but is inaccurate for group F. This might be because the group R and F measurements are taken in different climates, and the stopwatch is affected by a difference in temperature or humidity (or is damaged in transit from region R to region F). But it could also be the case that it is windier in one region, and this factor gives the edge to one group. In the former case, the stopwatch itself measures differently in the two groups (an intrinsic problem), and in the latter, the stopwatch works fine, but the running speeds are affected by an extraneous factor (an extrinsic problem). If it is further established that the stopwatch runs too fast (because of humidity) or there was a headwind present for group F, then the timing information provided by the stopwatch is unfair in the sense that group F is disadvantaged—although group R could also be disadvantaged by the same circumstances. In psychometrics, the stopwatch is a metaphor for an item or test score, and geographic region is a metaphor for group background characteristics. In any event, an item score should be the same on average for comparable groups of individuals, and so should predictions of success that are made on the basis of a test score. Two fundamental ideas are given in extrapolating the metaphor to test data.

An average difference between two groups (also known as disparate impact) does not imply unfairness in a test item. If a group of native speakers had a generally higher level of mathematics performance than a group of non-English speakers, this difference would be reflected across most test items. Thus, one essential idea is that comparisons can be made only among comparable individuals, and there is no presumption that groups as a whole are equal in the quality being measured. At this point, the astute reader might wonder why the capacity to identify comparability is presumed with DIF analysis. After all, presuming that individuals can be identified as comparable is tantamount to presuming that true running times have been measured by a perfect stopwatch. In response, a psychometrician or measurement specialist would offer the following rationale.

Briefly, test items are investigated to determine whether some items show relatively larger or smaller group differences than other items on a test for individuals who are similar on an *available* overall score from the test in question. This total test score is taken as a pragmatic substitute or a first approximation for a true measure of comparability, and the resulting DIF only raises a flag for further investigation. For example, for examinees with the same total test score, men might have a higher proportion of correct answers than women, or non-English speakers may have a higher proportion correct than native speakers. When the difference in question between groups can be attributed to factors irrelevant to the test construct, then the item could be considered

biased and either removed from the test or modified. Thus, the requirement that the difference be interpretable is a safeguard against using an imperfect measure of comparability. DIF is often found, but it is rarely interpretable. This limitation is well understood, but DIF techniques are generally considered to be useful as one step within a process of establishing a fairness argument.

With differential prediction, a valid measure of the criterion is presumed for the standard analysis. The goal is to determine whether the identical true criterion score for individuals from different groups results in about the same observed score on the test in question. Usually in differential prediction, this test (or predictor variable) is considered to be an independent variable (on the *x*-axis), and the criterion to be a dependent (or criterion) variable (on the *y*-axis). The standard presentation of axes has been reversed herein so that the concepts of DIF and differential prediction can be examined in parallel; in both cases, individuals who are comparable in terms of a true measurement (*x*-axis) should on average have about the same level of performance on the item or test outcome (*y*-axis) being assessed for violation of invariance. In both cases, the analysis proceeds pragmatically by assuming a proxy for the true measurement. In the stopwatch metaphor, the criterion is the actual running time (as measured by a perfect stopwatch), and this serves as the basis for comparability—while the available stopwatch is parallel to the item or test score. Consequently, steps roughly parallel to those of DIF analysis are taken if differential prediction is observed for comparable individuals from different groups. First, the group difference is estimated; second, the direction of bias is determined; and, third, the test construct is reexamined to understand potential causes of the observed difference. Only then would the effects of the differential performance on a selection rule be determined and corrective steps taken. Thus, for both DIF and differential prediction, the stopwatch metaphor provides a procedural (or syntactic) understanding of the bias metaphor. Ultimately, however, judgment and experience must used to interpret the results and arrive at an assessment of fairness, and this analysis requires a substantive understanding of the constructs and the measurement context.

A second essential idea involves the nature of the measurement problem when differences are found between "comparable" groups of individuals. In terms of the metaphor, the issue could be intrinsic to an item or score (parallel to a stopwatch being affected relatively more by temperature or humidity for one group), or it could be due to an extrinsic factor (parallel to running times being affected relatively more by headwinds for one group). In DIF, an intrinsic problem could arise from how a question is worded (e.g., a double negative), whereas an extrinsic problem might arise from secondary abilities required to answer a question (e.g., verbal reasoning in English for a mathematics problem). In the case of differential prediction, an intrinsic problem might arise as the incorrect choice of construct (e.g., abstract verbal reasoning for selecting music majors), whereas an extrinsic problem could result from a speeded test (e.g., one group is accustomed to speeded tests and the other is not). Practically speaking, the line between intrinsic and extrinsic factors is not sharp, and a useful distinction typically involves degree more than type.

A different way to consider fairness in the context of test items and scores can be framed in terms of equitable and humane treatment. From this perspective, fairness concerns equitably and the avoidance of insensitive item content. All test takers should be treated equitably throughout the testing process, and item content should not include language or other material likely to be offensive to some individuals. The impact of inequitable treatment or insensitive material may not distort the measurement process, but these concerns derive from the more fundamental notion of preserving respect and dignity in the assessment process. Test scores are developed to measure specific content knowledge or performance capacity as conveyed in a test blueprint or set of specifications for an intended construct. Psychometric or statistical choices, as will be argued, may lead to incomplete alignment of a test score with measurement intent—to the advantage of some individuals and to the detriment of others. In a real sense, this is also an equity issue, but the standard for comparison is the *intended test content* rather than the ubiquitous "groups."

This issue can be framed in terms of the decision in *Debra P. v. Turlington* (1981). Initially heard in

Federal District Court, which enjoined the State of Florida from using a new high school graduate test, the case was subsequently appealed to the U.S. Court of Appeals. On the basis of these cases, three basic legal standards were established for test fairness in high school exit examinations. Of interest in this paper is the requirement that a test must measure knowledge and skills that are taught before the test. This is has become known as "curricular validity," which typically defines the match between test content, on the one hand, and the content represented in both curricular materials and actual instruction, on the other. Thus, for an exit examination to be fair, teachers must be teaching what the test is testing. Given this framing, the psychometric issue is how well the test-scoring procedure represents the intended content, and this matter arises because tests can be scored in different ways, and not all scoring methods represent the intended content equivalently. For example, in some assessment programs, a test score is obtained as the simple sum of item scores, but in other programs, the score is obtained in a more complex fashion. We argue that the test-scoring method is equitable when the test content is fairly represented in a derived score for an individual.

This chapter examines one such *psychometric* equity issue and two issues that are concerned with the group comparison approach to test fairness. All three issues are woven together by the notion of shrinkage estimation, and the unintended consequences that may arise from this approach to estimation. In simple nontechnical terms, all measurements are thought to contain some degree of error, and the gist of shrinkage estimation is that when this error is considerable, estimates should be adjusted back toward a group or population average. This dampens the impact of measurement error on any given value estimated for some target unit of analysis. The chapter demonstrates that although the statistical reasoning behind this approach may be sound in theory, the associated rationale can be unconvincing or incompatible with the context in which a measurement or statistic is obtained.

The first of the three issues examined in this paper is the estimation of ability by means of the three-parameter logistic item response theory (3PL IRT) model (Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968; see also Volume 1, Chapter 6, this handbook). Specifically, it can be shown that if a 3PL model is used to score a test, then item responses do not contribute independently to an estimate of the underlying latent variable; or, stated differently, that two examinees may receive different credit for correct responses to the same item—depending on their patterns of responses to other items (Chiu & Camilli, in press). Here the phenomenon of *credit shrinkage* can be derived directly from the mathematics of the 3PL model, although the phenomenon of differential item credit could conceivably arise for other reasons, such as estimation error or model misfit.

For instance, proficiency is defined differently in the one-, two-, and three-parameter logistic models of IRT. Although pragmatists point out the high intercorrelations of the resulting proficiency estimates and the simple total test score, scores based on the 3PL model have exceptional consequences for test takers of low ability. For example, 3PL score estimates do not uniformly map onto a test's design specifications because of a kind of shrinkage—and this has a direct impact on score interpretation and equity—even if this model accurately fits the data. More generally, studies of the match between formal models (e.g., a regression or IRT model) and desired interpretations fall under the rubric of *psychometric fairness*, which is distinct from more traditional methods for evaluating test fairness.

The second issue is the role of shrinkage estimators in DIF. Zwick, Thayer, and Lewis (2005) and Sinharay, Dorans, Grant, and Blew (2009) proposed related methods for estimating the amount of DIF in test items that involve shrinking an initial estimate of DIF toward zero, thus obtaining a lower (closer to zero), more conservative estimate of DIF. Consistent with statistical theory, Zwick et al. (2005) found, in a simulation study, that shrinkage estimates of Mantel–Haenszel (MH) DIF statistics tended to be closer to their true values than the standard MH statistics and that this effect tended to be most pronounced in small samples. Sinharay et al. (2009) found that the shrinkage method results in dramatically more conservative estimates of DIF and concluded that shrinkage estimates are preferable to

classical estimates despite the irony that once shrinkage methods were applied, virtually no DIF was detected—even when it did exist—in any of the simulation conditions examined.

The third issue is selection based on fallible test scores and issues of measurement invariance. Briefly, tests are often used to select individuals for a limited number of positions, such as enrollment at a highly regarded university. The validity of such a test is partially established by showing a positive correlation with a criterion, such as success in college. Yet no selection test measures a criterion (such as college preparedness) perfectly because of *measurement error*, a phrase that conveys the notion that an individual score might be higher or lower than the true value for unknown or random reasons. All test scores to some degree are affected by random error, which is not considered a direct threat to test fairness. (The standard definition of *true score* is the average score of an individual in the long run, where the error is that any single observed score is simply a deviation from this average. Because this definition poses a hypothetical long run, true and error scores cannot be observed at the individual unit of analysis.) In contrast, one property of a fair test is that test items should measure an individual's level of qualification the same way, regardless of any group to which that individual may belong (e.g., male or female). Psychometricians refer to this property as *measurement invariance* (see Engelhard, 2008, for a general discussion). A second property concerns the rates at which individuals from different groups are predicted to succeed on the criterion. For example, two randomly chosen individuals from different groups who are selected based on the same cutoff score should have the same probability of success in college. This is referred to as *selection invariance*.

Borsboom, Romeijn, and Wicherts (2008) considered the situation in which the groups in question have different distributions on the selection test and criterion. Following results summarized by Millsap (2007), they showed that the presence of measurement error in test scores implies generally that selection invariance cannot be satisfied even when measurement invariance holds. In the current paper, this incompatibility is illustrated in the framework of IRT, but after Petersen and Novick (1976), the chapter shows that selection invariance does not generally hold even when groups share the identical criterion–predictor regression. Several solutions to this problem were proposed by Borsboom et al. (2008), but an alternative solution for minimizing this problem exists in the use of shrinkage estimators for proficiency, and the legal and societal reasons that constitute barriers to the implementation of this solution are explored.

These three topics are useful to illustrate how psychometrically and socially responsible requirements of test use may intersect. The flow of these topics is organized according to the following rationale. The first topic (3PL estimation of proficiency) comes before topics concerning test score use. The second topic (DIF) concerns the measurement properties of individual test items and also is a concern that exists logically before test score use. The third topic (selection) provides a logical segue from discussion of measurement to the first direct use of test scores. The chapter closes with a summary and a recommendation that psychometricians and measurement specialists should provide greater clarity in applying shrinkage estimators to societal issues. Before more detailed investigation of these issues, the basic idea of shrinkage estimators is first introduced.

SHRINKAGE ESTIMATORS: A BRIEF INTRODUCTION

As opposed to classical estimators such as the mean, shrinkage estimators exploit additional sample information to improve precision. As will be shown, when multiple groups are available in a data set, an estimate of a particular group mean can be improved by using (or "borrowing") information from the other groups. Likewise, a regression coefficient can be improved by borrowing information from the regression coefficients of other groups (Braun, 2006; Novick, 1970; Sloane, 2008). The resulting improvement of the estimator can be measured in terms of a decrease in mean squared error (*MSE*), where *MSE* is mathematically described for an estimator $\hat{\varphi}$ of the parameter φ by the equivalence $MSE(\hat{\varphi}) = Var(\hat{\varphi}) + bias^2$. The statistical justification for using a shrinkage estimator is minimizing the criterion of *MSE*, by sacrificing a

(hopefully) small increase in bias for a large decrease in $Var(\hat{\varphi})$.

The general univariate form of a shrinkage estimator, say $\tilde{\varphi}$, is given in two familiar forms by the equations

$$\begin{aligned}\tilde{\varphi}_j &= \rho\varphi_j + (1-\rho)\varphi \\ &= \varphi + \rho(\varphi_j - \varphi)\,,\end{aligned} \tag{27.1}$$

where φ_j is the parameter to be estimated for unit j before shrinkage, the unsubscripted φ is the global or average population parameter, and ρ is a reliability coefficient. Because ρ ranges between 0 and 1, $\tilde{\varphi}_j$ is closer than φ_j to the global value φ. The reliability ρ also has the general form

$$\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}, \tag{27.2}$$

where σ_τ^2 is a true (or between) variance component, and σ_ε^2 is the error (or within) variance component (for elaboration of the notion of reliability, see Volume 1, Chapter 2, this handbook). Different types of shrinkage estimators can then be represented by substitutions for φ_j, φ, σ_τ^2, and σ_ε^2. Estimates of these four parameters are required before the calculation of estimates for $\tilde{\varphi}_j$. Three estimators taking this general form are shown in Table 27.1.

A common example of a shrinkage estimator in psychometrics is known as the Kelley (1927) regression estimator for true scores. The components of this estimator are shown in the second column of Table 27.1. According to classical measurement theory, an observed total score x on an n-item test can be broken down into a true component t and an error component e, where $x = t + e$. Observed score variation also can be decomposed into two pieces: one due to true scores (σ_t^2), and another due to random measurement error (σ_e^2 / n). Although a number of properties of this coefficient are important, what is critical here is that as the number of items n goes up, the reliability of the total score given in Equation 27.2 goes up (although the increase is not linear). Shrinkage (which is a kind of regression to the mean) in this case refers to the movement toward mean performance of an observed score x_j for individual j: When the reliability is less than 1, then as can be seen in Equation 27.1, the regressed true score estimate shrinks toward the sample mean, which is generally the best available estimate of the population mean. In fact, when the reliability is zero, the shrinkage estimate of any individual's true score *is* the sample mean, which is clearly a biased estimator for any nonzero true score. One could argue that the sample mean is the only conceivable score that makes sense in such contexts, which

TABLE 27.1

Representation of Different Shrinkage Estimators

General form	$\tilde{\varphi}_j = \rho\varphi_j + (1-\rho)\varphi$		$\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}$
	Kelley	3PL Item	DIF
Raw Effect			
φ_j	t_j	P_{ij}	D_j
Component[a]		Estimate[b]	
φ_j	x_j	u_{ij}	$\hat{D}_j$
φ	$\bar{x}$	zero	zero
σ_τ^2	$\hat{\sigma}_t^2$	$\hat{\sigma}_{ij}^2$	$\hat{\sigma}_D^2$
σ_ε^2	$\hat{\sigma}_e^2/n$	$\left(1 - \hat{P}_{ij}\right)^2 / \hat{c}^{-1}$	σ_j^2

Note. 3PL = three-parameter logistics; DIF = differential item functioning.

[a] $\sigma_{ij}^2 = P_{ij}Q_{ij}$ and $Q_{ij} = 1 - P_{ij}$.

[b] $\hat{P}_{ij}$ is obtained by substituting estimates of a_i,b_i and θ_{ij} into P_{ij}.

illustrates that statistical bias is not necessarily an undesirable property of an estimator.

The Kelley (1927) regression estimator can also be understood in a Bayesian framework in which the information regarding a true score is a combination of collateral information (i.e., the "prior distribution") and the observed data for a particular person. A simple, but relevant, analogy can be made for estimating the adult facial features of a 10-year-old child. Suppose the prior (sometimes referred to as *collateral*) information is the facial features of the child's parents (and possibly other relatives), and the observed data are the measured facial features of the child. Both sources of information can be combined to obtain a prediction (i.e., the "posterior distribution") of the child's facial likeness. In fact, this approach roughly describes actual aging algorithms developed for the identification of missing children (ANSER Analytic Services, Inc., 2000). Accordingly, the child's likeness as measured in current photographic information is "regressed" to the features of the parental (and adult age) distribution.

Novick (1970) described statistical shrinkage as obtaining score estimates that are regressed (or shrunken) to different group means, depending on the "group" affiliation of the individual in question. Thus, collateral information is used to obtain a better estimate of an individual score. It follows that if a large average difference exists between two groups, there may be a substantial difference between the regressed true scores for the same observed score—because the scores are regressed to different group means. In computing shrinkage estimates, trading some small bias for additional precision is conceptually a persuasive argument. Yet, in a broader societal context, justification should be required for the values, purposes, and presumptions underlying the "group" classification scheme that provides the basis for borrowing information. Additional evidence may be required to establish the credibility and accuracy with which individuals are classified.

In any event, an observed achievement test score is often interpreted as the level of proficiency attained relative to the content and design of a test. A regressed true score estimate is no longer uniquely connected with a level of proficiency defined in terms of test specifications. It is also a function of collateral information. Thus, a regressed score may be incompatible with the original purpose of a test. For example, some tests are developed with criterion levels for student performance such as partially proficient, proficient, and advanced. In some contexts, this classification is analogous to a legally binding contract, and failure to attain proficiency may have enforceable consequences. Because of statistical bias, shrinkage estimators may reduce classification accuracy. By analogy, their use in an evaluation would modify the terms of the contract, possibly rendering it unenforceable. As shall be argued in this chapter, shrinkage may make sense in some contexts but not in others.

ISSUE 1: 3PL IRT SCORING

Three IRT models, or variations of them, are typically used in many assessment or licensure programs. For a 0 (wrong) or 1(right) response (say U_{ij}) on item i by person j, the 3PL model for the probability of a correct response λ is given by

$$\lambda(U_{ij} = 1 \mid \theta_j) = c_i + (1 - c_i)P_{ij}$$
$$P_{ij} = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}, \tag{27.3}$$

where a, b, and c are the discrimination, difficulty, and (so-called) guessing parameters, respectively (for elaboration, see Volume 1, Chapter 6, this handbook). This nonstandard notation is used so that one can represent the counterfactual probability that would be obtained without guessing as P. A 1PL model is obtained by setting $c = 0$ and a to a constant for all items. The 2PL model allows both a and b to vary across items, with $c = 0$ fixed. For logistic IRT models, it can be shown that optimal scoring weights to obtain a total test score (Birnbaum, 1968; Chiu, 2010) take the following general form:

$$\tilde{S}_j = \sum_{i=1}^{n} a_i \rho_{ij} u_{ij}, \tag{27.4}$$

where

$$\rho_{ij} = \frac{1}{1 + c_i Q_{ij} / P_{ij}} = \frac{\sigma_{ij}^2}{\sigma_{ij}^2 + (1 - P_{ij})^2 / c^{-1}}, \tag{27.5}$$

and the observed item response u_{ij} is either 1 (correct) or 0 (incorrect). As suggested in Table 27.1,

the term $(1 - P_{ij})^2$ in Equation 27.5 is parallel to the error variance of the Kelley estimator because shrinkage only occurs for correct answers, and c^{-1} plays a role roughly equivalent to that of the sample size.

Under the 1PL and 2PL models $\rho = 1$, the resulting weighted total scores in Equation 27.4 are sufficient statistics, labeled *S*, and items make independent contributions to *S*. For the 1PL model, *S* is the simple sum of item responses, and for the 2PL model, the item scores are weighted by the discrimination coefficients *a*. Not so in the 3PL model. As Chiu and Camilli (in press) explained, as the probability of answering a test item *incorrectly* increases, the factor $c_i Q_{ij} / P_{ij}$ in the denominator of Equation 27.5 increases; in turn, the credit given to a correct item response diminishes. For example, suppose two individuals take a 50-item multiple-choice test with items ordered from least to most difficult. The first individual responds correctly only to item 50 and to items 1 to 5, whereas the second individual responds correctly to item 50 and to the majority of the first 49 items. According to the 3PL scoring result, the first individual will receive less credit for a correct response to the 50th item. Equation 27.4 is completely consistent with the standard maximum likelihood estimate of θ. That is, the formula is derived by taking the first derivative of the likelihood function and setting the result to zero as shown in Appendix 27.1. It is simply easier to understand the impact of the *c* parameter in a raw score metric.

With 3PL scoring, credit for a correct response shrinks toward zero as a function of the guessing parameter and the odds of an incorrect response, which in turn depends upon θ. The issue here is not the internal consistency of the 3PL model. In fact, the shrinkage may be useful in reducing the effect of unlikely correct answers for some individual test takers. Rather, the main concern here is whether total scores should be obtained with item weights involving the proficiency of an examinee, rather than item weights based strictly on test design specifications.

ISSUE 2: IMPLICATIONS FOR CONTENT-BASED INTERPRETATION OF SCORES

The 3PL item weights have significant implications for interpreting scores. Item credit shrinkage will de facto alter the design specifications of an assessment, which generally consist of content and processes. For instance, the content areas of a mathematics test might be defined in terms of number sense, measurement, algebra, probability, and so forth, whereas processes might be defined as procedural, conceptual, and problem solving. These two facets are typically crossed in a test design matrix, and a test is initially designed by determining what proportion of total items will fall into each cell of the matrix. Typically, each cell of the test specifications has a weight by fiat—based on the number of items specified for each cell. With 2PL and 3PL scoring, the cells are implicitly reweighted to some degree. Consider the case of the 3PL model in the presence of guessing on geometry items. If more difficult items are associated with geometry on a test, then this content is essentially omitted from the test specifications for low-scoring individuals: They receive little credit for such items even for correct responses. In other words, some cells of the content–process domain are essentially inoperative for some individuals in the presence of guessing, and score interpretation should be suitably amended.

Scoring based on the 2PL or 3PL IRT model reweights the item specification blueprint. For the 2PL model, the weight of any cell *K* of the test specification matrix for a test with *N* items is

$$w_{jK} = \sum_{i \in K} a_i \rho_{ij} \Big/ \sum_{i=1}^{N} a_i \rho_{ij} = \sum_{i \in K} a_i \Big/ \sum_{i=1}^{N} a_i \quad (27.6)$$

because $\rho = 1$ for all items. Content represented by more highly discriminating items is prioritized in determining individual scores, but the reweighting of the test specifications is identical for all individuals *j*. For the 3PL model, Equation 27.6 does not simplify. The reliabilities ρ_{ij} vary by individual as do the cell weights w_{jK}. This implies some degree of individual variation across identical cells of the content–process matrix, which defines the domain of generalization for test score interpretation.

The shrinkage of item credit with the 3PL model may produce a more accurate estimate of proficiency, especially if the assumption of unidimensionality is appropriate. But reweighting may work against individuals with uneven profiles of ability across the represented test content. For example,

suppose a student named Corrado never memorized the multiplication tables but enjoyed the mathematics of geometry. With many incorrect answers to easy multiplication questions and a few correct answers to harder abstract questions that have nonzero *c* parameters, Corrado would receive less credit under 3PL scoring for geometry items than his raw score might indicate. For a more extreme example, suppose Corrado answered correctly the four easiest items on a test as well as the four most difficult. In this case, the credit given for the latter items would be substantially downweighted.

This example may make logical sense for some purposes, and if so, the 3PL estimate of proficiency provides a sensible scoring option. For other purposes, the proficiency estimates of 2PL and 3PL models may mistakenly shift test score interpretation away from the test content–process specifications. Moreover, 3PL proficiency estimates may conflict with an unspoken assumption that students receive the same credit for the same correct answers, an assumption that is consistent with nominal test specifications (that are usually carefully described in technical manuals). Although the issue of 3PL unfairness—or fairness, depending on one's perspective—is easily ignored, a testing staff should be highly knowledgeable on scoring issues. Thus, an ethical issue arises in the extent to which potential reweighting occurs with 2PL and 3PL proficiency estimation. How should such information be taken into account by assessment staff, and how should such information be communicated to stakeholders? This is not a purely conceptual question because a number of computer-assisted testing programs and standardized achievement test batteries rely on the 3PL model. Moreover, because *c* parameters are more poorly estimated than other item parameters, estimation error and item misfit may exacerbate item-weighting issues.

ISSUE 3: DIFFERENTIAL ITEM FUNCTIONING

One well-established set of procedures in test quality control is that of DIF. To obtain indices of DIF, groups must first be defined. In the two-group situation, these are usually termed the *reference* and *focal* groups, where the focal group often represents a racial or ethnic minority. As implied by the acronym DIF, the units of analysis are the individual items constituting a test. Accordingly, the main concern is an examination of measurement invariance, which basically holds that any individual with the same proficiency should receive, on average, the same credit for an item response regardless of group membership (Camilli & Shepard, 1994). Qualitative procedures for examining potential test items for cultural sensitivity (see Camilli, 2006) are usually combined with statistical methods of DIF for examining the suitability of test items.

Holding proficiency constant, DIF analysis is used to identify items whose properties (usually the difficulty or the IRT *b* parameter) change across defined groups. Estimators of DIF are commonly implemented by comparing the proficiencies for different groups of individuals at the same observed total score on a test. Such procedures fall under the rubric of observed-score indices of DIF. In contrast, IRT procedures are conceptually based on the latent proficiency θ.

For example, consider a test item intended to measure reading comprehension. Students first read a passage having to do with snowstorms and then reply to a multiple-choice item based on the passage. If students from different groups who are equally proficient—that is, who have the same total score—have differential probabilities of a correct response, then the item is said to exhibit DIF. Further investigation is required to determine the cause of DIF. If, for example, it is discovered that students in tropical and subtropical climates have a lower probability of a correct response than students from temperate regions, then the item exemplifies unwanted DIF, possibly because of cultural or geographical insensitivity. Such items are marked for modification or deletion in the test assembly process.

Shrinkage Estimators for DIF

Many kinds of DIF statistics can be expressed as shrinkage estimators, and it is not a current purpose to provide an exhaustive review. Instead, the goal is to examine a proposal advanced by several researchers (Sinharay et al., 2009; Zwick et al., 2005) who have investigated the stability of

shrinkage estimators based on the notion of mean square error. Given a generic standard DIF estimator for item j, say D_j, the shrinkage formula is given in the fourth column of Table 27.1. The population mean of the DIF estimators for a set of items is assumed to be zero, and must be, as explained, close to zero empirically for technical reasons (Camilli, 1993). The effect of shrinkage is then to move all DIF estimators much closer to zero—because ρ is often substantially less than 1.0. In this case, the coefficient ρ has a less straightforward interpretation. Parallel to the Kelley estimator, here the DIF estimator is an observed "score" that is being regressed back toward a value of zero. In Table 27.1, note that σ_j^2 is the variance error of $\hat{D}_j$, and σ_D^2 is the variance of D_j across items (Camilli & Penfield, 1997; Longford, Holland, & Thayer, 1993).

Trading off False Positives and False Negatives

The issue here is not whether DIF shrinkage formulas are internally consistent (they are) or whether they reduce mean square error (they do). Rather, two other distinct considerations extend beyond the mathematics of DIF. The first is whether the shrinkage strategy is consistent with the purpose and function of DIF statistics in an assessment program, given the function of DIF as an indicator of potential cultural bias on tests. Shrinkage always reduces the absolute level of DIF in an item, and according to commonly applied classification rules (Educational Testing Service, 2002), fewer items will be identified for further scrutiny. Thus, a narrow focus on statistical properties omits altogether consideration of the relative importance of false positive (incorrectly identifying an item as exhibiting DIF) and false negative error rates (missing items with true DIF). This discussion should be at the core of shrinkage estimators for DIF. In fact, as shown by Sinharay et al. (2009), virtually no items would be identified as having actionable amounts of DIF after shrinkage has been applied. If the object of an investigation is to improve the fairness of a test, what good is a rule that minimizes false positive errors at the expense of false negative errors?

The second issue is that the shrinkage is always toward zero, but it is not possible to determine whether zero is the correct value. A nonzero parameter δ for DIF could be added to an IRT model by setting $b_i' = b_i - \delta$, but any value of b_i' could be paired with $\theta_i' = \theta_i - \delta$ without altering the IRT probability. Thus, $\theta_j - b_i = \theta_j' - b_i'$ and δ is subsequently not identified. It could be argued that $\delta = 0$ is a highly plausible value for the distribution of DIF, but there is no way to avoid the fact that the result of the DIF analysis is presumed to obtain a shrinkage estimate. This would seem to be a particularly unfortunate decision if the twin purpose is to discover DIF *and* to convey the information that an *unbiased* process of discovery was used.

Social Issues with DIF and Item Selection

If the aim of DIF analyses is to identify potential cultural bias, why shrink at all? Estimators of DIF are difficult enough to communicate to a public vested in test outcomes. But with shrinkage estimators, DIF can no longer be defined as differential group probabilities of a correct response at the same total score, and the efforts to create fair tests will be further obfuscated. It would be better to live with standard DIF indices and to accept higher rates of false positive errors to reduce the rate of false negative errors. In research studies, however, shrinkage may result in DIF statistics with more desirable properties, especially if the goal is to avoid Type 1 errors.

From a social rather than statistical point of view, shrinkage may be an undesirable property. Rightly or wrongly, DIF statistics are used to make the claim that test developers have taken steps to minimize test bias. This need must be weighed against the need to minimize the role of sampling variability. But if the many facets of measurement and sampling error were taken into account, it is likely that any DIF estimate could be shrunken to virtually zero. This result would certainly make life easier in test development: Fewer items would present DIF problems, and expenses associated with replacing items or rescoring tests might be more readily contained. It is not clear, however, that this approach to DIF adequately addresses the issue of false negatives in the broader context of test fairness.

ISSUE 3: SELECTION

Drawing on previous work by Millsap (1998, 2007), Borsboom et al. (2008) provided a deeper look into the impossibility of selection variance. Much telescoped, the argument is that in the presence of measurement error, different types of selection rates cannot be invariant across groups differing in proficiency. Before discussing the fairness implications of this argument, some background on selection methods is useful.

There are four types of decision relative to criterion scores in selection applications, and a true positive decision is characterized by an examinee being selected on the basis of a predictor (the "positive" part) while also exceeding the criterion (the "true" part). Likewise, there are also false positives, false negatives, and true negatives. This set of outcomes is given in the 2 × 2 table in Figure 27.1. Four types of conditional probabilities can then be represented as the marginal probabilities in this table as shown in Figure 27.2. For example, I/(I + II) is the probability of selection given success defined as the criterion score exceeding the criterion cutoff (sensitivity), with III/(III + IV) as the probability of being below the predictor cutoff given a score below the criterion cutoff (specificity). Selection models are defined by setting cutoff scores for groups that result in the same conditional probability across groups. For example, according to the equal probability model of Linn (1973), the proportion of successful performers for selected applicants, I/(I + IV), should be the same across groups. The equivalence is established by fixing a value for the proportion and then implementing different cut scores for different groups (see Petersen & Novick, 1976, for a more complete treatment).

According to Borsboom et al. (2008), all of the marginal probabilities are required to be the same across two or more populations for selection invariance to hold. However, as Petersen and Novick (1976) realized,

> Since it can be shown that only under certain special conditions equating K [the conditional probability of selection given success] among subpopulation leads to equating $\overline{K}$[the conditional probability of rejection given failure], and vice versa . . . it might be suggested that in order to take into account both aspects of the culture-fair selection issue . . . into consideration, we should at least contemplate equating some combination of K and $\overline{K}$ instead of trying to equate, independently, either K or $\overline{K}$ among populations. (p. 21)

In other words, selection invariance holds only under special and rare circumstances, even in the case of measurement invariance framed in terms of item characteristics (see Petersen & Novick, 1976, pp. 21–22, for a mathematical demonstration). Petersen and Novick also evaluated the equal risk model of Einhorn and Bass (1971) and determined that only this model and its converse imply the same selection strategy. What Millsap (1998) and Borsboom et al. (2008) have essentially shown is that with measurement error, not even the equal risk model maintains invariance across populations. Thus, for no currently known method of selection can invariance be maintained across groups with different proficiency distributions. Previously, the regression invariance assumption and its relationship to measurement error had been demonstrated in the context of the analysis of covariance by Lord (1960) and Cronbach, Rogosa, Floden, and Price (1977).

To illustrate this in a simulation, suppose that measurement invariance is defined such that item IRT parameters are equivalent across groups of interest. Suppose further that 0–1 (wrong vs. right) item responses for a 25-item test are simulated by a standard IRT model. According to IRT, the actual proficiency, labeled θ (rather than *Y*, which is more commonly encountered in the selection literature) is measured by a number of items, and item responses are conditionally independent given θ. In this simulation study, θ is taken to represent an ideal

	Reject $< X_c$	Accept $\geq X_c$
Suitable $\geq Y_c$	II: False Negative	I:True Positive
Not Suitable $< Y_c$	III: True Negative	IV: False Positive

FIGURE 27.1. Selection outcomes with A–D as unconditional selection probabilities.

Description	Quantity	Label	Selection Model
P(S|A)	I/(I+IV)	Positive Predictive Value	Equal Probability
P(¬S|¬A)	III/(II+III)	Negative Predictive Value	Converse Equal Probability
P(A|S)	I/(I+II)	Sensitivity	Conditional Probability
P(¬A|¬S)	III/(III+IV)	Specificity	Converse Conditional Probability

Note: Suitable or Success=S, Accept=A, ¬ = not.

FIGURE 27.2. Evaluation probabilities for four selection models.

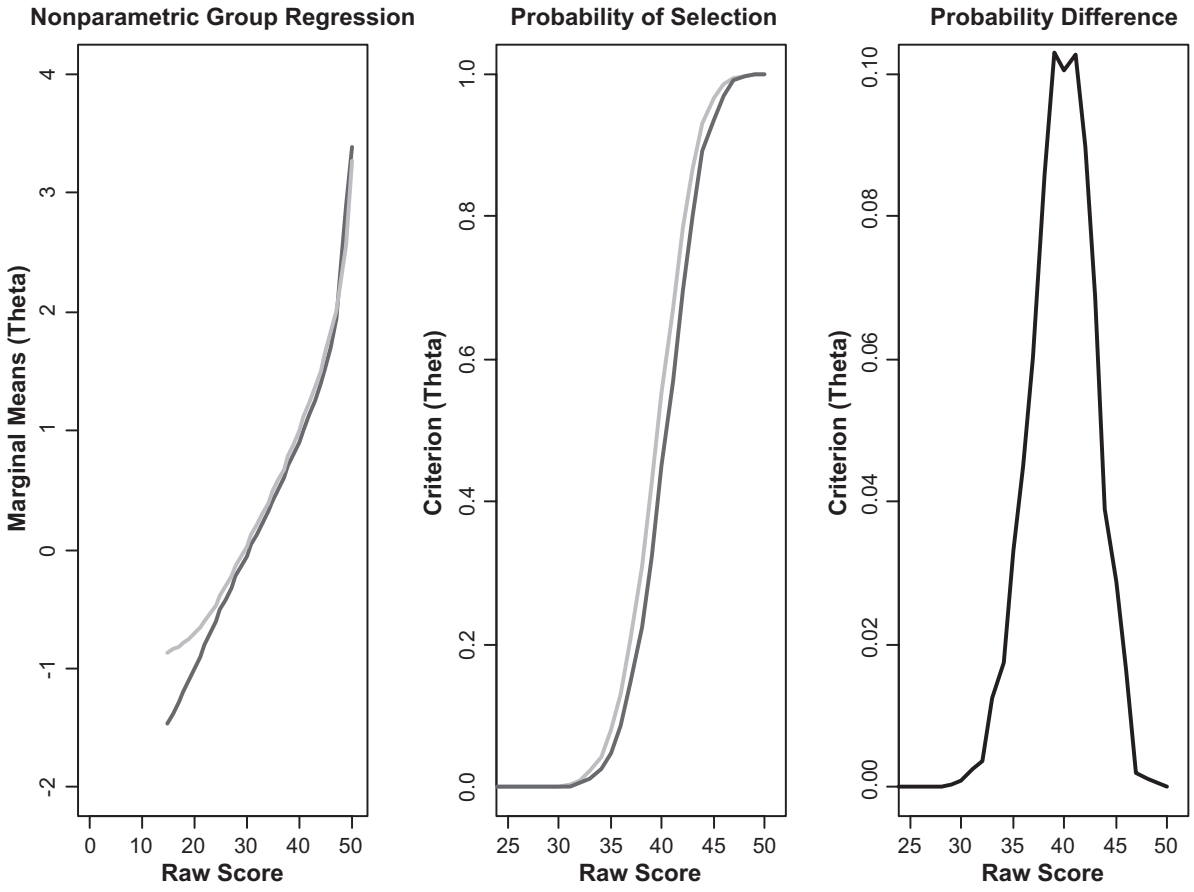


FIGURE 27.3. Regressions and success probabilities for the 1PL model.

criterion, and the total score is taken to represent the predictor. Accordingly, a test with measurement error is used to predict a criterion in a selection situation. Finally, suppose there are two groups of 25,000 with a one standard deviation difference on the criterion θ. This scenario is intended to be illustrative only and the problems it reveals are likely to be less serious than what otherwise would occur in more realistic predictor–criterion setups. A large sample size is used to minimize the effects of random variability in graphic representations.

In the first panel of Figure 27.3, θs (the criterion values) are plotted against raw scores (the predictor values) for two groups using a 1PL model with a one standard deviation difference on θ. All figures were obtained by averaging selection probabilities over 50 iterations. Nonparametric regression curves are added to speed interpretation. Item responses were generated under the condition of measurement invariance: The item parameters were exactly the same for the two groups by design. It is immediately apparent in the first panel that the group regression

lines are different, with the regression curve for the group with the higher average on the criterion ($\mu_\theta = 0$) above that for the group with the lower average ($\mu_\theta = -1$). As Borsboom et al. (2008) observed, the predicted criterion scores are regressing to their different group means. Thus, for members of the lower scoring group, the regression curve shifts to the right along the *x*-axis, or equivalently for the higher scoring group, the regression curve shifts to the left. In contrast to the analysis offered by Petersen and Novick (1976), these figures demonstrate that with measurement error, Cleary's (1968) definition of test fairness as equal group regression lines cannot be theoretically obtained in the presence of group differences. Empirically, differential prediction is the exception rather than the rule; however, the differences examined in the following paragraphs are rarely examined explicitly in practice.

In the second panel of Figure 27.3, the cumulative probabilities of success are given for each group across the range of observed scores, *x*. This is specified as the proportion of criterion scores $\theta > 1.5$ at each value of *x*, where $\theta = 1.5$ separates individuals that exceed the criterion from those who do not. The difference in regression curves results in a notable difference in the probability of exceeding the criterion for the two groups for any particular observed score. In the third panel of Figure 27.3, it is shown that the difference in the group probabilities of success is largest (almost .1 units) at some locations on the *x*-axis, despite the apparently small difference in regression curves in the middle panel. Because the overall reliability of the test is quite high at about $\rho = .9$, the result seems particularly striking. Parallel charts are given in Figure 27.4 using a 3PL model, and the results are highly similar. Thus, the phenomenon does not depend on the type of IRT model chosen.

In introducing the topic of differential prediction conceptually at the beginning of this chapter, the

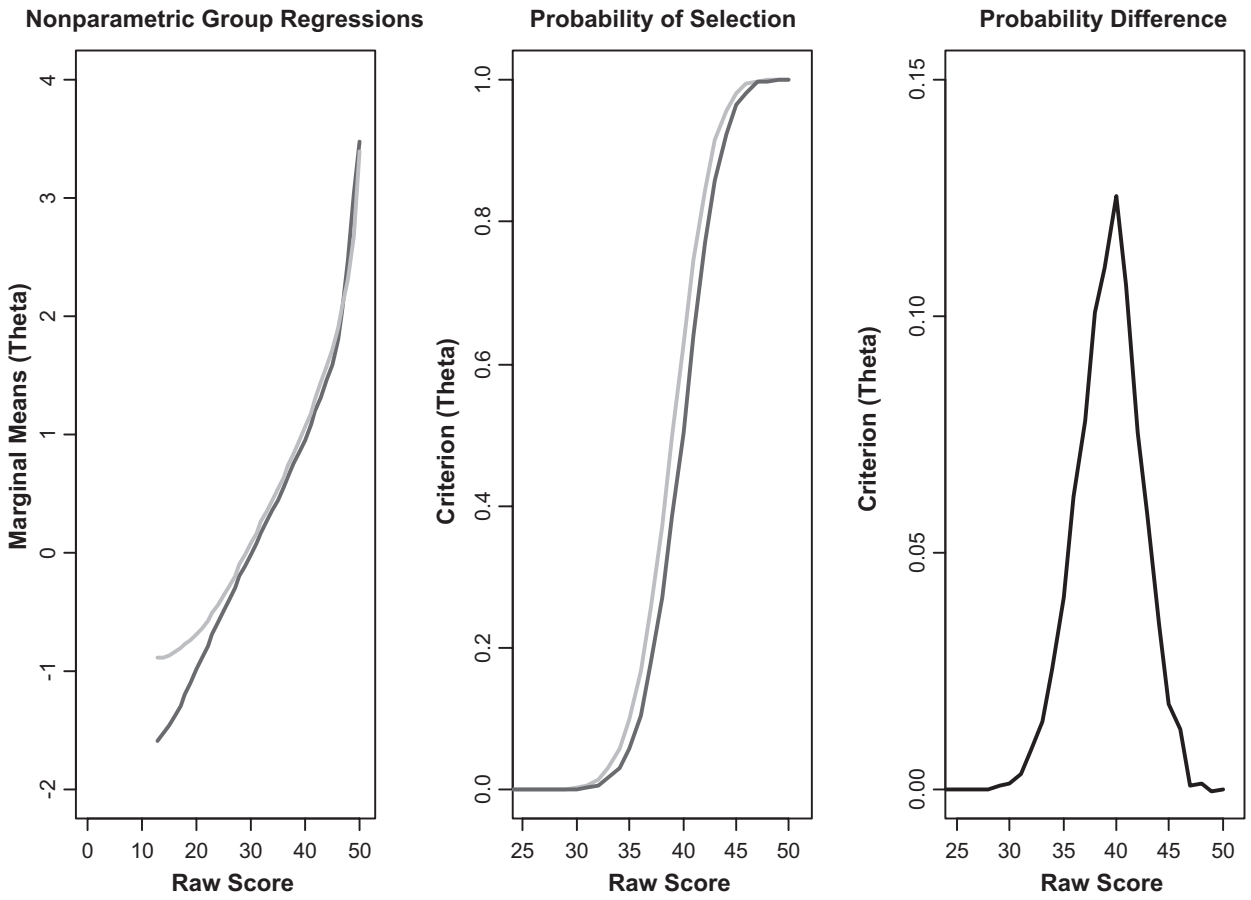


FIGURE 27.4. Regression and success probabilities for the 3PL model.

axes of the standard presentation were reversed. Here we add that if θ appeared on the *x*-axis and Y on the *y*-axis, then in the simulations described earlier, the group regression lines are nearly the same. That is, when one conditions on the "true" measurement of interest θ, the group averages on Y are about equal—even though two regressions emerge when the axes are switched. The difference is that in the standard prediction model as used in the simulations, Y is used as the conditioning variable and group averages on θ are compared. In this case, a group difference on θ emerged precisely because Y is an imperfect variable for conditioning due to measurement error.

The basic regression-to-the-mean result has been well understood (Kelley, 1927) for some time, and the question of interest to this discussion is whether there is an acceptable solution to the problem. Borsboom et al. (2008) made several recommendations for addressing the regression-to-the-mean problem in selection, including the use of multiple indicators and test design. In this section, the potential use of a Kelley shrinkage estimator to adjust for artifacts of regression to the mean is considered. Theoretically, different cuts cores could be used to adjust for regression artifacts, but it is conceptually more elegant to make the adjustment to test scores. Because of the different group distributions of θ, regressed scores of the higher group would shift up, and regressed scores of the lower group would regression down. An alternative way of describing this phenomenon is that the higher group's regression curve would shift up the θ axis, and the lower group's regression would shift down. For example, in the first panel of Figure 27.3, using a regressed true score for the predictor would shift the regression curve of the higher scoring group up, and the two curves would converge.

Could regressed true scores be implemented in assessment and licensure programs to reduce the effect of measurement error? Probably not, for several reasons that illustrate how measurement theories collide with social realities, that is, the concept of "groups," and societal perceptions of the meaning of test scores.

The Existence of Groups

The traditional selection situation has a simple setup in which all stakeholders understand the conditions for success, namely, that individuals at any value of the predictor score have an equal chance of being selected. Modifying this process to take into account group membership would have the effect of adding a group membership label to the selection criteria. Yet the Supreme Court in *Grutter v. Bollinger* (2003) and *Gratz v. Bollinger* (2003) ruled that an explicit point system for student selection based on group membership was unconstitutional. The argument in this case concerned adding points based on race or ethnicity, and it seems fairly straightforward to conclude that the Court would take an equally dim view of subtracting points.

In particular, any use of racial classification (which is considered as *suspect* from the outset in cases involving constitutional issues) must withstand the test of "strict scrutiny" (see Raza, Anderson, & Custred, 1999). Moreover, in the United States and many other countries, the groups in question exist as a distinction in race, ethnicity, or some combination of the two. In practice, the task of identifying the group membership of a person is fraught with difficulty (Camilli, 2006).

Societal Perceptions

Societal perception of the meaning of test score may conflict with the statistical efficiency of the regressed true score estimate. As summarized by Petersen and Novick (1976),

> Despite our well-displayed fondness for the Bayesian model . . . estimation of means, we must acknowledge there can be a problem. It may add to overall efficiency to reduce our estimate of a person's true score because we identify him with some population that has a lower mean true score, but it may not appear fair. Suppose in a selection situation, one person has his score lowered by this regression to the population mean and a second person from a population with a higher mean true score has his score raised. Suppose further this results in an inversion in the ordering of the reported scores and that, as a result, the second person is selected for college admissions and the first is not. We would

> certainly be hard put to convince the first examinee, his parents, *and his lawyer* that he had been treated fairly. (p. 5)

Rather than a legal issue, the concern in the preceding is whether the potential benefits of regressed true scores could be communicated well enough to allow for the possibility of general societal acceptance. This seems doubtful. A narrative in which equal performances on a qualification test do not necessary equate to equal scores for predicting success runs a high risk of signaling an arbitrary and capricious selection process.

DISCUSSION

In all of the topics discussed in this chapter, a statistical theory initially seems to provide the technology to garner better measures of student, candidate, or test item effects. Yet shrinkage estimators may be inconsistent with their intended use as tools for improvement, and in some cases, unintended negative consequences may ensue (Messick, 1989, 1994). Moreover, this problem is especially difficult to remedy if there is no construct for distinguishing positive from negative consequences. Very little effort has been expended to formulate a conceptual map of how shrinkage estimators in certain educational application relate to the purpose for which a quantitative estimator is used. Rather, rationales for these estimators have simply been borrowed from a general statistical theory. In any event, the role of validity theory has necessarily become more challenging given the prevalence and degree of technological change to educational evaluation.

Without an adequate concept map, there is no intention or purpose that can be verified or challenged empirically with respect to a statistical index. The onus is thus on stakeholders to establish guiding principles for the intelligent use of shrinkage estimators in educational contexts. This requires clear specification of strengths and weaknesses of shrinkage estimators in a language that is accessible to policy implementers as well as to those being evaluated. Toward this effort, concluding suggestions and reflections are proffered.

3PL Proficiency Estimation

Other than improvement in model fit, little justification for the 3PL has been forthcoming. Item credit shrinkage—or more accurately, variable reduction in credit—is a virtually unknown phenomenon. Scoring with this model to some degree disconnects nominal and actual test specifications. Moreover, a satisfying solution to 3PL shrinkage is not obtained by simply reverting to a 1PL model: Differences in the suitability of measurement models are constrained by differences in the design of instrumentation to collect item responses.

The properties of the 3PL model bring the *Rasch debate* (Fisher, 1994) into focus. Fisher (1994) ascribed an objectivist view of measurement to the users of multiparameter IRT models. In particular, he singled out Lindquist (1953) who argued that

> from the point of view of the tester, the definition of the [test] objective is sacrosanct; he has no business monkeying around with that object. The objective is handed down to him by those agents of society who are responsible for decisions concerning objectives, and what the test constructor must do is to attempt to incorporate that definition as clearly and exactly as possible in the examinations that he builds. (p. 53)

Although Fisher (1994) contended that such a point of view has led to the popularity of 2PL and 3PL models, as shown in this chapter, the 1PL model is more—not less—faithful to Lindquist's criterion. To avoid potential distortions of the measurement map inherent in the test blueprint, standard tests must be assembled with items *without lower asymptotes*, such as open-ended questions. They must also be constructed with items having *equal discrimination parameters*. It is likely that only constructed-response items can satisfy these criteria, and only then would test scoring based on a 1PL-family model improve the interpretability of test scores. Regardless of the impracticality of this approach to assessment, the essential point of disagreement in the debate is the degree to which test item content should supersede a test development process centered on *measurement of a construct*. As

Fisher might argue, the simple act of using a 1PL model to assemble and score a test does not ensure the empirical existence of a nominal construct.

Shrinkage and DIF

Shrinkage estimation in the case of DIF, using the minimum mean square error justification, involves an associated trade-off between false positive or Type 1 errors (deciding whether the item shows DIF when it actually does not) and false negative or Type 2 errors (the item actually has DIF that was not detected). It is true that many scientists prefer the practice of minimizing Type 1 errors, but the only way to prevent such errors from occurring with near certainty is to reject the hypothesis of nonzero DIF in virtually all cases. Shrinkage estimators of DIF would serve this purpose most efficiently. Thus, a deeper look at the rationale for shrinkage estimators of DIF reveals the need for conceptual justification in the framework of test fairness.

Shrinkage and Selection

There is little chance that shrinkage estimators for individuals will ever be used in high-stakes selection situations. They may be useful, however, in other situations in which guidance is the main concern. Especially if there is a route for the self-correction of selection decisions in educational programs, then shrinkage estimators could improve the efficiency with which educational services are delivered. In research applications, where the goal is to evaluate hypotheses about the effects of selection, the use of shrinkage estimators may reduce Type 1 errors resulting from regression to the mean.

One topic that this chapter has not considered is the issue of shrinkage in estimates associated with value-added models (VAM). Recently, a great deal of interest has been generated regarding the use of VAM, and shrinkage estimators resulting from random-effects models have been advocated for obtaining measures of teacher and school effects. As noted by McCaffrey et al. (2003),

> the random-effects method "shrinks" the estimate based on the given teacher's students toward the overall mean for all students. On average, shrinking the estimate has optimal statistical properties across teachers but can be sub-optimal for teachers whose effects are far from the mean. Fixed-effects estimates can be highly sensitive to *sampling error*. (p. xvi)

Much of the discussion in this chapter could be extended easily to VAM estimators with the conclusion that statistical theories of "optimal" estimation do not necessarily provide a substantive rationale in applications to student and teacher evaluations. An important fairness issue is involved because the match may be incomplete between the intended construct and the statistical estimator. That is, the VAM score may contain either irrelevant variance or construct underrepresentation in the sense of Messick (1994).

Some have considered the VAM approach to have great potential for disentangling the myriad influences on student achievement (e.g., Harris, 2009), whereas others (e.g., Briggs, 2008; Hill, 2009; Tate, 2004) have been more circumspect. Briggs (2008) argued that

> when the educational intervention under investigation is parameterized as a teacher or school, the interpretation of the associated VAM residual as a descriptive measure rather than a causal effect shifts the technical conversation from a consideration of *internal validity* to a consideration of *construct validity*; from *statistics* to *psychometrics* (p. 5).

To be interpreted as intended, for example, a VAM estimate (or residual) might require a vertical scale or a test designed to measure developmental change. The sum total of all such requirements, and the relationships between them, comprise the VAM construct. Likewise, Hill (2009) argued that if a VAM estimator is cast as a measure of teacher quality, it should, in turn, be validated for this purpose just as any other test score is validated against a proposed interpretation. We would add that if a shrinkage estimator is adopted, a policy-relevant justification should be provided. It should be recognized that validation is not attained by simply listing the desirable and undesirable properties of VAM estimators, and

in addition, that particular estimators should be validated for particular purposes. Indeed, it is the validity argument in this case that is the fairness issue in high-stakes applications of VAM.

In all of the issues considered herein, a tension exists between the need to employ statistical procedures pragmatically and the desire to provide a scientifically defensible rationale. Yet the theoretical statistical properties of an estimator are not sufficient justification for its use. Rather than performing in restricted roles as technicians, psychometricians are well qualified to take more active roles in examining the theory of shrinkage estimators relative to their intended uses and interpretations in assessment contexts. Indeed, theoretical and practical training in test validity would seem to be prerequisite for this important work. Research linking theory and practice is certainly under way in some areas, such as VAM, but more work in other areas is necessary.

APPENDIX 27.1: ESTIMATION OF 3PL PROFICIENCY

In a 3PL IRT model, the probability λ_{ij} for the examinee with a certain ability level θ_j to answer a particular item right can be represented as

$$\lambda_{ij}\left(u_{ij}=1\middle|\theta_j,a_i,b_i,c_i\right)=c_i+\left(1-c_i\right)P_{ij}, \quad (27.7)$$

where P_{ij} is defined in Equation 27.3. For a test of n dichotomous items, the log likelihood of a response pattern for an examinee is given by

$$\begin{aligned} F &= \ln\prod_{i=1}^{n}\lambda_{ij}^{u_{ij}}(1-\lambda_{ij})^{1-u_{ij}} \\ &= \sum_{i=1}^{n}\left[u_{ij}\ln\lambda_{ij}+(1-u_{ij})\ln(1-\lambda_{ij})\right], \end{aligned} \quad (27.8)$$

with $u_{ij}=1$ or 0 for a correct or incorrect response, respectively. An estimated proficiency is obtained by maximizing this function with respect to θ. The log likelihood function can then be written as

$$F=\sum_{i=1}^{n}u_{ij}\ln\left[P_{ij}\left(1+\eta_{ij}\right)\right]+\left(1-u_{ij}\right)\ln\left[1-P_{ij}\left(1+\eta_{ij}\right)\right]$$

$$\eta_{ij}=c_i\left(\frac{1-P_{ij}}{P_{ij}}\right). \quad (27.9)$$

Differentiating F with respect to θ, setting the result equal to zero, and simplifying gives

$$\sum_{i=1}^{n}a_iP_{ij}=\sum_{i=1}^{n}a_iu_{ij}\left(1+\eta_{ij}\right)^{-1}, \quad (27.10)$$

where the left side of Equation 27.10 gives the weighted true score estimate with adjustment for the lower asymptote parameters (Chiu & Camilli, in press). The quantity

$$w_{ij}=a_i\left(1+\eta_{ij}\right)^{-1} \quad (27.11)$$

can also be derived as the locally best weight as shown by Birnbaum (1968, Section 19.3).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

ANSER Analytic Services, Inc. (2000, September). *Technologies for identifying missing children, final report* (National Institute of Justice Report 97-LB-VX-KO25). Retrieved from http://www.ncjrsgov/pdffiles1/nij/grants/186277.pdf

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Borsboom, D., Romeijn, J. W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods, 13*, 75–98. doi:10.1037/1082-989X.13.2.75

Braun, H. (2006). Empirical Bayes. In J. Green, G. Camilli, & P. Elmore (Eds.), *Complementary methods for research in education* (pp. 243–258). Washington, DC: American Educational Research Association.

Briggs, D. C. (2008, November 13–14). *The goals and uses of value-added models.* Paper presented to the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation and Educational Accountability, National Research Council and the National Academy of Education, Washington, DC.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice.* (pp. 397–417). Hillsdale, NJ: Erlbaum.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport, CT: American Council on Education/ Praeger.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on the Mantel-Haenszel log-odds ratio. *Journal of Educational Measurement, 34*, 123–139. doi:10.1111/j.1745-3984.1997.tb00510.x

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Hollywood, CA: Sage.

Chiu, T.-W., & Camilli, G. (in press). Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement*.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. doi:10.1111/j.1745-3984.1968.tb00613.x

Cronbach, L. J., Rogosa, D. R., Floden, R. E., & Price, G. G. (1977). *Analysis of covariance in nonrandomized experiments: Parameters affecting bias.* Occasional Paper, Stanford Evaluation Consortium, Stanford University, Stanford, CA.

Debra P. v. Turlington, 644 F. 2d 397 (5th Cir. 1981).

Educational Testing Service. (2002). *ETS guidelines for fairness review of assessments*. Princeton, NJ: Author.

Einhorn, H. J., & Bass, A. R. (1971). Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, *75*, 261–269. doi:10.1037/h0030871

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research and Perspectives, 6*, 155–189. doi:10.1080/15366360802197792

Fisher, W. P., Jr. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective measurement* (Vol. 2, pp. 36–72). Norwood, NJ: Ablex.

Gratz v. Bollinger. (2003). 39 U.S. 244.

Grutter v. Bollinger. (2003). 539 U.S. 306.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management, 28*, 693–699. doi:10.1002/pam.20464

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management, 28*, 700–709. doi:10.1002/pam.20463

Kelley, T. L. (1927). *Interpretation of educational measurements*. Yonkers on Hudson, NY: World Book.

Lindquist, E. F. (1953). Selecting appropriate score scales for tests (Discussion). In *Proceedings of the 1952 Invitational Conference on Testing Problems* (pp. 160–169). Princeton, NJ: ETS.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43*, 139–161.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–197). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1960). Large-scale covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55*, 307–321.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Millsap, R. E. (1998). Group difference in intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33*, 403–424. doi:10.1207/s15327906mbr3303_5

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*, 461–473. doi:10.1007/s11336-007-9039-7

Novick, M. R. (1970). Bayesian considerations in educational information systems. In G. V. Glass (Ed.), *Proceedings of the 1970 Invitational Conference on Testing Problems* (pp. 77–88). Princeton, NJ: Educational Testing Service.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13*, 3–29. doi:10.1111/j.1745-3984.1976.tb00178.x

Raza, M. A., Anderson, A. J., & Custred, H. G., Jr. (1999). *The ups and downs of affirmative action preferences.* Westport, CT: Greenwood Press.

Sinharay, S., Dorans, N., Grant, M., & Blew, E. (2009). Using past data to enhance small sample DIF estimation: A Bayesian approach. *Journal of Educational and Behavioral Statistics, 34*, 74–96. doi:10.3102/1076998607309021

Sloane, F. (2008). Multilevel models in design research: A case from mathematics education. In A. Kelly & R. Lesh (Eds.), *The handbook of design research in education* (pp. 459–476). New York, NY: Taylor & Francis.

Zwick, R., Thayer, D. T., & Lewis, C. (2005). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1–28. doi:10.1111/j.1745-3984.1999.tb00543.x

CHAPTER 28

FUTURE DIRECTIONS IN ASSESSMENT AND TESTING IN EDUCATION AND PSYCHOLOGY

John Hattie and Heidi Leeson

When Gauss, in the 1800s, described the normal distribution, thence establishing theories about measurement error, he opened the door for developments of classical test theories. The beginning of the current measurement models commenced in the 1900s when classical test theory (CTT) was developed. Traub (1997) noted that this theory was born of a ferment relating to three remarkable achievements: recognition of the presence of errors in measurement, a conception of error as a random variable, and a conception of correlation and how to index it. Spearman (1904) was most involved in these three contributions, and for the next 100 years, many others improved, adapted, researched and used this classical model. There were developments of the correction for attenuation, estimates of reliability (e.g., Cronbach's synthesis of these ideas in 1951), concurrent and criterion concepts of validity, and so many methods about how to create items to best meet the optimal criteria for classical modeling. Embretson (2004) in her review of measurement in the 19th century noted that with few notable exceptions, most fundamental principles of the classical model were available by 1930, and the remainder of the century was devoted to applying or refining these principles.

In 1968, Lord and Novick produced the seminal high watermark of CTT. Using the Spearman notions, they began with the simple notion (perhaps a tautology) that the True score is equal to the Observed score plus error. The claim is that systematic effects between responses of examinees to test items is due only to the variation in the ability (True score) of interest. Thus, the score for a student (X_{jg}, the observed score for examinee j on test g) is the true score (τ_{jg}) plus the error score (E_{jg}). Thus for each student

$$X_{jg} = \tau_{jg} + E_{jg}. \tag{28.1}$$

From such a simple notion was the classical model born, and Lord and Novick provided a detailed and substantive development of the key principles of this model. They demonstrated that it was simple, elegant, and surprisingly durable. The faults of the classical model, however, were well known during its development and extensive use. For example, under the classical model, the estimate of the difficulty of an item in a test can vary depending upon the sample of examinees to which it is administered. It follows that a given individual will appear quite able when compared with one sample of test takers, but much less so when compared with another sample. Thus, an examinee's characteristics and test characteristics cannot be separated—as under the classical model an examinee's score is defined only with respect to a class of (typically hypothetical) parallel tests as opposed to being defined in terms of an underlying latent trait continuum. Furthermore, the classical model averages the various error components and does not apportion error to the various components (each student, each item, etc.). There is a reliance on estimates of reliability defined in terms of parallel forms, there is the lack of an argument as to how a student might perform when confronted with a test item, and there is a lack of methods to detect item bias adequately and defensibly equate

DOI: 10.1037/14049-028
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

test scores. Lord and Novick (1968) realized these concerns, and it is fascinating to note that they also included in their book a chapter by Birnbaum (1968) that provided the fundamental principles to the item response model, which by this time already had a rich history. Because item response theory (IRT) has blossomed and has become the model of choice in most instances of large-scale educational testing, although the classical model dominants in many practice-based domains (e.g., classrooms, clinics, companies; for a more elaborate description of the IRT models, see Volume 1, Chapter 6, this handbook).

The fundamental premise of the IRT model is a specification of a mathematical function relating the probability of an examinee's response on a test item to underlying abilities. As the name implies, IRT attempts to model (in probabilistic terms) the *response of an examinee to an individual test item*. An examinee's ability, θ, and an item's inherent difficulty, *b*, are scaled along the same dimension. To the extent that the examinee's ability "exceeds" the item's difficulty along this dimension, the examinee is said to have an increasingly better than 50–50 chance of getting the item right. If the examinee's ability falls "below" (i.e., to the left) of the item's difficulty, then the examinee is said to have a less than 50–50 chance of getting the item right. When the two parameters coincide, the probability of the examinee getting the item correct is one half. Thus, IRT attempts to model in probabilistic terms the difference between θ and *b*. To scale this difference so that it is a probability function with a range from 0 to 1, it is first necessary to carry $(\theta - b)$ into an exponent. By convention, the base of the exponent is the number *e* (the base of natural logarithms). The resulting exponential expression, $e^{(\theta - b)}$, when divided by the scaling constant, $[1 + e^{(\theta - b)}]$, gives the desired result, the logistic function:

$$p(u = 1|\theta) = e^{(\theta - b)}/[1 + e^{(\theta - b)}]. \qquad (28.2)$$

Equation 28.2 says that the probability of an examinee with ability θ getting an item correct, $p(u = 1)$, follows the familiar logistic function.

The difficulty parameter *b* in IRT is one of three parameters that characterize any given item. Items differ not only in their difficulty but also in their ability to discriminate between examinees who are high on the attribute and those who are low on the attribute. The discrimination parameter *a* indexes the ability of the item to discriminate between examinees of differing ability. For multiple-choice tests of cognitive ability and achievement, even persons very low on the construct being measured have a nonzero chance of getting the item correct by guessing. The so-called "guessing parameter," *c*, indexes the probability that an examinee with very low ability will get the item right by chance alone. The full three-parameter item response model is therefore as follows:

$$p(u = 1 \mid \theta) = c + (1 - c)\frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}.$$

This development of IRT models led to major advances in test score equating, detection of item bias, the refinements of adaptive testing, and the development of multidimensional models. More recently, there have been extensions to polytomous items, attitudes and personality attributes, and modeling misconceptions and cognitive processes. Unlike the classical theory, IRT starts at the item level and adjusts response data for various properties of test items (such as difficulty, discrimination, or propensity for guessing). There are many sources for the development of these models (see Hambleton, Swaminathan, & Rogers, 1991).

In 1997, van der Linden and Hambleton published a handbook of modern IRT. Probably their most important claim was the distinction between "classical item response models" and the newer models based on items with polytomous item formats, models for response item or multiple attempts on items, models for multiple abilities or cognitive components, nonparametric models, models for nonmonotone items, and models with special assumptions about the response process. These modern item response models are now becoming widely used—certainly in most large-scale testing, and they also are beginning to be used in classrooms and other workplaces (thanks to technological advances). More common uses relate to computer-based testing in many forms; the development of reporting

engines; the major changes deriving from computerized essay scoring; and the integration of cognitive science, model-based measurement, and the move to merging these testing advances with instruction. The merging of other statistical advances is also continuing, the most exciting of which is the remarriage of IRT and structural equation modeling (SEM). SEM allows for more control over the patterns and possibly causal paths from one set of variables to others as well as multidimensional models, whereas IRT has emphasized the parameters of the items and the person-fit measures. McDonald (1967, 1999) has long shown that IRT and SEM are closely related and much more can be gained by modeling based on bringing these two powerful, and at times separately developed, notions (and computer programs) together.

The developments noted from classical through item response models could be conceived, until about 10 years ago, as footnotes to the work of the pioneers of these models. This advance has been helped by the ready availability of many sophisticated computer programs. Additionally, the fundamental estimation solutions have become well known, and there are many examples of excellent test development methodology based on both the classical and IRT models. The argument in this chapter is that the coming together of five major domains of research has opened a new and exciting vista for the upcoming future of tests and measurement. These dimensions include (a) advances in technologies, (b) a move from the user as major recipient of tests and test scores to the user as recipient of reports (and less on the scores alone), (c) the realization of a distinct best test design, (d) the differentiation involved based on different forms of interpretation (formative, ascriptive, diagnostic, summative [FADS]), and (e) the measurement of cognitive processing and contexts. Because of the advances of these five directions, some of them rapid over the past decade, newer and exciting methods for testing personality, workplace, and higher order competencies are emerging. Each of these is described in the following sections.

ADVANCES IN TECHNOLOGIES

In most areas of society, technology is playing a central role on changing existing behaviors, functions, and practices. The impact of technology on assessment has seen such innovative practices as paper-and-pencil tests transferred to screen; automated scoring (e.g., of essays); item presentation adaptive to ability levels; the design of items that can be dynamic, adaptive, and interactive; and automatic item generation. In addition to current uses, future developments may see assessments conducted in a virtual reality in which the test taker experiences or participates in a simulation of an event or problem, and they give a "virtual" response from within the simulation. Real-time manipulation of screen entities and objects (e.g., within an interactive science lab) would permit tracking of the steps taken to answer a problem or produce a result. There are two major categories of innovation: those specific to hardware devices and components, and those relating to technological in software applications.

Hardware Devices and Components

Technology is playing an important role in enabling data-driven decision making to occur, often in real-time. For example, web-based reporting systems can allow information relating to state and national curriculum standards to be aligned with the raw data collected from classroom assessments. Observations (e.g., classrooms, workplaces) can be collected via handheld devices (e.g., smartphones and tablets), linked to achievement and processed by sophisticated interpretative software in real time (F. Smith & Hardman, 2003). The use of mobile computing devices provides users with faster and more efficient ways both to administer and analyze data from assessments. Schools, for example, have seen their technology infrastructure become more developed, particularly in relation to student-to-computer ratios. As handheld computer devices now provide computing performance similar to previous generation laptops, they are being seen as a viable alternative to delivering some assessments (Trinder, Mahill, & Roy, 2005). Since their arrival onto the market in 1994, personal digital assistant (PDA) technology (e.g., PocketPCs, iPads, mobile phones) has arguably made the most significant impact because these devices offer distinct advantages over desktop computers and laptops. There is no required boot-up time (switch on and use), making them ideal to

utilize at a moment's notice; their battery lifetimes are longer; they are more portable; and, most important, their price is lower. An example of the interaction of software and hardware technology is demonstrated by Wireless Generation (Resnick & Berger, 2010). Their handheld computer (PDA) is loaded with the specific software platforms (e.g., Dynamic Indicators of Basic Early Literacy Skills [DIBELS]), which guide teachers through the administration of the assessment, allow for assessment information to be recorded simultaneously by the teacher as the student completes each aspect of the assessment tasks, and then provide immediate reporting. Performance information is uploaded and synchronized with preexisting information (e.g., previous performance benchmarks, state learning objectives and standards), thus providing analysis (e.g., response pattern tracking), student- and item- level reporting, and individual- and group-level progress monitoring. Obviously, the development of this type of assessment technology has greatly enhanced the flexibility that teachers as test administrators have when collecting data.

Software Applications

The greatest change in computing over the past 20 years has been related to the Internet and virtual realities (e.g., Second Life). The availability of multimedia and hypermedia technologies permit learner-controlled interactive solutions, where multiple media formats (text, video, audio, still images) can be displayed simultaneously. This accessibility has allowed many more forms of *innovative item* or *innovative test design,* ranging from essentially static passage and text editing (e.g., Parshall, Davey, & Pashley, 2000) to interactive video computerized assessment (Drasgow, Olson-Buchanan, & Moberg, 1999). Many of these methods optimize the advantages of computer scoring, such as inclusion of multimedia, interactivity, simulation, novel ways to record or score answers (e.g., drag and drop, graphing, highlighting, graphing), and the use of alternatives to test scoring (i.e., a, b, c, or short constructed-response answers), including a light pen, a microphone, and a mouse.

There are still too few published studies about tests that have been created using a combination of multimedia features, such as integrated audio, computer graphic technology, and video (e.g., Drasgow et al., 1999). Although most attention has focused on how tests administered across different media differ or not, it may be desirable to ask about the opportunities and extra information that can be derived by capitalizing on the features only available in one media compared with another. For example, Leeson (2008) explored the use of interactive multimedia assessment, with particular reference to showing the additional measurement information that could be gained from this method compared with pencil-and-paper methods. The students engaged in a web-based multimedia (integration of both graphics and video) tool featuring a scenario in which onscreen characters had to complete a challenging reading test. Throughout the scenario, one of the onscreen students turned to the user and asked, "Is this how you feel?" in reaction to a vignette played out in the scene. In an analysis of the test information function, there were only slightly higher levels of information for the students with lower proficiency in reading self-efficacy and very little difference for students with higher proficiency (see Figure 28.1), begging the question of what is added by these newer methods. That the medium is different is not sufficient to make the case for such media providing additional information. More research is needed on whether and how much additional information is provided by these newer methods, and more advice is needed on how to exploit the advantages of these newer technologies to capture additional information.

CHANGING CONCEPTIONS OF VALIDITY AND REPORTING

There has been a remarkable change in conceptions of "validity," which have moved from validity about the test to validity about the interpretations and actions made from test scores (e.g., Shepard, 1993). Messick (1989) has been the most forceful in providing a unified approach to the notion of validity. He commenced by stating,

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales

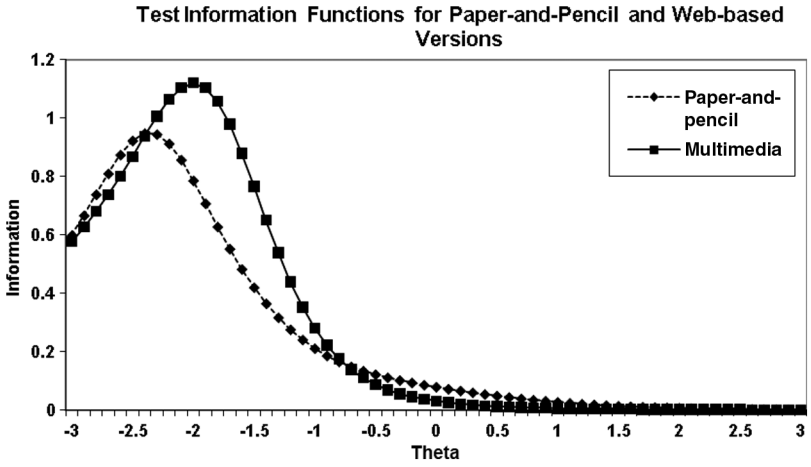


FIGURE 28.1. Average item response theory information across theta levels for paper-and-pencil and web-based versions of the Reader Self-Perception Scale Progress subscale (Leeson, 2008).

> support the "adequacy" and "appropriateness" of "inferences" and "actions" based on test scores or other modes of assessment. As is delineated below, the term test score is used generically here in its broadest sense to mean any observed consistency, not just on tests as ordinarily conceived, but on any means of observing or documenting consistent behaviors or attributes. Broadly speaking, then, validity is an inductive summary of both existing evidence for and the potential consequences of score interpretation and use. Hence what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators—inferences about score meaning or interpretation and about the implications for action that the interpretation entails. (Messick, 1989, p. 13)

A fundamental import of this message is the importance of providing evidence about how users make inferences and take actions, and whether the test developer (and report developer) can provide evidence for the adequacy and appropriateness of these interpretations (Maguire, Hattie, & Haig, 1994). Although most test score reports seem reasonable and at times pretty, there is far less research on optimal methods for developing reports that minimize interpretation accuracy and maximize appropriate precision (for more information on this topic, see Chapter 23, this volume).

Hattie (2010) derived seven major principles in the development of defensible reports based on human–computer interface research, graphics design, and the limited but fascinating work on visual interpretation: (a) readers need a guarantee of safe passage and destination recovery; (b) each report needs to have a major theme (anchored in the task domain, maximizing interpretations and minimizing the use of numbers); (c) reports should minimize scrolling, be uncluttered, and maximize the "seen" over the "read"; (d) reports should provide justification of the test for the specific applied purpose and for the utility of the test in the applied setting; (e) reports should include the meaning and constraints of any report; (f) reports should be timely to the decisions being made; and (g) reports need to be conceived as actions not as screens to print. This topic of optimal report design is still in its infancy, and much more attention is needed on how to devise reports that enable users to correctly interpret them and to make accurate inferences about and take appropriate actions from the reports (see also M. R. Roberts & Gierl, 2010).

A necessary future direction is to provide validity evidence of the *impact* of score reports and to suggest theories of action as to the link between observed behaviors and traits and the reports on

these. Bennett (2010), for example, has challenged test developers to outline and defend their "theory of action"—not only what are the intended and unintended consequences of any testing program but also what are the causal paths between the developers' tests and the outcomes desired—what is the program logic, the theory of action? Bennett outlined some intended outcomes for the school-based assessment he is creating, starting with (a) a clearer, deeper, and better organized understanding on the part of the teachers of the content domain in which they wish to teach; (b) an increase in teachers' pedagogical knowledge and assessment skills; (c) the routine use of formative assessment in the classroom to make appropriate adjustments to instruction; and (d) improved student engagement in learning and assessment. These should then lead, he argued, to improved student learning with respect to content standards and more meaningful information about the effectiveness of schooling. Evidence of this theory of action is that the students who are measured by the tests are then appropriately guided to the next level of instruction, strengths and gaps are found and dealt with, achievement and follow-up claims have similar meaning across population groups, instruction is indeed adjusted by empirical evidence, and the quality of inferences suggested and adjustments made are similar across population groups. Developing such theories of action is a big step for both measurement developers and users, but a reasonable one, especially in light of the exciting directions in assessing cognitive processes.

This stance of providing validity evidence relating to the actions or inferences and also developing theories of action that can provide indicators of validity claims is exciting and could be revolutionary, especially if the cognitive processes and contexts are taken into consideration. Additionally, many related innovations over the past decade can assist in this revolution—for example, those relating to the nature of computer scoring, knowing more about how items provide differential information relating to format or context, developing newer ways of beginning the debates about theories of action based on effective taxonomies, considering different purposes of assessing and reporting, and devising more effective methods of measuring progress.

Open-Ended Scoring

Given the time, resource costs, reliability, and generalizability issues associated with scoring open-ended items, it is no surprise that there have been significant attempts to automate the scoring requirements. Computerized essay scoring has moved a long way from parsing and measuring conditional clauses and other grammatical and syntactic structures (e.g., Page, 1966). Now, content is king. For example, the ETS Criterion (Burstein, Chodorow, & Leacock, 2003) scoring engine uses natural language processing techniques to identify the syntactical and lexical cues in the writing. It allows the teacher to select from more than a hundred writing tasks that span all grade levels, across narrative, expository, and persuasive writing purposes. The engine produces total scores and can provide real-time feedback reporting on mechanics, style, and organization of the writing as well as surface features (grammar, punctuation, and spelling). Vantage Learning has developed IntelliMetric (Elliot, 2003), which uses artificial intelligence, natural language processing, and statistical technologies to "learn" the writing response characteristics that human raters would apply (Shermis & Burstein, 2003). Besides scoring the essays, the software allows students to see their own progress, receive feedback reports, and learn from various diagnostic details. E-rater® (ETS) uses a sample of human-scored essays to ascertain the features and the feature weights and the scoring is thus peculiar to each prompt (and grade level). The major scoring relates to content, word variety, grammar, text complexity, and sentence variety. An encouraging feature of this computer technology is the high agreement rates that have been established between automated scoring systems and subject matter experts (e.g., Shermis & Burstein, 2003; Shermis, Burstein, Higgins, & Zechner, 2010).

The existing developments of essay scoring also can relate to how writing can be assessed and learned. Students can write part or a complete essay, submit the response through these programs, and get reports immediately about their writing proficiencies as well as specifically tailored feedback about their essay, how to improve such essays, branching to instructional modules on the deficiencies in the essay, and prompts about potential improvements.

Short-Constructed Responses

The long and well-known set of criticisms to multiple-choice items often leads many users (especially practitioners) to prefer constructed responses—such as essay or short-answer formats. The typical claim is that with these constructed responses, students are asked to generate rather than recognize correct answers. Computer scoring many of these shorter answers, however, is problematic as the amount of information is usually very small (a few words) and the creativity of users to invent spellings, grammar, and alternate answers seems at times infinite. Many methods have been tried, including assigning a "best-guess-mark," converting words to a phonetic algorithm (e.g., soundex or metaphone, which can reduce the effects of minor spelling mistakes and require less data storage), and utilizing many of the advances from Bayesian classifications (as in spam detection that is predicting the probability that an e-mail is spam, based on the occurrence of certain words in the message, the probability of any message being spam, and the probability of those words appearing in a nonspam message). The advantages are that over time, the "learning" of what is a correct or incorrect answer builds. Much success was achieved when the answers were one or two words, but beyond two words, the probability of success dropped remarkably.

Instead a simple method can be used. Student responses are stored, and teachers are invited to score them as correct or incorrect (at this stage, the item is coded as a red light, indicating that it needs to be scored by the teacher). When 25 teachers score the response correct or incorrect, then an orange light is turned on, and future teachers who see this response can see how previous teachers scored the item and accept the majority interpretation of the response. When 75 teachers score the response, a green light is turned on, which indicates that there can be much confidence in the reliability of the scoring. Over time, most student responses are coded orange or green and most teachers automatically accept their fellow teachers' interpretations. There are nearly always some unique or red light responses, and these beg interpretation (thus adding to the database toward a 25- or 75-teacher concordance). If users are prevented from involvement in these kinds of interpretative decisions (i.e., the green automatically accepted), then they tend to lose confidence in the scoring, but when users are given the choice, they nearly always accept the previous teachers' responses. Thus, scoring open-ended items of a class or even larger group could be conducted in a very short time because only the "less than 25" responses need their attention. Two major bonuses of this method are that the reliability of scoring open-ended items soars and the cost of implementation (no artificial intelligence or probabilistic modeling) is very small.

Differential Item Functioning

One of the sources of variance that can affect the validity of the interpretations relates to differential item functioning (DIF). Zumbo (2007) detailed three generations of DIF. In the first, the concern was item bias, and the research centered around differential performance of groups based on some moderator (e.g., sex, race). The second generation started to distinguish between bias and adverse impact. Bias is an attribute of a specific inference, not of a test, and refers to some distortion in scores that undermines the validity of a particular interpretation. According to Hattie, Jaeger, and Bond (1999),

> For example, in the assessment of mathematical proficiency, if the intent is to assess proficiency in mathematics as "purely" as possible, without confounding the measurement with linguistic ability, it would be important to keep the required level of competence in the language in which the test is written to a minimum. If the examination includes word problems, then the vocabulary and linguistic demands of the problems should be as simple as possible. Otherwise, persons less proficient in the language of the test, such as those for whom the language is a second language or persons who speak specific dialects within the general population, may be unfairly penalized because of purely linguistic, as distinct from mathematical, considerations. (p. 432)

Adverse impact relates to determining the negative effects a group may suffer as a result of test scores.

An assessment is said to exhibit adverse impact with respect to examinee race if, for example, the rate at which African American examinees are certified is substantially below the certification rate of White, non-Hispanic examinees. It is not the differential certification rate that is sufficient; it is the impact of the difference. The presence of adverse impact need not indicate bias and vice versa. There are now many options for detecting bias, such as contingency methods (e.g., Mantel–Haenszel), IRT (relating to variations in item characteristics curves for different groups), and multidimensional models (Shealy & Stout, 1993; for more information on item fairness and bias, see Chapter 27, this volume).

The third generation is "conceiving of DIF as occurring because of some characteristics of the test item *and/or testing situation* that is not relevant to the underlying debility of interest" (Zumbo, 2007, p. 229). This perception then introduces discussion about fairness and equity in testing, dealing with possible threats to internal validity, investigating the comparability of translated or adapted measures, trying to understand the item-response processes, and investigating lack of invariance. There are many defensible methods to detect DIF, and the research needs to move toward understanding the causes or differential processes used in answering items when DIF is present.

Changing Test Specifications

It has been long known that test quality can be linked to the specifications devised to develop the items and component parts. The most well-known taxonomy in education is Bloom's taxonomy, which refers to the type of thinking or processing required in completing tasks or answering questions—that is, know, comprehend, apply, analyze, synthesize, and evaluate (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). The more recent adaption has two dimensions—cognitive processes (i.e., remember, understand, apply, analyze, evaluate, and create) by knowledge types (i.e., factual, conceptual, procedural, and metacognitive; Anderson & Krathwohl, 2001).

Despite its widespread use, this taxonomy has major problems that restrict its effectiveness and value. Reviews of Bloom point to the overdependence on the model despite the lack of research as to its value. Most of the evaluations are philosophical treatises noting, among other criticisms, that there is no evidence for the invariance of these stages or claiming that the taxonomy is not based on any known theory of learning or teaching (Calder, 1983; Furst, 1981). The empirical evidence provides little support for the use of the taxonomy in organizing instruction, curriculum, or assessment (Hattie & Purdie, 1998); the false assumption that these categories represent a hierarchically ordered set; the presupposition of a necessary relationship between the questions asked and the responses to be elicited (see Schrag, 1989); the problem that nearly every item can be made to fit a Bloom level and thus the model does not help improve or exclude items; a tendency to use Bloom for the categorization of items and less for the interpretation of user's responses to these items (Ennis, 1985); and, most seriously, the lack of support for the fundamental claims of the invariance and hierarchical nature of the various levels (see Hattie & Purdie, 1998).

The newer Bloom model introduced cognitive processing distinctions, and these may prove to be sufficient and the most valuable aspect of the taxonomy. It moves closer to the model developed by Biggs and Collis (1982), which they called structure of observed learning outcomes (SOLO). There are four levels termed *unistructural*, *multistructural*, *relational*, and *extended abstract*—which are simply an idea, many ideas, relating ideas, and extending ideas. The first two levels concern surface, and the latter two concern deeper processing (see Figure 28.2 for an example).

Uni-structural.	Who painted *Guernica?*
Multi-structural.	Outline at least two compositional principles that Picasso used in *Guernica.*
Relational.	Relate the theme of *Guernica* to a current event.
Extended Abstract.	What do you consider Picasso was saying via his painting of *Guernica*?

FIGURE 28.2. An item exemplifying the four levels of the structure of observed learning outcomes model.

This taxonomy can lead to improvement of items (as each item should be aimed for one and not multiple levels of the taxonomy); for scoring and interpreting responses, there is higher consistency in allocating items to SOLO than to Bloom (Hattie & Purdie, 1998; Meagher-Lundberg & Brown, 2001; Thomas, Tagg, Holton, & Brown, 2002). It can be used to develop optimal items for testlets (all at one level, all at surface or deep, or an item at each level), it can provide interpretations of the strength and weakness of respondents in the item content as well as processing demands, and it can be used for scoring open-ended items (e.g., essays; see Coogan, Hoben, & Parr 2003; Glasswell, Parr, & Aikman, 2001). SOLO has been used to classify study skills programs (Hattie, Biggs, & Purdie, 1996), identify expert teachers (T. W. Smith, Baker, Hattie, & Bond, 2008), evaluate gifted programs (Maguire, 1988), and evaluate teacher education students (see Hattie & Brown, 2004).

More attention is needed to develop even more effective taxonomies to devise and evaluate items and reporting optimally. Neither Bloom nor SOLO begins to tap beyond the edges into cognitive processing, error analyses, or the skills participants bring to the testing occasion. Without a more defensible taxonomy, the questions of validity regarding constructive alignment (of items to tests to curriculum or attribute specifications) is unlikely to progress. Any taxonomy needs to be evaluated not only in terms of the proportion of items in a test that reflect the appropriate parts of the taxonomy but also to the extent it aids in developing, scoring, crediting partial knowledge, and distinguishing complexity and difficulty. Most important, it can be used for meaningful reporting to teachers and students.

Whatever the test specification, the way tests are constructed has been the topic of major changes over the past 20 years. van der Linden (2005) noted that despite 100 years of developments since classical theories first started being developed, a technology that enables us to engineer tests rigorously to our specifications still has not been developed. Instead, tools from test theories (classical and IRT) are still used to detect attributes of items, and the "primary mode of iteration is not to create good tests, but only to prevent bad tests" (van der Linden, 2005, p. xi). van der Linden (1996; van der Linden & Luecht, 1996) argued that item development (particularly within computer-adaptive testing but also in general) involves designing a test that is optimal in some of its attributes while simultaneously requiring that the test assumes certain prespecified levels of values for a set of other attributes. Formally, this approach involves a test assembly process as an instance of constrained optimization, and thus it involves linear programming methods to choose items that maximize the specification of some objective function while satisfying various constraints placed on the desired test. By specifying various objectives and constraints, it is possible to use the linear programming methods to create a feasible test—that is, a test that meets all the constraints while maximizing the objectives. Objective functions include maximizing test reliability, minimizing number of items in a test, maximizing information at a given cut point, matching a test characteristic function to a target function, minimizing the items with explicit gender or content orientation (for example), and ensuring minimal items on various subdimensions. Examples of constraints include setting the number of items at some value, setting the number of items in the total test on a particular topic to some value, setting the mean item difficulty at some value, ensuring that certain items or types of items must appear in the test, ensuring that certain items not appear simultaneously in the test, and setting the total length of the tests to some number of lines. The technologies to deliver tests using these methods are becoming more readily available. For example, in the development of a large-scale testing tool (Assessment Tools for Teaching and Learning; Hattie, Brown, & Keegan, 2005), "optimal" tests were created using linear programming; in 2005, this took between 5 min and 7 min, and in 2010, it took 5 s to 7 s over the Internet to create optimal tests from more than 5,000 items based on about 20 constraints, mainly chosen by the user. There are also interesting implications in these "assessment engineering" models for the automatic generation of items from quite sophisticated clones (Luecht, 2008). By attending to the reasoning behind the concepts being assessed and identifying

the reusable elements that can be adapted for new items, it is possible to devise cloning of items in many efficient and effective ways (Alves, Gierl, & Lai, 2010; Mislevy & Riconscente, 2006).

Different Purposes of Testing

More than 50 years ago, Scriven (1967, 1990) introduced the notions of formative and summative, which he developed to refer to interpretations in evaluation but quickly transferred to interpretations in testing. A major and widely made mistake has been to consider that these terms referred to assessment methods, but it is the case that a particular assessment can be interpreted in a formative or a summative manner—it is not the test but the interpretations that are formative or summative. As Scriven argued when he introduced these terms, *formative* applies to the nature and timing of the interpretation (during the learning), and *summative* applies to interpretations made about and at the end of the learning.

It does seem that many of the current accountability models in education, for example, tend to prefer tests that can more readily be interpreted from a summative standpoint. To devise systems that can provide more formative interpretations seems contradictory to the needs of many of these accountability movements. Of course, this focus does not stop many from providing teachers with what are termed "formative assessments" (as if formative and summative apply to the tests). Most recently, these have been relabeled "predictive" tests, as their fundamental purpose is to help teachers predict the final summative decision. It may be, however, that some tests used *during* instruction may have low or zero correlation to the desired outcome from a summative interpretation. Consider a student struggling with reading and a teacher using a test to assist in diagnosing the problem. If the diagnosis is correct and then remediation is provided, then it is likely that the student would have a higher chance of being successful in a test at the end of the unit—there may be in this case no correlation between the score on the predictive and end-of-year assessments (i.e., no predictive correlation or even a negative correlation may result). An excellent use of testing, however, may have occurred.

Scriven (2005) more recently introduced a further type of interpretation—*ascriptive*—which relates to using tests retrospectively, generally for documentation, rather than to support any decisions: indeed, a common use. Given the four major purposes of interpretations—formative, ascriptive, diagnostic, and summative—there may need to be greater attention given to "best test design" for each, defensible taxonomies, interpretation more related to the desired purposes, description of the best item construction methods for each purpose, and different measurement models for each type of interpretation. Furthermore, as many would argue, formative decisions are more critical to the learning outcomes of students (if for no other reason that summative interpretations are made at the end of unit of work). If this is the case, then perhaps the tests that lead to formative interpretations should be of higher quality.

A major use of testing in education has been to determine the "levels" of performance, with far less attention placed on methods to measure progress. Such progress testing may also be of use in clinical settings. At best, pre–post testing (such as gain scores) has served as the basis for measuring growth. The literature related to measuring progress with only two points of estimate demonstrates that such interpretations are fraught with difficulties, particularly given that three points is minimal to determine most trend information. It seems surprising that there has been so little use of time series models (although some excellent modeling is available using these methods; Swaminathan, 1984), and dependable measurement over many occasions is needed to implement these methods successfully. Glass (1978) demonstrated that the use of interrupted time series models could be a powerful adjunct to methods that ask about the effects of interventions (e.g., changing teaching methods, counseling interventions), but they are rarely used as testing is still considered an "event" of such magnitude that successive testing seems too daunting and problematic. At the individual level, growth curve analyses can be used to measure progress over multiple points of time—and, in this sense, these models are conditional on time (Raudenbush & Bryk, 2002; Singer & Willet,

2003). As Betebenner (2009) has shown, the rates of growth to reach or maintain proficiency differ by participant based on the current and prior level of attainment for that participant. Thus, interpretations of growth and progression typically require both a normative and criterion basis (see also Briggs & Betenbenner, 2009).

There have been important advances in measuring progress in recent years. One series of advances is the measurement of progress added by a leader (e.g., teacher, counselor, team leader) or by an institution (e.g., school, workplace). There are now many variants of value-added models, which compare the actual level of performance of participants with the level predicted on the basis of their prior attainment or background characteristics. Such statistical modeling, usually via hierarchical linear modeling or growth mixture models, allows therapists, leaders, or institution effects to be expressed as the difference between the actual results and those that are mathematically predicted, taking into account the differences in prior achievement. There are now many forms of value-added modeling: gain score models, regression discontinuity models, covariate adjustment models, layered–complete persistence models, and variable persistence models (Lockwood et al., 2007; Mariano, McCaffrey, & Lockwood, 2010).

ASSESSING COGNITIVE PROCESSES AND CONTEXT

One exciting trend is the assessment of cognitive processes to measure specific knowledge structures and processing skills and to provide information about strengths and weaknesses in learning. This has led to new methods for developing items, reports, and simulations of various contexts, including observations, video, and various technologies. This section outlines these enhancements relating to job performance, clinical psychology, and personality assessment.

Assessment of Cognitive Processes

Embretson (1995) began a major search for new measurement models to model how some individuals process information. Her multicomponent and general latent trait models were devised to measure individual differences in underlying processing components on complex aptitude tasks. She was particularly concerned with the parts of a solution that must be successfully completed for the task to be solved. For example, verbal analogies typically involve two general components: rule construction (how the attributes in the analogy relate to each other) and response evaluation (evaluating the alternatives available—either from recognition or constructed). Leighton and Gierl (2007) have further developed these ideas in their cognitive diagnostic assessment (CDA) models, primarily designed to measure specific knowledge structures and processing skills to provide information about strengths and weaknesses. A major aim is to provide users with more information from test results about how participants here responded to items so that instructional changes can be implemented. CDA is focused on at least three aspects of cognitive processes: skill profiles that represent the most important skills and concepts of the domain, structural procedural networks that provide the building blocks of understanding, and cognitive processes that are invoked in addressing an item. CDA moves the priority of measurement from summative and formative to diagnostic interpretation and places greater emphasis on the appropriateness and granularity of the construct representation, the design and selection of observable indicators, the ability of the items to measure the construct, and the appropriateness of the theoretical measurement foundations related to the specific purposes of diagnostic assessment (Embretson & Yang, 2006).

These models have led researchers, such as Rupp and Mislevy (2007), to more "structured item response" models (Mislevy, Almond, & Lukas, 2003). These models include parameters not only for the latent variables (as in classical IRT models) but also for the unobservable mental operations required to answer the items. That is, they include the various processes (e.g., pattern recognition, comparisons, developing production rules, different levels of integration of knowledge, different degrees of procedural skills, speed of processing) that enable learners to perform constituent cognitive processes

for solving items (see also Junker & Sijtsma, 2001). A most important consequence of these developments is the nature of how to write items. Such a principled approach to test design and reporting requires identification of a valid cognitive processing model, developing items to a template that relates to the feature of such a model, and devising reports to provide feedback relative to the processes as well as the proficiencies of attempting an item. The test specifications require not just content specifications but also processing specifications. Such models could provide much more richness in not only establishing the quality and quantity of answering items correctly but also in the misconceptions and misspecifications students have when addressing these items (see Luecht, 2007).

These processing models can lead to reports providing examinees with specific information about their cognitive strengths and weaknesses and can be used not only for diagnostic testing but also to activate instructional content in something like a "skills tutorial." Such tutorials can have test items linked to the content and curricular standard for an item and can contain instructional videos and links to teaching and learning resources. With the move toward computer-based reporting, tests and reports will become much more instructionally rich and useful (imagine a test written and a report provided on an iPad circa 2010, for instance), become more continuous, and provide students and teachers with immediate diagnosis and scoring, and then it is feasible for dynamic computer-based reporting to affect regular teaching and learning decisions.

Luecht (2009) and Leighton and Gierl (2007) are developing many of these ideas into an "assessment engineering" model. This approach starts with the development of one or more construct maps that describe concrete and ordered performance expectations at various levels of a proposed scale; developing empirically driven evidence models and cognitive task models for the processes and for each level of the construct; developing multiple assessment task templates for each task model to control item difficulty, covariance, and residual errors of measurement; and developing psychometric quality assurance procedures for holding item writers and test developers accountable for adhering to the intended test and task design. The items are not fixed entities but rather are a set of assessment "data objects" that offer a distinct collection of assessment features that can be manipulated to produce a large number of items, each with predictable difficulty and other psychometric characteristics (Luecht, 2009).

These "evidence-centered" or "assessment engineering" procedures promise to be the greatest step forward in assessment since the IRT models were introduced (and many other teams are working on these issues; Ripley, 2007; Wilson, 2005; Wilson et al., 2012). Examples include the Virtual Performance Assessment project, with an emphasis on scientific inquiry skills (http://www.virtualassessment.org/wp); eViva, with an emphasis on describing and annotating milestones and explaining answers; Cascade (from Luxembourg), which measures the confidence between first and final answer; Peerwise (Denny, Hamer, Luxton-Reilly, & Purchase, 2008), which allows students to create items that are then answered and rated by others; and Primum, which is one of the parts of the U.S. Medical Licensing Examination that assesses would-be doctors' capability to treat patients in a practical setting (Dillon, Clyman, Clauser, & Margolis, 2002). The latter provides simulations in which candidates are presented with authentic problems and are asked to treat a simulated patient on screen. They can talk to the patient, receive information, conduct examinations, and order tests and treatments, and the performance is assessed against model responses using a regression-based automatic scoring procedure. These newer models can allow much more interpretative information about processes as well as the formative, ascriptive, diagnostic, and summative interpretations. They are likely to move the attention back to "how" users complete test items as well as providing information on "what" it is they know and can do; this opens opportunities for richer assessment of more complex tasks, higher order modeling of processes, and the measurement of these processes and outcomes as they interact with differing conditions of assessment.

Moving to Newer Demands on Test Outcomes

A major trend in employment and promotion assessment is the desire for more transferable proficiencies,

and there is thus much debate about higher order thinking, collective intelligence, and so on. Consider, as an example, the developments by Schleicher (2009), who is leading the Program for International Student Assessment (PISA), which is an internationally standardized assessment of 15-year-olds carried out every 3 years across 74 countries. The PISA team is aiming to develop assessment relating to three major "deeper" sets of attributes: versatilists, who have excellent thinking skills but not necessarily much content knowledge; personalizers, who have excellent interpersonal skills and can work with others to find ways to resolve problems; and localizers, who have high levels of content information and can build deeper processing within a local domain. The question is thus raised as to how to devise measures for these various types of surface and deep thinking both in the individual and in groups. Rupp, Gushta, Mislevy, and Shaffer (2010) have outlined assessment models of "epistemic games," which model how learners may learn what it is like to think and act like many professionals (journalists, artist, engineers, teachers) and thus facilitate the emergence of disciplinary thinking. These kinds of approaches lead to important debates about what counts as defensible, trustworthy, and convincing evidence and highlights the skills required in the new world of versatilists, personalizers, and localizers.

Assessment of Context

As well as assessment moving to encapsulate proficiencies, progress, processes, and people, there is also a resurgence of interest in assessing the interactions with the context of learning. In comparison to cognitive assessment, investment in measuring contexts and how they interact with individual proficiencies has been negligible. Traditionally, the assessment of contexts has been conducted via the use of standardized observation systems. Many of these systems have been quite limited and at best more directed toward assessing surface aspects of the context and providing information on participant's strengths and weaknesses (Goh & Khine, 2002). Hardman and his colleagues (see F. Smith & Hardman, 2003) have used innovative technology to help teachers understand their students. Their computerized interaction system works via a handheld device about the size of a calculator. It then enables observation and recording of the lesson in real-time and the results are available for immediate analysis. Compared with the older pencil-and-paper systems, it is quicker, more mobile, and more immediate; it is possible to reanalyze classrooms retrospectively from the videos captured; it is highly adaptable (coding can be changed on the run); it is relatively nonintrusive; and the coding is highly reliable. Ackers and Hardman (2001) showed the dominance in most classrooms of the transmission mode of teaching and the three-part exchange that tends to dominate interactions (teacher question, student response, teacher reaction); they have provided evidence of how the U.K. literacy strategy has led to more not less prescriptive teaching, especially when the strategy demands that there be greater student participation and involvement (Hardman, Smith, & Wall, 2003); and they have shown that effective teachers have greater frequency but similar types of interactive discourse to less effective teachers (Smith, Hardman, Mroz, & Wall, 2004). Such assessments of context, in this case of classrooms, move the field beyond the surface features (where the teacher stands, who speaks, and how often) toward richer and more predictive attributes of learning that may be related to the other outcomes of measurement.

Shadel (2010) also saw the assessment of context as a critical source of information about personality functioning, particularly given the importance of appraisals (evaluative judgments about the relation between self and some contextualized environmental events). Cervone (2004) has provided much evidence about self-knowledge as important for regulating coherence in self-efficacy appraisals across a variety of contexts. Any job situation can have situational constraints (e.g., insufficient time, lack of essential assistance from others, inadequate equipment, unforeseeable crises). These need to be appropriately measured and included in determinations about performance. Indeed, Kane (1997) argued that constraints should be assessed on the same aspects of a job, over the same period of time as those on which performance is assessed. Situational constraints on performance may be substantially

greater than many have thought and may explain the majority of variance in performance.

Job Performance and Selection

A recent advance in job performance and selection has been the utilization of information technology. Data mining is now being applied to discover the patterns between the characteristics of personnel and work behaviors, work performance, and retention (Chien & Chen, 2006). This method involves numerous techniques, such as decision trees, algorithms, and neural networks, to discover and explore meaningful patterns and rules in large quantities of data (Beckers & Bsat, 2002). For example, Chien, Wang, and Cheng (2007) used decision tree analysis to assess the latent knowledge required for various positions of employees in a semiconductor foundry company in Taiwan. This approach led to the development of optimum decision rules to establish human resource management strategies and recruitment policies that were then integrated into company policy.

An indication of the applicability of data mining to other areas within psychology is evident in the establishment of the *Journal of Educational Data Mining* in 2009. In the first issue, the editors Baker and Yacef (2009) presented a review of the current status and future trends of data-mining methods and how this approach fits with other interdisciplinary areas, such as psychometrics, statistics, artificial intelligence, concept visualization, and computational and user modeling. Relating specifically to testing and assessment, Madhyastha and Hunt (2009) utilized a method of mining multiple-choice data to examine the similarity of concepts being shown in the item responses. Their analysis produces a similarity matrix in which the visual distance between the concepts tested gives an indication of their relative difficulty based on the underlying reasoning strategies students make when selecting responses.

Rothstein and Goffin (2006) completed a recent review of the use of personality measures in organizations and showed that 30% of U.S. organizations use personality tests to screen applicants, with 40% of Fortune 100 companies using such measures for positions from frontline workers to chief executive officers (cf. Erickson, 2004; Faulder, 2005). The two main motivations for such use are the perception that they contribute to the reduction of employee turnover and that they aim to improve employee fit to the position and organization. They caution that human resource personnel and recruiters often lack the understanding to choose and use the correct measure and make defensible interpretations of the personality measures. Knowing which personality attributes are better predictors is critical when making these interpretations. For example, the various meta-analytical investigations have provided much support for considering conscientiousness and emotional stability when predicting various behaviors, such as integrity, teamwork, customer focus and service, in-role performance, group performance, creativity, and turnover (e.g., Hoel, 2004; Pace & Brannick, 2010; Raja & Johns, 2010; Robert & Cheung, 2010).

Clinical Psychology and Assessment

Recent estimations suggest that an overwhelming majority of clinical psychologists utilize psychological assessment as part of their core clinical practice, although it is estimated that they spend only 0 to 4 hours a week in this activity (Daw, 2001; Groth-Marnat, 2009). However, over the past 70 years, assessment practice in mental health settings has been capricious in nature. Butcher's (2006) review highlighted how the popularity of objective assessment use between the 1930s and 1960s was followed by a distinct reduction in test use given the rise of behavioral therapy in the 1970s. Although the 1990s saw a comeback in the use of and research surrounding psychological tests (Butcher & Rouse, 1996), dramatic changes to health services across many countries saw yet another a decline in the use of such methods in clinical practice (Eisman et al., 1998). Despite this history, Butcher (2006) argued that the growing perceived utility of clinical assessment into the 21st century have been due to five recent developments in this area.

Feedback to clients. First, beyond being an ethical requirement as outlined by the American Psychological Association (2010; see Ethical Standard 9.10), there has been empirical research establishing the therapeutic effects

of clinical assessment feedback to clients (e.g., Allen, Montgomery, Tubman, Frazier, and Escovar, 2003; Finn, 1996, 2003; Finn & Kamphuis, 2006; Finn & Tonsager, 1992; Tharinger et al., 2008). For example, Allen et al. (2003) investigated the effects of client assessment feedback on the rapport built between the client and therapist, and the client's understanding of themselves. Results showed that the experimental group who received personalized feedback on their personality assessment reported significantly higher scores on subsequent measures of self-esteem, self-competence, and self-understanding when compared with the control group that had received only information about the personality measure itself. Allen (2005) found similar outcomes when clients were given feedback on their performance on the Million Clinical Multiaxial Inventory-III. In this study, findings indicated that this feedback positively affected rapport with the therapist and led to self-verification and self-discovery within the adult participants.

Assessment feedback on non-personality-based measures also appears to be beneficial to the client. Krieshok, Ulven, Hecox, and Wettersten (2000) examined the impact of feedback from client's performances on both personality inventories and cognitive functioning measures. They found that such feedback provided a beginning point for the discussion of difficult or sensitive clinical issues. More recent studies have examined whether the benefits found among adult clients is replicable among children and adolescence via more creative approaches. An example of this is the use of personalized fables to provide assessment feedback to younger persons and their family. Although stories and fables have been long used as a technique in psychotherapy, application of this approach has only recently been examined. On the basis of Fischer's (1985/1994) work, Tharinger et al. (2008) presented an illustrative case study examining how to construct individualized fables that are designed to take into account the psychological and emotional status of the child, while also benefiting the informational needs of the family.

Psychometric properties of tests. The second development to which Butcher (2006) has referred is the increased research surrounding the psychometric properties of the tests and assessment procedures widely used in clinical settings (e.g., Hunsley & Mash, 2007; Meyer et al., 2001; Purpura, Wilson, & Lonigan, 2010; Witt & Donnellan, 2008; Witt et al., 2010). In a report initiated by the American Psychological Association's Psychological Assessment Work Group (PAWG), Meyer et al. (2001) provided an extensive meta-analysis of the test validity of clinical psychological testing and assessment. Furthermore, they provided evidence of the assessment efficacy of psychological measures when compared with well-utilized general and neurological medical tests. Results showed that both psychological and medical tests had similar variability in their validity coefficients. For example, the ability to detect dementia via neuropsychological tests ($r = .68$) is comparable with magnetic resonance imaging (MRI) procedures ($r = .57$) assessing the same condition. Although a diagnosis should never be made solely from a test score, there is enough evidence to suggest that many psychological measures provide valid and reliable information and are comparable to medical tests and that these measure can be used to make a legitimate contribution toward the activity of clinical assessment.

Another psychometric review was recently conducted within the subspecialty of clinical pediatric psychology. From survey results, Holmbeck et al. (2008) conducted an investigation of the 37 most commonly used instruments by pediatric psychologists. On the basis of their criteria for evidence-based assessments, 34 of the measures reviewed were classed as demonstrating strong psychometric properties, albeit that most had at least one psychometric weakness (Holmbeck et al., 2008).

Clinical utility of tests. Beyond the psychometric properties of specific measures, Butcher proposed that another development has been the substantial increase in research surrounding the clinical utility of general psychological inventories and strategies. This finding has been particularly apparent in the clinical use of behavioral assessment methods and personality measures constructed to assess the general population (e.g., Beutler & Groth-Marnat, 2003; Butcher & Rouse,

1996; Groth-Marnat, 2009; Thomas & Locke, 2010; Wise, 2010). A related development has been the newly found applicability of specialized clinical tests and assessment strategies to other applied areas of psychology (e.g., forensic, personnel, health) and to nonclinical samples (e.g., Crawford & Henry, 2003; Crawford, Henry, Crombie, & Taylor, 2001; Kessler et al., 2005).

Growth of computer-based testing. The fourth development that Butcher (2006) posited was the ongoing growth of computer-based testing and assessment capabilities. Clinical psychologists are increasingly using computer-based clinical assessments for test administration, collection, and reporting as well as for the interpretation of test results (e.g., the Minnesota Multiphasic Personality Inventory—2; Atlis, Hahn, & Butcher, 2006). In addition, the pervasiveness of the Internet over the past decade has seen many of these services associated with online versions of their instruments, particularly so with personality and neuropsychological assessments (Atlis et al., 2006; Caspar, 2004). Such advances are not without limitations, however. As mentioned earlier in this chapter, online security is an issue at both test developer and respondent levels. At a test level, item exposure has yet to be fully solved, and similarly legitimate respondent information is difficult to validate. Furthermore, if security cannot be guaranteed, increased measurement error may result if respondents are selective about the information they are disclosing online. Given it is only recently that traditional measurement tools have been analyzed for their psychometric properties, an issue with the newly developed web-based versions of these measures will lie in the lack of research examining their equivalence (validity) and efficiency (reliability) as comparable tools. Furthermore, because of this probable lag in such analyses, the clinical psychologist may lack confidence in the applicability of paper-and-pencil-based test norms when aligning online results. This is a point argued by Caspar (2004) who suggested that although a few initial empirically based studies have indicated partially comparable results across modes (e.g., Butcher, Perry, & Hahn, 2004; Percevic, Lambert, & Kordy, 2004), the development of separate norms might be necessary for a test's online version.

Cultural diversity. Like in the other areas of psychology mentioned in this chapter, cultural diversity, the fifth development Butcher (2006) identified, will play an important role in the future use of psychological assessments of mental health. The increasing diversity of populations within countries, coupled with the increasing use internationally of measurement and assessment tools originally developed in the United States, will require a more global perspective of the relevance and adaptations of such approaches. Where much research has investigated the test mode effects of transferring paper-and-pencil item to screen, a *cultural* mode effect might become a required area of focus. Here the equivalence, bias, and efficiency of administering assessment instruments to diverse populations should be empirically tested to establish the psychometric stability of an instrument that has been adapted or revised for a new population. If clinical psychologists are to continue to make clinical assessment a core part of their practice, then it requires test developers and researchers to empirically prove that such tools and procedures are producing what Butcher (2006) referred to as "fair and effective multicultural assessment" (p. 206).

Personality Assessment

Perhaps the greatest advance in personality assessment has been the development and popular use of the Big Five personality model (openness, conscientiousness, extraversion, introversion, neuroticism; Costa & McCrae, 2005). However, more specific subdimensions of these five may be invoked in certain situations. For example, B. W. Roberts, Chernyshenko, Stark, and Goldberg (2005) investigated the underlying hierarchical structure of conscientiousness (industriousness, order, self-control, responsibility) and found that their more differentiated model of conscientiousness was more predictive of behaviors. Such use of more specific dimensions within the broader framework of the Big Five also may lead to a more valid understanding of a person's functioning and personality within the organization when compared with the traditional categorical approach, such as that provided by the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2000; see also Huprich & Bornstein, 2007).

Although IRT has been extensively adopted in different areas of psychology, its application in personality assessment has lagged behind other domains. Over the past 3 decades, considerable research has examined the use of IRT approaches to modeling response data from various personality measures (e.g., Reise, 1999; Reise & Henson, 2000; Rouse, Finger, & Butcher, 1999; Steinberg & Thissen, 1995). Typically, the one- and two-parameter logistic (2PL) models (or graded response model) have been applied to personality scales to find a more appropriate treatment of response data than that provided by the more traditional approach in which values are simply summed across items. Although the application of traditional IRT models has provided a significantly more flexible and robust approach than CTT's handling of personality responses, it cannot be deduced that such models necessarily provide the best *representation* of this data. Instead, recent conjecture has reemphasized Thurstone's (1927, 1928, 1929) earlier argument that the response behaviors exhibited by test takers were different when responding to personality (or attitude) items than when responding to cognitive ability items. Cronbach (1949) and Thurstone proposed that *maximum* behavior occurred when test takers were aware that their performance outcomes were to be measured against prescribed standards. Thus, in many cognitive-ability testing environments, test-taker response behavior is largely constrained to right–wrong responses, often under the pressure of time limits. Conversely, Cronbach referred to the less constrained, but more complex, testing environment of non-cognitive-ability tests as inducing a *typical* performance behavior from test takers. Here, because of the lack of pressure (e.g., typically no or reduced pressure of time limits) and requirement for specific knowledge recall and recognition, test takers exhibited variability in effort and motivation. As a result of these response and behavior distinctions, it has been proposed (Coombs, 1964; J. S. Roberts, Laughlin, & Wedell, 1999) that the resulting response data would be modeled more appropriately using an "ideal point approach" to scoring (e.g., generalized graded unfolding model [GGUM]) rather than the traditional "dominance-based" models (e.g., graded-response model [GRM]).

Thurstone (1928, 1931) posited that an ideal point approach assumes that respondents endorse items based on how closely they believe that the item reflects their own position (e.g., their ideal point). On the basis of the premise of proximity, an "agree" (or a "strongly agree") response is determined by the extent that their own position (e.g., attitude, viewpoint) is reflected by the content of the item. In other words, if an item is positioned below (more negative than the individual's attitude) or above (more positive than the individual's attitude) the individual's position on the trait continuum, then the probability increases that the individual will disagree with the item (J. S. Roberts, Donoghue, & Laughlin, 2000). As such, the *typical* performance modeled under ideal point assumptions produces nonmonotonic response functions that have a single peak and are symmetric about the origin $\left(\theta_j - \delta_i\right) = 0$, where θ_j denotes the location of the *j*th individual on the continuum, and δ_i denotes the position of the *i*th item on the continuum (J. S. Roberts et al., 2000). It is from Thurstone's (1928, 1931) initial ideal point procedures that Coombs (1964) coined the term *unfolding* to represent the process of locating items and respondents' positions on the trait continuum.

Although J. S. Roberts et al. (2000) noted that numerous parametric (Andrich, 1996; Andrich & Luo, 1993; DeSarbo & Hoffman, 1986; Hoijtink, 1990, 1991; Verhelst & Verstralen, 1993) and nonparametric (Cliff, Collins, Zatkin, Gallipeau, & McCormick, 1988; van Schuur, 1984) unfolding models have been devised, the GGUM is the only parametric model for graded responses that allows the discrimination parameter (α_i) of the item to vary and thus permit response category threshold parameters to vary across items (see J. S. Roberts et al., 2000, p. 6). As such, the GGUM leads itself to be compared with other IRT models where varying α_i parameters are a feature.

Unlike the lesser know GGUM, the GRM has become one of the best known and widely applied IRT models for polytomous (e.g., Likert) response models (Hambleton et al., 1991). Samejima (1969) developed the GRM as an extension of the 2PL model to allow for the analysis of items with polytomous ordered response categories, for which

parameters are estimated for *m* ordered response options (see Samejima, 1997, p. 89, Equations 9 and 10). Under this model, analysis identifies the relationships between the item or option parameters, the person parameters, and the particular option that has been selected (Scherbaum, Finlinson, Barden, & Tamanini, 2006). Thus, it is assumed that given the ordered response set under the graded response model, the latent trait value is smaller for test takers who respond "strongly disagree" than it is for test takers who respond "disagree." This assumption highlights the dominance response process that underlies this model (Scherbaum et al., 2006).

Generally, the conjecture has fallen into two camps, one arguing that findings suggest that traditional polytomous IRT models, typically the GRM, are unsuited to the specific nature of non-cognitive-ability scales (e.g., Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; J. S. Roberts et al., 1999, 2000), and the other finding that given the establishment of unidimensionality, the GRM is the preferable option for modeling personality data (e.g., Maydeu-Olivares, 2005).

Chernyshenko et al. (2001) investigated how well a selection of parametric models (2PL, three-parameter logistic [3PL], and GRM) typically used for analyses of personality scales fit the data when compared with the fit provided by Levine's (1984) nonparametric multilinear formula score (MFS) model. Analyses of data from the Sixteen Personality Factor Questionnaire and the Big Five personality scale revealed that overall none of the parametric dominance models applied provided adequate fit. In comparison, the more complex nonparametric MFS model provided fewer misfits across both personality scales. Although Chernyshenko et al. (2001) highlighted some issues that may have affected the poor performance shown by the parametric models, they argued that the results probably reflect the inappropriateness of applying dominance-based models to non-cognitive-ability-based items. Thus, Chernyshenko et al. concluded that the good fit provided by the MFS nonmonotonic function indicates that an ideal point response process may underlie noncogntive responses processes.

Maydeu-Olivares (2005), however, challenged this recommendation, suggesting instead that a distinction between the measurement of *personality* and *attitude* may exist. Although an ideal point approach may be appropriate for attitudinal scales, Maydeu-Olivares suggested that the traditional dominance-based process may be more appropriate for personality-based scales. Specifically, when respondents are required to ascertain the degree to which a personality-based description applies to them, then an individual endorsement of an item will represent their standing at that theta level or higher, as with cognitive-based responses (Maydeu-Olivares, 2005). In addition to using the models and methodology used in the Chernyshenko et al. (2001) study, Maydeu-Olivares also included Bock's (1972) nominal model, Masters's (1982) partial credit model, an extension of Masters's partial credit model (Thissen & Steinberg, 1986), and a normal ogive model version of Samejima's (1969) GRM using the limited information estimation. Results showed that the GRM outperformed all the other models, with the full information version providing the best fit. Although not as successful as the limited and full versions of the GRM, the large number of parameters estimated by the MFS model provided reasonable fit to both scales, which concurs with the findings from the Chernyshenko et al. study.

The good fit provided by the GRM, however, may be a reflection of the scale's construction procedure, rather than a reflection on the appropriateness of dominance IRT models to personality response patterns. Where the Likert procedure to scale construction seeks to avoid statements reflecting either neutral or extreme positions, the ideal point method requires that all locations on the trait continuum be represented. Given that previous applications of an ideal point model has only ever been applied to scales developed based on a dominance approach (e.g., Likert), it seems more appropriate that the effectiveness (or otherwise) of the ideal point approach should be applied to a personality scale constructed given the same theoretical assumptions.

Chernyshenko (2002) and Leeson (2008) provided the only two investigations whereby an ideal

point model has been applied to a personality scale constructed using ideal point assumptions. Using three different scale construction procedures (e.g., dominance CTT, dominance IRT, and ideal point IRT), Chernyshenko constructed three six-facet measures of conscientiousness. Comparing the fit of each conscientiousness scale to its respective model, he found that the ideal point approach showed greater fit to the ideal point constructed scale than did the other two approaches to their respective scales. Across the three scales, items from the ideal point scale provided more information, and hence, greater measurement precision. Leeson (2008) extended Chernyshenko's study by also comparing the modeling capabilities of a dominance IRT model with response data generated from an ideal point constructed scale. Although the scale was developed using the ideal point methodology, the ideal point model (via GGUM) did not provide the expected superior fit to the response patterns. Instead, the graphic results from this study showed that for nearly half of the items, the fit provided by the GGUM estimations showed little if any difference to that provided by the GRM. Although the GGUM did not outperform the GRM in regards to model–data fit, the GGUM estimates did fit the data well. Statistical fit analysis supported the fit plot findings, with chi-square statistics showing comparable magnitudes of fit across all four subscales. Furthermore, chi-square fit analysis revealed that GGUM produced outright superior model–data fit in only a quarter of the item combinations across all scales. The fit performance by the GRM challenges previous conjecture and findings (e.g., Chernyshenko et al., 2001; J. S. Roberts et al., 2000), suggesting that dominance models may not provide best fit for non-cognitive-ability data. A possible explanation may lie at the scale construction level, with the prevalent fit from the GRM evidence of a scale that did not accurately reflect a true ideal point structure. Even given adherence to construction of an ideal point scale, both Chernyshenko (2002) and Leeson (2008) found that the majority of GGUM-estimated items displayed monotonic item response functions.

To date, research has not yet conclusively demonstrated if an ideal point approach to analyzing non-cognitive-ability data is worthy of becoming the more valid approach. If the choice of a model is based on its ability to maximize information and minimize errors, then it has not yet been established that a model such as the GGUM is anymore competent in modeling personality and attitudinal data than the GRM. However intuitively correct the theoretical premises of Thurstone (1929) might appear, clearly a significant amount of empirical evidence is still needed before the ideal approach can be modeled with confidence. An important start, as posited by Leeson (2008), is the examination of the actual ideal point approach to scale construction. The development of an ideal point framework requires that items span the entire range of intervals on the trait dimension (e.g., low, medium, and high). Given this structure, items are not reversed scored, and an ideal point model trait estimate (or trait score) is derived by computing the mean item location of the items endorsed (Drasgow, Chernyshenko, & Stark, 2010b). By comparison, the dominance framework requires that items reflect positive and negative aspects of the trait, with any neutral items removed. For such a Likert scale, negative items are reversed scored, and trait scores are derived from three methods: a computation of the proportion of items endorsed, the sum of the item scores, or the trait estimate generated by the logistical IRT model used (Drasgow et al., 2010b). Brown and Maydeu-Olivares (2010), however, have argued that fair and construct valid intermediate items required by the ideal point framework are challenging to write. Thus, it is much easier to write positive or negative items than the more ambiguous middle or average reflections of a trait. Leeson (2008) noted that although 30% of the developed items in her scale were judged by subject matter experts as being good representations of average academic self-worth, GGUM parameters showed that almost all of the item parameter estimates indicated extreme locations on the trait continuum.

Future work needs to determine whether the difficulty of developing a "true" ideal point scale is in part responsible for the ideal point model's (e.g., GGUM) indifferent performance when compared with dominance models (i.e., the graded response model; Samejima, 1997) as found by recent studies (e.g., Maydeu-Olivares, 2005; Leeson, 2008).

It is proposed here that before investigating some of the other issues that have been posited in the dominance versus ideal point debate (e.g., see Brown & Maydeu-Olivares, 2010; Credé, 2010; Dalal, Withrow, Gibby, & Zickar, 2010; Drasgow et al., 2010a, 2010b; Oswald & Schell, 2010; Reise, 2010; Waples, Weyhrauch, Connell, & Culbertson, 2010) the establishment of clear, valid, and reliable scale construction procedures to produce items representing an ideal point scale is required to determine whether there is indeed a difference in responses to cognitive and noncognitive items.

GLOBALIZATION OF MEASUREMENT IN PSYCHOLOGY

That measurement now crosses national boundaries is quite evident. Nearly all assessment in job selection now is done via the Internet; the move to the cloud has allowed even more web-based assessment, because the constraints of connection speed and pipe size no longer matter as much. There is now more awareness of cultural, language, and country translations, and the increase in Internet testing has led to concerns about security and more opportunities for global cooperation in assessment.

The Growth of Global Testing

When the above trends in testing are combined, the opportunities are quite exciting for the status and future of assessment. Testing has begun to cross borders at a rapid speed, and this internalization has been most noticeable in the global growth of national assessment and international educational testing (Kamens & McNeely, 2010). Indeed, by the end of the 20th century, participation among both developing and industrial countries had increased dramatically, with Kamens and McNeely (2007) predicting that more than a third of all countries would be assessing middle and high school students using standardized tests within the next decade. Benavot and Tanner (2007) claimed that as of 2006, 81% of industrial countries had national assessment systems for their schools. Indeed, technological advances have made large-scale assessment and testing programs logistically possible, coupled with the ongoing development and availability of sophisticated testing models. Kamens and McNeely (2009), however, have argued that technological improvements cannot be the only explanation of this dramatic and continuing growth. A strong motivator is the perceived benefit that nations may gain from an educated population that has the skills to compete, invent, and innovate on the world's stage.

The development in schools, for example, of Progress in International Reading Literacy Study, Trends in International Mathematics and Science Study, PISA, and Adult Literacy and Lifeskills have led to not only educational outcomes rank listed of by country but also many advances in how to optimally test across nations—such as controlling for the use of "inappropriate" words and concepts across countries, translation issues, and comparative validity. The sociocultural issues become all the more important (Padilla & Borsato, 2008; Suzuki & Ponterotto, 2008). Changing a test from one language and culture to another is thus not merely translating the words and concepts but also dealing with the cultural appropriateness of items. Although there have been decades of research addressing the validity issues related to the administration of tests in one language to another language (Padilla & Borsato, 2008), more recent developments in the validity and standards for test adaptation have moved to consider the sociocultural dimensions of assessment as well (see http://www.intestcom.org/guidelines/index.php). The International Test Commission has been the single most important leader worldwide in this process (for more information on test adaptation, see Chapter 26, this volume).

Padilla and Borsato (2008) went further and claimed that specific cultural variables should be examined when making assessment decisions, such as the ways individuals having different cultures come into continuous firsthand contact with each other (Aşçı, Fletcher, & Çağlar, 2009). For example, in many countries, a major concern is the applicability of testing among many cultures within that country (and indeed the interactions can then change the "culture" of that country; Cabassa, 2003; van de Vijver & Phalet, 2004). Following an extensive review of trends in industrial and organizational

psychology research over the past 4 decades, Cascio and Aguinis (2008) have argued that significant innovative research needs to be focused on understanding and measuring global or cultural intelligence. This step would not be a reversion to intelligence tests (see Fletcher & Hattie, 2011), as these measures are often seen as based on Western intelligence theories and measures and it often is assumed that a universal set of abilities or intelligences can be adequately translated, administered, and interpreted globally. Given the ease of travel, and global communication systems, the variance between countries is likely to be reduced while at the same time the variance within countries will become larger.

The issues outlined here in relation to the global measuring of aptitude and cognitive skill also are germane in personality, behavioral, and neuropsychological assessment although far less research has been focused in these areas. Reynolds (2000) found that although neuropsychologists in the United States were being presented with a greater proportion of patients from vastly different ethnicities and cultures, there is a paucity of cross-culturally valid or culturally valid test instruments. As P. Smith, Lane, and Llorente (2008) argued, it is essential that the degree of bias and the effects of multiculturalism existing in neuropsychological assessments be adequately understood if effective diagnostic evaluations are to occur. Furthermore, even before a neuropsychological assessment begins, it is important to prepare the patients from culturally diverse backgrounds for the event. This practice helps to address the expectations, and potential confusion and anxieties, regarding the assessment procedure and can assist in maximizing the quality of consequential interpretations of the assessments by the patient (Uomoto & Wong, 2000).

The growth of access to the Internet has primarily forced testing to become more global. Many traditional psychological tests are now available on the Internet and the greatest growth in this regard has been certification, licensing, and job selection instruments (Bartram, 2008). In 2003, Naglieri et al. (2004) found a million results referencing psychological testing, and in a Google search more than 865,000 results with the phrase *psychological tests* and 900,000 with *educational tests* were found—with many providing free tests (often with free or paid scoring and reporting). These tests covered the gamut of domains from neuropsychological, industrial and organizational, educational, personality and psychodiagnostic, and clinical and counseling testing. It is likely that many of these tests would not follow professional guidelines (e.g., the Joint Technical Standards; for more information about these standards, see Volume 1, Chapter 13, this handbook) in their development and interpretations in a way that is responsible, helpful, and unlikely to cause harm. Naglieri et al. asked that Internet-based tests be subjected to the same defensible standards for assessment tools as paper-and-pencil tests when their results are used to make important decisions, that scores on Internet-based tests (and via report) be supported by evidence of validity, and the qualities of the administrator and interpreter of the tests meet the same high, rigorous standards as laid down by test standard and ethics proposals. These requirements seem too much too late because none of these criteria are likely to be realized now that the genie is out of the bottle. Many tests and reports are already on the Internet—unproctored, with little or no evidence and a comparable degree of interest in psychometric details, with uncorroborated reports and interpretations—and are widely used by many. This scenario certainly requires a different approach to research about Internet testing and its usage, and about regulating the developers, users, and interpreters of testing.

A concern related to the increase in global testing is the nature of a normative basis for interpretations. In addition to the issues of test translation and adaption, there are research questions about the applicability of local norms that often form the basis for interpretations and comparisons. van de Vijver (2008) has identified four major issues: the cultural and linguistic aspects of the test; the necessity to demonstrate the appropriateness of local norms to each candidate; an attention to both internal (any factor that could challenge the comparability of tests scores) and external (the identity of the predictor) bias; and the necessity to incorporate the literature on diversity management, intercultural competency,

and acculturation that can play an important part in the predictive power of tests. The days of global norming are near.

Security in the Global World of Testing

Given the increase of tests on the Internet, the issues of test security have also become greater. Tests available via the Internet may be inappropriately answered by some users; the tests may be modifiable without acknowledgment, copyright, or correctness established; and tests may be administered unproctored. Cizek (2006), among many others, has noted the enormous rise of cheating, or maybe it is the rise of improved methods for the *discovery* of cheating (that may have always been prevalent). He cited research showing that 64% of students admitted to cheating at least once in the past year, up from 60% in a similar study in 2004.

Foster and Miller (2010) outlined many types of cheating: preexposure to test content, using a proxy to take a test (e.g., hire an expert), receiving help from another person during the test, using inappropriate aids during testing, hacking into scoring databases to raise test scores, copying from another test taker during a test, capturing test question files, using a camera or cell phone to take pictures of items, video recording an Internet test session, and retaking a test inappropriately. Some recent software can assist to mitigate these concerns, or at least identify potential abusers of tests. For example, Foster (2009) outlined a program whereby users type a 15-character phrase 12 to 15 times, and the patterning, keystroke timing, and pacing is used to then compare with test-taking behavior during the test. Also, webcams can be used to note or check whether there is suspicion of deviant behavior, stopping the use of some keys to prevent accessing other programs on a computer, or running test pattern analyses. Segall (2001) proposed open Internet testing (in his case for the U.S. Armed Services) and those passing this first selection test are then invited to visit an assessment center for first a confirmation tests and, if necessary, further testing (plus using matching statistical procedures between the Internet and confirmation tests).

Global Cooperation in Assessment

The changes toward more global solutions in reporting and availability of assessment are likely to lead to major changes in how tests are produced, interpreted, and commercially made available. It is possible to consider an Internet-based assessment model, in the same sense as Google Earth, Facebook, and so on. Imagine scrolling to "Internet Testing." The user (teacher, student, psychologist) could create a test controlling the length, nature of items, and mode of delivery (paper, on screen, computer-adaptive testing; via iPad, mobile phone, etc.). This approach would require an excellent evidenced-based or assessment engineering model underlying the development of items and tests as well as an item–data management system that would not only be a repository of items and their signatures—and for links of items to curricula—but also create a process to allow users to build their own items, learn about the attributes and success of their items, and perhaps allow for successful items to be added to the bank. A "rendering" tool could take the optimal set of items that meet these specifications (e.g., via linear programming) to create a test. Upon completion, the scoring and reporting engine, which would need to be at the heart of the application, could then provide reports that maximize the adequacy and appropriateness of interpretations for the user (e.g., teacher, student).

Such a model has major implications for commercial selling of tests and reports, for reducing the burden of the current major expenses (developing new items, reports), and for competencies required to interpret the reports appropriately and adequately. Such a model could also improve the assessment capabilities of students and teachers and may be one of the more successful mechanisms for improving the quality of teaching and learning, particularly in developing countries where quality tests linked to local curricula are often sparse. Testing would move from its current emphasis on the quality of items and tests to the quality of the reports and the constructive alignment of items, curricula, diagnoses, job requirements, and reporting. All of the components still need to be of high value, but the experience of testing is moving from a script administered by a trained or registered administrator to a test taker, to an opportunity for users to create tests for specific and local purposes more directly and for users to derive appropriate and dependable reports

about a trait or attribute. The person making the interpretations becomes more important than the test, items, or developer. The consequences become more important as the interpretations are linked to "what next?" decisions and diagnoses.

CONCLUSION

Ten years ago, a paper on the persistent issues in measurement theory concluded that many aspects of measurement theory have not kept pace with the advances in psychology and how people process ideas (Hattie, Jaeger, & Bond, 1999). Hattie et al. (1999) saw much promise in the improvements in modern technology to advance such measurement modeling and to provide more effective ways to communicate the complex information from better designed tests. The underlying assumptions of the major measurement models (classical and item response) seemed to be grounded in Bloom's first two levels of knowledge and comprehension but were less capable of modeling the higher order processing proficiencies. The promise by van der Linden and Hambleton (1997) when they predicted that IRT may adopt a new identity "as an enthusiastic provider of psychological models for cognitive processes" as it may "help to better integrate measurement and substantive research—areas that to date have lived too apart from each other" (p. 22) was shared in this chapter. There is hope in the development of performance assessments; however, largely due to the pressures for accountability, more computer-based, quick scored, summative tests have been utilized and performance assessments have waned. The message of this earlier review was more an evolution rather than a revolution—more use of current methods with little new or novel.

Over the past 10 years, there have been major advances of technology (Internet, m-technologies), reduced costs of this technology, important developments in reporting and validity, progress in modeling cognitive processing, more recognition of the importance of contexts, increased sharing of testing methods, and advances in a global context. These developments are exciting, promise much, and indeed could be considered revolutionary. What seems needed at present is a new psychometric model that brings all of these developments together and allows for better modeling of the component and interactive parts. Such a model would need to look closely at many current assumptions. The current IRT and CTT models are mainly concerned with dichotomous or polytomous items, whereas any new model needs to be concerned with the processing that leads to the responses and needs to consider the relation between this processing and the context of learning and development. The current IRT and CTT models are based on assumptions about ability, whereas any new model needs also to be concerned about change and not assume invariance over time. Furthermore, the current IRT and CTT models rarely take into consideration the opportunity to learn, assuming that proficiency information for all students is based on an equal opportunity—whereas in many situations, this assumption is not realistic. Hence the scores currently include a mix of ability, achievement, processing strategies, and opportunity to learn. Any new model would need to consider these separately as well as how they relate together to form a statement about proficiency. The newer measurement models need to include much more about levels of proficiency *and* progress over time especially relating to the cognitive processing that optimally cause changes over time. Finally, the new measurement models need to be sensitive to the differing "best test" design, item attributes, and reporting relating to the various purposes of assessment (FADS).

To develop a new model, especially including modeling of cognitive processing, some of the fundamental assumptions that have lead to many of the current models should be questioned. For example, a key assumption of many IRT models is unidimensionality, although the newer multidimensional IRT models can assist in dealing with this assumption to some extent (Frey & Seitz, 2009; Hattie, 1984; see also Volume 1, Chapter 6, this handbook). Traub (1983) raised the issues as to whether differences in instruction can create multidimensionality where before there had been unidimensionality. His ineluctable conclusion was that no unidimensional item response model is likely to fit educational achievement data. It may be that there are some people who do not fit the models, and there have been important advances in

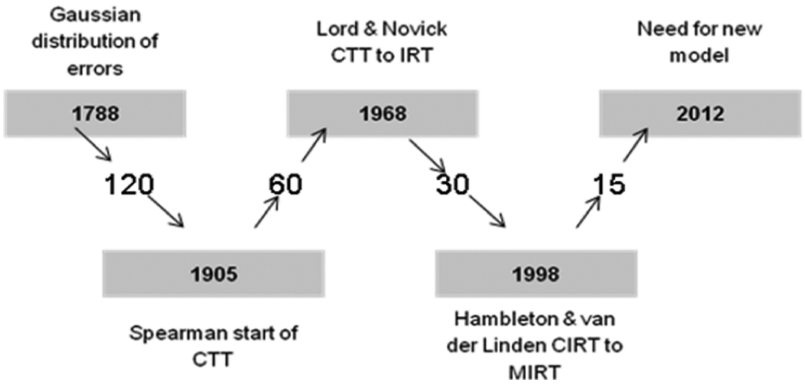


FIGURE 28.3. A timeline of major developments in measurement.

developing techniques to detect item-score patterns that are improbable given the IRT model or given the other patterns in the data (Emons, Sijtsma & Meijer, 2005; Meijer & Sijtsma, 2001).

Possibly the greatest potential for change in education testing from the innovations in technology and the Internet is that tests can be built relating to the learning process rather than to modify learning processes that best relate to current tests. So much current learning involves teachers talking, students reading and writing, and closed-form assessment methods. The newer technologies allow for more doing, exploring, making choices, and unpredictability. The gaming industry, for example, may have much to teach the assessment industry. Developing games starts by asking about the outcomes and goals and then asking how to get there particularly by invoking appropriate challenges to keep the user on task, wanting to invest more effort, and providing immediate feedback. Games are built around problem solving; place a premium on user proficiencies to create, innovate, and produce; can involve others in competitive or collaborate actions to solve problems and create "personal bests"; collect information on users as they develop through the game; track information across time; integrate learning and assessment; provide feedback during the task so that the user becomes more efficient and effective; and are equitable in that they do not favor slow or fast learners and ignore background variables such as the home, socioeconomic status, or race (for an elaboration of each of these ideas, see Shaffer & Gee, 2008). There are many directions for more effective testing that engages learners and emphasizes testing students' strategies. Furthermore, so much learning can accrue from this involvement in testing.

This chapter began with a brief acknowledgement to the history of test theory. There were 120 years between Gauss's contributions to the distribution of errors and Spearman's start ofCTT, 60 years then transpired before Lord and Novick's contributions and the beginning of the dramatic rise of IRT models, 30 years later Hambleton and van der Linden's compilation of the many new IRT models occurred, and so maybe in 15 years there may be a new model that brings together knowledge of testing surface level attributes (CTT and IRT) with the more deeper or higher order processing (Figure 28.3). This "15 years" culminates in 2012, and whether then or soon after, the current research indicates that a new model is not too far away. The early decades of this century are an exciting time for measurement. Hambleton (personal communication, 2010), summed up this excitement:

> Indeed, the greatest challenge is the blurring of any distinction between summative and formative interpretations in large-scale and classroom testing along with the proliferation of research in these two areas stemming from the influx of ideas from cognitive science, instructional technology, the learning sciences, mathematical statistics, computing science, and educational psychology. In other words, educational assessment is truly becoming an interdisciplinary field (in the past, it was largely test development and statistical analysis).

References

Ackers, J., & Hardman, E. (2001). Classroom interaction in Kenyan primary schools. *Compare, 31*, 245–261. doi:10.1080/03057920120053238

Alves, C. B., Gierl, M. J., & Lai, H. (2010, May). *Using automated item generation to promote principled test design and development.* Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2002, Amended June 1, 2010)*. Retrieved from http://www.apa.org/ethics/code/index.aspx

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York, NY: Longman.

Andrich, D. (1996). A general hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49*, 347–365. doi:10.1111/j.2044-8317.1996.tb01093.x

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17*, 253–276. doi:10.1177/014662169301700307

Aşçı, F. H., Fletcher, R. B., & Çağlar, E. D. (2009). A differential item functioning analysis of the PSDQ with Turkish and New Zealand/Australian adolescents. *Psychology of Sport and Exercise, 10*, 12–18. doi:10.1016/j.psychsport.2008.05.001

Atlis, M. M., Hahn, J., & Butcher, J. N. (2006). Computer-based assessment with the MMPI–2. In J. N. Butcher (Ed.), *MMPI–2: The practitioner's handbook* (pp. 445–476). Washington, DC: American Psychological Association. doi:10.1037/11287-016

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*, 3–17.

Bartram, D. (2008). Global norms: Towards some guidelines for aggregating personality norms across countries. *International Journal of Testing, 8*, 315–333. doi:10.1080/15305050802435037

Beckers, A. M., & Bsat, M. Z. (2002). A DSS classification model for research in human resource information systems. *Information Systems Management, 19*, 1–10. doi:10.1201/1078/43201.19.3.20020601/37169.6

Benavot, A., & Tanner, E. (2007). *The growth of national learning assessments in the world, 1995–2006.* Background paper for the Education for All Global Monitoring Report 2008: Education for all by 2015: Will we make it? Paris, France: UNESCO.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, *8*, 70–91. doi:10.1080/15366367.2010.508686

Betebenner, D. W. (2009). *Growth, standards and accountability*. Center for Educational Assessment. Retrieved from http://www.nciea.org/cgi-bin/pubspage.cgi

Beutler, L. E., & Groth-Marnat, G. (2003). *Integrative assessment of adult personality*. New York, NY: Guilford Press.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. New York, NY: Academic Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The cognitive domain.* New York, NY: McKay.

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51. doi:10.1007/BF02291411

Briggs, D., & Betenbenner, D. W. (2009, April). *Is growth in student achievement scale dependent?* Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Brown, A., & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology, 3*, 489–493. doi:10.1111/j.1754-9434.2010.01277.x

Burstein, J., Chodorow, M., & Leacock, C. (2003). *Criterion: Online essay evaluation: An application for automated evaluation of test-taker essays.* Paper presented at the 15th annual conference of innovative applications of artificial intelligence. Acapulco, Mexico. Retrieved from http://www.ets.org/Criterion

Butcher, J. N. (2006). Assessment in clinical psychology: A perspective on the past, present challenges, and future prospects. *Clinical Psychology: Science and Practice, 13*, 205–209. doi:10.1111/j.1468-2850.2006.00025.x

Butcher, J. N. (2010). Computer-based test interpretation. In I. B. Weiner & W. E. Craighead (Eds.),

Corsini encyclopedia of psychology (pp. 1–3). Hoboken, NJ: Wiley.

Butcher, J. N., Perry, J., & Hahn, J. (2004). Computers in clinical assessment: Historical developments, present status, and future challenges. *Journal of Clinical Psychology, 60*, 331–345. doi:10.1002/jclp.10267

Butcher, J. N., & Rouse, S. V. (1996). Personality: Individual differences and clinical assessment. *Annual Review of Psychology, 47*, 87–111. doi:10.1146/annurev.psych.47.1.87

Cabassa, L. J. (2003). Measuring acculturation: Where we are and where we need to go. *Hispanic Journal of Behavioral Sciences,* 25, 127–146.

Calder, J R. (1983). In the cells of Bloom's taxonomy. *Journal of Curriculum Studies, 15*, 291–302.

Cascio, W. F., & Aguinis, H. (2008). Staffing twenty-first century organizations. *Academy of Management Annals, 2*, 133–165.

Caspar, F. (2004). Technological developments and applications in clinical psychology and psychotherapy: Introduction. *Journal of Clinical Psychology, 60*, 221–238. doi:10.1002/jclp.10260

Cervone, D. (2004). The architecture of personality. *Psychological Review, 111*, 183–204. doi:10.1037/0033-295X.111.1.183

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523–562. doi:10.1207/S15327906MBR3604_03

Chien, C. F., & Chen, L. (2006). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications, 34*, 280–290. doi:10.1016/j.eswa.2006.09.003

Chien, C.-F., Wang, W.-C., & Cheng, J.-C. (2007). Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Systems with Applications, 33*, 192–198. doi:10.1016/j.eswa.2006.04.014

Cizek, G. J. (2006). Detecting and dealing with academic dishonesty. In W. Buskist & S. F. Davis (Eds.), *Handbook of the teaching of psychology* (pp. 238–243). Boston, MA: Blackwell. doi:10.1002/9780470754924.ch41

Cliff, N., Collins, L. M., Zatkin, J., Gallipeau, D., & McCormick, D. J. (1988). An ordinal scaling method for questionnaire and other ordinal data. *Applied Psychological Measurement, 12*, 83–97. doi:10.1177/014662168801200108

Coogan, P., Hoben, N., & Parr, J. M. (2003). *Written language curriculum framework and map: Levels* 5–6 (Tech. Rep. No. 37). Auckland, New Zealand: University of Auckland/Ministry of Education.

Coombs, C. H. (1964). *A theory of data.* New York, NY: Wiley.

Costa, P. T., Jr., & McCrae, R. R. (2005). A five-factor model perspective on personality disorders. In Strack, S. (Ed.), *Handbook of personology and psychopathology* (pp. 257–270). Hoboken, NJ: Wiley.

Crawford, J. R., & Henry, J. D. (2003). The Depression Anxiety Stress Scales (DASS): Normative data and latent structure in a large non-clinical sample. *British Journal of Clinical Psychology, 42*, 111–131. doi:10.1348/014466503321903544

Crawford, J. R., Henry, J. D., Crombie, C., & Taylor, E. P. (2001). Normative data for the HADS from a large non-clinical sample. *British Journal of Clinical Psychology, 40*, 429–434. doi:10.1348/014466501163904

Credé, M. (2010). Two caveats for the use of ideal point items: Discrepancies and bivariate constructs. *Industrial and Organizational Psychology, 3*, 494–497. doi:10.1111/j.1754-9434.2010.01278.x

Cronbach, L. J. (1949). *Essentials of psychological testing.* New York, NY: Harper & Row.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Dalal, D. K., Withrow, S., Gibby, R. E., & Zickar, M. J. (2010). Six questions that practitioners (might) have about ideal point response process items. *Industrial and Organizational Psychology, 3*, 498–501. doi:10.1111/j.1754-9434.2010.01279.x

Daw, J. (2001). Psychological assessments shown to be as valid as medical tests. *Monitor on Psychology, 32*, 46–47.

Denny, P., Hamer, J., Luxton-Reilly, A., & Purchase, H. (2008, September). *PeerWise: Students sharing their multiple choice questions*. In Proceedings of the Fourth International Workshop on Computing Education Research, ACM, New York, NY.

DeSarbo, W. S., & Hoffman, D. L. (1986). Simple and weighted unfolding threshold models for the spatial representation of binary choice data. *Applied Psychological Measurement, 10*, 247–264. doi:10.1177/014662168601000304

Dillon, G. F., Clyman, S. G., Clauser, B. E., & Margolis, M. J. (2002). The introduction of computer-based case simulations into the United States Medical Licensing examination. *Academic Medicine,* 77, S94–S96. doi:10.1097/00001888-200210001-00029

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010a). Improving the measurement of psychological variables: Ideal point models rock! *Industrial and Organizational Psychology, 3*, 515–520. doi:10.1111/j.1754-9434.2010.01284.x

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010b). 75 years after Likert: Thurstone was right! *Industrial*

and Organizational Psychology, 3, 465–476. doi:10.1111/j.1754-9434.2010.01273.x

Drasgow, F., Olson-Buchanan, J. B., & Moberg, P. J. (1999). Development of an interactive video assessment: Trials and tribulations. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 177–196). Mahwah, NJ: Erlbaum.

Eisman, E., Dies, R., Finn, S. E., Eyde, L., Kay, G. G., Kubiszyn, T., . . . Moreland, K. (1998). *Problems and limitations in the use of psychological assessment in contemporary healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group, Part II*. Washington, DC: American Psychological Association.

Elliot, S. (2003). *How does IntelliMetric score essay responses?* Newtown, PA: Vantage Learning.

Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: Application and mathematical reasoning. *Journal of Educational Measurement, 32*, 277–294. doi:10.1111/j.1745-3984.1995.tb00467.x

Embretson, S. E. (2004). The second century of ability testing: Some predictions and speculations. *Measurement: Interdisciplinary Research and Perspectives, 2*, 1–32.

Embretson, S. E., & Yang, X. (2006). Multicomponent latent trait models for complex tasks. *Journal of Applied Measurement, 7*, 335–350.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods*, *10*, 101–119. doi:10.1037/1082-989X.10.1.101

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*, 44–48.

Erickson, P. B. (2004, May 16). Employer hiring tests grow sophisticated in quest for insight about applicants. *Knight Ridder Tribune Business News*, p. 1.

Faulder, L. (2005, January 9). The growing cult of personality tests. *Edmonton Journal*, p. D6.

Finn, S. E. (1996). *Manual for using the MMPI–2 as a therapeutic intervention*. Minneapolis: University of Minnesota Press.

Finn, S. E. (2003). Therapeutic assessment of a man with "ADD." *Journal of Personality Assessment, 80*, 115–129. doi:10.1207/S15327752JPA8002_01

Finn, S. E., & Kamphuis, J. H. (2006). The MMPI–2 Reconstructed Clinical (RC) scales and restraints to innovation, or "What have they done to my song?" *Journal of Personality Assessment, 87*, 202–210. doi:10.1207/s15327752jpa8702_10

Finn, S. E., & Tonsager, M. E. (1992). Therapeutic effects of providing MMPI–2 feedback to college students awaiting therapy. *Psychological Assessment, 4*, 278–287. doi:10.1037/1040-3590.4.3.278

Fischer, C. T. (1985/1994). *Individualizing psychological assessment*. Mahwah, NJ: Erlbaum.

Fletcher, R. B., & Hattie, J. A. C. (2011). *Intelligence and intelligence testing*. Oxford, England: Routledge.

Foster, D. (2009). Secure, online, high-stakes testing: Science fiction or business reality? *Industrial and Organizational Psychology, 2*, 31–34. doi:10.1111/j.1754-9434.2008.01103.x

Foster, D., & Miller, J. L. (2010). *Global test security issues and ethical challenges*. Paper presented at the International Test Commission Conference, Hong Kong.

Frey, A., & Seitz, N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, *35*, 89–94. doi:10.1016/j.stueduc.2009.10.007

Furst, E. J. (1981). Bloom's taxonomy of educational objectives for the cognitive domain: Philosophical and educational issues. *Review of Educational Research, 51*, 441–453.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237–261. doi:10.1111/j.1745-3984.1978.tb00072.x

Glasswell, K., Parr, J., & Aikman, M. (2001). *Development of the asTTle writing assessment rubrics for scoring extended writing tasks*. (Tech. Rep. No. 6). Auckland, New Zealand: University of Auckland, Project asTTle.

Goh, S. C., & Khine, M. S. (Eds.). (2002). *Studies in educational learning environments: An international perspective*. Singapore: World Scientific. doi:10.1142/9789812777133

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: Wiley.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hardman, F., Smith, F., & Wall, K. (2003). *An investigation into the impact of the national literacy strategy on the literacy learning of pupils with special educational needs in literacy in mainstream primary schools*. University of Newcastle upon Tyne: A report for the Nuffield Foundation.

Hattie, J. A. (1984). Decision criteria for assessing unidimensionality: An empirical study. *Multivariate Behavioral Research, 19*, 49–78. doi:10.1207/s15327906mbr1901_3

Hattie, J. A., Biggs, J., & Purdie, N. (1996). Effects of learning skills intervention on student learning: A meta-analysis. *Review of Educational Research, 66*, 99–136.

Hattie, J. A., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education, 24*, 393–446.

Hattie, J. A., & Purdie, N. (1998). The SOLO model: Addressing fundamental measurement issues. In B. Dart & G. Boulton-Lewis (Eds.), *Teaching and learning in higher education* (pp. 145–176). Melbourne, Australia: Australian Council for Educational Research.

Hattie, J. A. C. (2010). The validity of reports. *Online Educational Research Journal*. Retrieved from http://www.oerj.org/View?action=viewPaper&paper=6

Hattie, J. A. C., Brown, G. T., & Keegan, P. (2005). A national teacher-managed, curriculum-based assessment system: Assessment Tools for Teaching and Learning (asTTle). *International Journal of Learning, 10*, 770–778.

Hattie, J. A. C., & Brown, G. T. L. (2004). *Cognitive processes in assessment items: SOLO taxonomy* (Tech. Rep. No. 43). Auckland, New Zealand: University of Auckland, Project asTTle.

Hoel, B. (2004). Predicting performance. *Credit Union Management, 27*, 24–26.

Hoijtink, H. (1990). A latent trait model for dichotomous choice data. *Psychometrika, 55*, 641–656. doi:10.1007/BF02294613

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement, 15*, 153–169. doi:10.1177/014662169101500205

Holmbeck, G. N., Thill, A. W., Bachanas, P., Garber, J., Miller, K. B., Abad, M., . . . Zukerman, J. (2008). Evidence-based assessment in pediatric psychology: Measures of psychosocial adjustment and psychopathology. *Journal of Pediatric Psychology*, *33*, 958–980. doi:10.1093/jpepsy/jsm059

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3*, 29–51. doi:10.1146/annurev.clinpsy.3.022806.091419

Huprich, S. K., & Bornstein, R. F. (2007). Dimensional versus categorical personality disorder diagnosis: Implications from and for psychological assessment. *Journal of Personality Assessment, 89*, 1–2. doi:10.1080/00223890701356854

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action. *Applied Psychological Measurement, 25*, 211–220. doi:10.1177/01466210122032028

Kamens, D. H., & McNeely, C. L. (2007, March). *International benchmarking and national curricular reform: Educational goal setting and assessment effects*. Paper presented at the Annual Conference of the Comparative and International Education Society, Baltimore, MD.

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review, 54*, 5–25. doi:10.1086/648471

Kane, J. S. (1997). Assessment of the situational and individual components of job performance. *Human Performance*, *10*, 193–226. doi:10.1207/s15327043hup1003_1

Krieshok, T. S., Ulven, J. C., Hecox, J. L., & Wettersten, K. (2000). Resume therapy and vocational test feedback: Tailoring interventions to self-efficacy outcomes. *Journal of Career Assessment*, *8*, 267–281. doi:10.1177/106907270000800305

Leeson, H. (2008). *Maximizing information: Applications of ideal point modeling and innovative item design to personality measurement*. Unpublished doctoral dissertation, University of Auckland, New Zealand.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511611186

Levine, M. V. (1984). *An introduction to multilinear formula score theory (Measurement Series 84–4)*. Champaign: University of Illinois, Model Based Measurement Laboratory.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Le, V., & Martinez, F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*, 47–67. doi:10.1111/j.1745-3984.2007.00026.x

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luecht, R. M. (2007, April). *Assessment engineering in language testing: From data models and templates to psychometrics*. Invited paper (and symposium) presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M. (2008, October). *Assessment engineering in test design, development, assembly, and scoring*. Paper presented at the East Coast Organization of Language Testers Conference, Washington, DC. Retrieved from http://www.govtilr.org/Publications/ECOLT08-AEKeynote-RMLuecht-07Nov08%5B1%5D.pdf

Luecht, R. M. (2009, June). *Adaptive computer-based tasks under an assessment engineering paradigm*. Paper presented at the Item and Pool Development Paper Session, Graduate Management Admission Council, Washington, DC.

Madhyastha, T. M., & Hunt, E. (2009). Mining diagnostic assessment data for concept similarity. *Journal of Educational Data Mining, 1*, 72–91.

Maguire, T. O. (1988, December). *The use of the SOLO taxonomy for evaluating a program for gifted students*. Paper presented at the Annual Conference of the

Australian Association for Research in Education, University of New England, Armidale, NSW.

Maguire, T. O., Hattie, J. A., & Haig, B. (1994). Construct validity and achievement assessment. *Alberta Journal of Educational Research, 40*, 109–126.

Mariano, L. T., McCaffrey, D. F., & Lockwood, J. R. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics, 35*, 253–279. doi:10.3102/1076998609346967

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174. doi:10.1007/BF02296272

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research, 40*, 261–279. doi:10.1207/s15327906mbr4002_5

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs, No. 15*.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Meagher-Lundberg, P., & Brown, G. T. L. (2001). *Item signature study: Report on the characteristics of reading texts and items from calibration 1*. (Tech. Rep. No. 12). Auckland, New Zealand: University of Auckland, Project asTTle.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135. doi:10.1177/01466210122031957

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education, Macmillan.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment. *American Psychologist*, *56*, 128–165. doi:10.1037/0003-066X.56.2.128

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence centered design* (ETS RR-03-16). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidenced-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist, 59*, 150–162. doi:10.1037/0003-066X.59.3.150

Oswald, F. L., & Schell, K. L. (2010). Developing and scaling personality measures: Thurstone was right-but so far, Likert was not wrong. *Industrial and Organizational Psychology, 3*, 481–484. doi:10.1111/j.1754-9434.2010.01275.x

Pace, V. L., & Brannick, M. T. (2010). Improving predication of work performance through frame-of-reference consistency: Empirical evidence using openness to experience. *International Journal of Selection and Assessment, 18*, 230–235. doi:10.1111/j.1468-2389.2010.00506.x

Padilla, A. M., & Borsato, G. N. (2008). Issues in culturally appropriate psychoeducational assessment. In L. A. Suzuki & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (3rd ed., pp. 5–21). San Francisco, CA: Jossey-Bass.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 28*, 238–243.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. J. van der linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129–148). Dordrecht, the Netherlands: Kluwer.

Percevic, R., Lambert, M. J., & Kordy, H. (2004). Computer-supported monitoring of patient treatment response. *Journal of Clinical Psychology, 60*, 285–299. doi:10.1002/jclp.10264

Purpura, D. J., Wilson, S. B., & Lonigan, C. J. (2010). Attention-deficit/hyperactivity disorder symptoms in preschool children: Examining psychometric properties using item response theory. *Psychological Assessment, 22*, 546–558. doi:10.1037/a0019581

Raja, U., & Johns, G. (2010). The joint effects of personality and job scope on in-role performance, citizenship behaviors, and creativity. *Human Relations, 63*, 981–1005. doi:10.1177/0018726709349863

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson & S. Hershberger (Eds.), *The new rules of measurement: What every psychologist should know* (pp. 219–241). Mahwah, NJ: Erlbaum.

Reise, S. P. (2010). Thurstone might have been right about attitudes, but Drasgow, Chernyshenko, and Stark fail to make a case for personality. *Industrial and Organizational Psychology, 3*, 485–488. doi:10.1111/j.1754-9434.2010.01276.x

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI–R. *Assessment, 7*, 347–364. doi:10.1177/107319110000700404

Resnick, L. B., & Berger, L. (2010, March). *An American examination system.* Paper presented at the National Conference on Next Generation Assessment Systems. Washington, DC.

Reynolds, C. R. (2000). Methods for detecting and evaluating cultural bias in neuropsychological assessment tests. In E. Fletcher-Janzen, T. L. Strickland, & C. R. Reynolds (Eds.), *Handbook of cross-cultural neuropsychology* (pp. 249–285). New York, NY: Kluwer Academic/Plenum Press. doi:10.1007/978-1-4615-4219-3_15

Ripley, M. (2007). *The four changing faces of e-assessment 2006–2016.* Retrieved from http://insight.eun.org/ww/en/pub/insight/thmatic_dossiers/articles/e_assessment/eassessment2.htm

Robert, C., & Cheung, Y. H. (2010). An examination of the relationship between conscientiousness and group performance on a creative task. *Journal of Research in Personality, 44*, 222–231. doi:10.1016/j.jrp.2010.01.005

Roberts, B. W., Chernyshenko, O., Stark, S., & Goldberg, L. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology, 58*, 103–139. doi:10.1111/j.1744-6570.2005.00301.x

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3–32. doi:10.1177/01466216000241001

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59*, 211–233. doi:10.1177/00131649921969811

Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessment. *Educational Measurement: Issues and Practice, 29*(3), 25–38. doi:10.1111/j.1745-3992.2010.00181.x

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155–180. doi:10.1016/j.hrmr.2006.03.004

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An Item Response Theory analysis of the MMPI–2 Psy-5 scales. *Journal of Personality Assessment, 72*, 282–307. doi:10.1207/S15327752JP720212

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment, 8*, 4–47.

Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response theory models. In J. Leighton & M. J. Gierl (Eds.), *Cognitively diagnostic assessment for education: Theory and applications* (pp. 205–241). Cambridge,England: Cambridge University Press. doi:10.1017/CBO9780511611186.008

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Samejima, F. (1997). Graded response model. In W. van Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer-Verlag.

Scherbaum, C. A., Finlinson, S., Barden, K., & Tamanini, K. (2006). Applications of item response theory to measurement issues in leadership research. *Leadership Quarterly, 17*, 366–386. doi:10.1016/j.leaqua.2006.04.005

Schleicher, A. (2009, September). *Benchmarking the performance of education internationally*. Paper presented at the Learning and Technology World Forum, Washington, DC.

Schrag, F. (1989). Are there levels of thinking? *Teachers College Record, 90*, 529–533.

Scriven, M. (1967). *The methodology of evaluation.* Washington, DC: American Educational Research Association.

Scriven, M. (1990). Beyond formative and summative. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education at quarter century* (pp. 19–64). Chicago, IL: University of Chicago/National Society for the Study of Education.

Scriven, M. (2005, October). *Key Evaluation Checklist: Intended for use in designing, evaluating, and writing evaluation reports on programs, plans, and policies; and for evaluating evaluations of them.* Retrieved from http://www.wmich.edu/evalctr/archive_checklists/kec_feb07.pdf

Segall, D. O. (2001, April). *Measuring test compromise in high stakes computer testing: A Bayesian strategy for surrogate test-taker detection.* Paper presented at the National Council on Measurement in Education, Seattle, WA.

Shadel, W. G. (2010). Clinical assessment of personality: Perspectives from contemporary personality science. In J. E. Maddux & J. P. Tangney (Eds.), *Social foundations of clinical psychology* (pp. 329–348). New York, NY: Guilford Press.

Shaffer, D. W., & Gee, J. P. (2008). *How computer games help children learn.* New York, NY: Palgrave Macmillan.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from

group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159–194. doi:10.1007/BF02294572

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, NJ: Erlbaum.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). A*utomated essay scoring: Writing assessment and instruction.* Educational Measurement.

Singer, J. D., & Willet, J. B. (2003). *Applied longitudinal data analysis.* London, England: Oxford University Press. doi:10.1093/acprof:oso/9780195152968.001.0001

Smith, F., & Hardman, F. (2003). Using computerized observation as a tool for capturing classroom interaction. *Educational Studies, 29*, 39–47. doi:10.1080/03055690303264

Smith, F., Hardman, F., Wall, K., & Mroz, M. (2004). Interactive whole class teaching in the national literacy and numeracy strategies. *British Educational Research Journal, 30*, 395–411. doi:10.1080/0141192 0410001689706

Smith, P., Lane, E., & Llorente, A. M. (2008). Hispanics and cultural bias: Testing development and applications. In A. M. Llorente (Ed.), *Principles of neuropsychological assessment with Hispanics: Theoretical foundations and clinical practice* (pp. 136–163). New York, NY: Springer. doi:10.1007/978-0-387-71758-6_7

Smith, T. W., Baker, W. K. Hattie. J. A., & Bond, L. (2008). A validity study of the certification system of the National Board for Professional Teaching Standards. In L. Ingvarson & J. A. C. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards* (pp. 345–380). Advances in Program Evaluation Series No. 11, Oxford, England: Elsevier.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101. doi:10.2307/1412159

Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A Festschrift honouring Donald W. Fiske* (pp. 161–181). Hillsdale, NJ: Erlbaum.

Suzuki, L. A., & Ponterotto, J. G. (Eds.). (2008). *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (3rd ed.). San Francisco, CA: Jossey-Bass.

Swaminathan, H. (1984). Factor analysis of longitudinal data. In H. G. Law, C. W. Snyder, J. A. Hattie, & R. P. McDonald (Eds.), *Research methods for multimode data analysis* (pp. 308–332). New York, NY: Praeger.

Tharinger, D. J., Finn, S. E., Hersh, B., Wilkinson, A., Christopher, G., & Tran, A. (2008). Assessment feedback with parents and pre-adolescent children: A collaborative approach. *Professional Psychology: Research and Practice, 39*, 600–609. doi:10.1037/0735-7028.39.6.600

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567–577. doi:10.1007/BF02295596

Thomas, G., Tagg, A., Holton, D., & Brown, G. T. L. (2002). *Numeracy item signature study: A theoretically derived basis.* (Tech. Rep. No. 25). Auckland, New Zealand: University of Auckland, Project asTTle.

Thomas, M. L., & Locke, D. E. C. (2010). Psychometric properties of the MMPI–2–RF somatic complaints (RC1) scale. *Psychological Assessment, 22*, 492–503. doi:10.1037/a0019229

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review, 34*, 273–286. doi:10.1037/h0070288

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554. doi:10.1086/214483

Thurstone, L. L. (1929). Theory of attitude measurement. *American Journal of Sociology, 33*, 529–545.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57–70). Vancouver, British Columbia, Canada: Educational Research Institute of British Columbia.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, *16*(4), 8–14. doi:10.1111/j.1745-3992.1997.tb00603.x

Trinder, J. J., Mahill, J. V., & Roy, S. (2005). Portable assessment: Towards ubiquitous education. *International Journal of Electrical Engineering Education, 42*, 73–78.

Uomoto, J. M., & Wong, T. M. (2000). Multicultural perspectives on the neuropsychology of brain injury assessment and rehabilitation. In E. Fletcher-Janzen, T. L. Strickeland, & C. R. Reynolds (Eds.), *Handbook of cross cultural neuropsychology* (pp. 169–184). New York, NY: Springer. doi:10.1007/978-1-4615-4219-3_11

van der Linden, W. J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement, 20*, 373–388. doi:10.1177/01466216 9602000405

van der Linden, W. J. (2005). *Linear models for optimal test design.* New York, NY: Springer.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.

van der Linden, W. J., & Luecht, R. M. (1996). An optimization model for test assembly to match observed-score distributions. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 405–418). Norwood, NJ: Ablex.

van de Vijver, F. J. R. (2008). Personality assessment of global talent: Conceptual and methodological issues. *International Journal of Testing, 8*, 304–314. doi:10.1080/15305050802435011

van de Vijver, F. J. R., & Phalet, K. (2004). Assessment in multicultural groups: The role of acculturation. *Applied Psychology*, *53*, 215–236. doi:10.1111/j.1464-0597.2004.00169.x

van Schuur, W. H. (1984). *Structure in political beliefs: A new model for stochastic unfolding with application to European party activists.* Amsterdam, the Netherlands: CT Press.

Verhelst, N. D., & Verstralen, H. H. F. M. (1993). A stochastic unfolding model derived from the partial credit model. *Kwantitativea Methoden, 42*, 73–92.

Waples, C. J., Weyhrauch, W. S., Connell, A. R., & Culbertson, S. S. (2010). Questionable defeats and discounted victories for Likert rating scales. *Industrial and Organizational Psychology, 3*, 477–480. doi:10.1111/j.1754-9434.2010.01274.x

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum.

Wilson, M., Bejar, I., Scalise, K., Templin, J., Wiliam, D., & Torres-Irribarra, D. (2012). Perspectives on methodological issues. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st-century skills* (pp. 67–142). Dordrecht, the Netherlands: Springer.

Wise, E. (2010). Introducing the MMPI–2–RF into clinical practice. *In Practice, 30*, 75–77.

Witt, E. A., & Donnellan, M. B. (2008). Furthering the case for the MPQ-based measures of psychopathy. *Personality and Individual Differences, 45*, 219–225. doi:10.1016/j.paid.2008.04.002

Witt, E. A., Hopwood, C. J., Morey, L. C., Markowitz, J. C., McGlashan, T. H., Grilo, C. M., . . . Donnellan, M. B. (2010). Psychometric characteristics and clinical correlates of NEO PI–R fearless dominance and impulsive antisociality in the Collaborative Longitudinal Personality Disorders Study. *Psychological Assessment*, *22*, 559–568. doi:10.1037/a0019617

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223–233.

Index

Volume numbers are printed in boldface type, followed by a colon and the relevant page numbers.

APA Handbook of Testing and Assessment in Psychology, Vol. 3: Testing and Assessment in School Psychology and Education, edited by K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, and M. C. Rodriguez