UNIT 6 CRITERIA OF A GOOD TOOL

Structure

6.1 INTRODUCTION

In Unit-5, you have studied about the techniques of evaluation such as observation, interview, rating scale, socio-metric techniques, seminar, group discussion, etc. From that Unit, you might have understood the techniques of using various assessment tools for assessment and evaluation. In this Unit, we will discuss the types of tools and the essential criteria of a good tool. During the course of discussion you will understand that there are different criteria to be taken into consideration while evaluating a tool or determining the worth of a tool to be used. In this regard, broadly we will discuss the criteria of an effective tool such as reliability, validity, usability, objectivity and norm of a tool and how to determine the index of reliability and validity for making it an effective tool. Again, the effective tools are used for satisfying the purpose of its evaluation. According to this purpose, we can classify the tools into two broad categories, that are self-made tool and standardized tool. This Unit will also make you understand the concept of self-made and standardized tools and also to differentiate between them.

6.2 OBJECTIVES

After going through this Unit, you should be able to:

- define self-made and standardized tools,
- differentiate self-made and standardized tool,
- discuss the different criteria of an effective tool,
- explain the relationship between validity and reliability,
- describe the procedure to find out reliability and validity of a tool,
- calculate the index of reliability and validity of a tool, and
- explain the criteria of an effective tool.

6.3 EVALUATION TOOLS : TYPES AND DIFFERENCES

We use a variety of tools for the purpose of assessment of performances of the students in the school. We also use different types of tools for collecting data to conduct various projects and research studies. It is also equally used for various administrative purposes, for example, recruitment for various jobs. That is why we say that for assessing something, we need to use a tool. Now the question is as to which type of tool can be used for what purpose. On the basis of the purpose, tools are classified into two broad types namely self-made and standardized tools. In this Section, you will learn about both types of tools and also you will be able to differentiate them.

6.3.1 Self-made and Standardized Tools

As discussed, we usually use two types of tools, self-made and standardized, in assessment and evaluation. Let us understand the concept of both the tools.

Self-made tool : Self-made tool is also popularly known as teacher-made tool. This type of tool is prepared for assessing performance of the students in various subjects. These types of tools are usually prepared by those class teachers who are engaged in teaching the subject to the particular class. Self-made tools are meant to limited purpose for the particular groups of students to whom the teacher teaches. You as a teacher, should be engaged in preparing and administering such type of tools in your class. You may be engaged in preparing a unit, quarterly or half-yearly test for assessing the performance of your students. This is an example of teacher-made tool.

Standardized tool : When we prepare a tool by following a set procedure of test construction and develop norms for its use, then it is called standardized tool. This is generally prepared keeping in view the large group. All the criteria for preparing an effective tool are followed in a standardized tool. A self-made tool can also be a standardized tool, provided the procedures and steps are followed in developing the same. Generally, standardized tools are developed by specialists or experts in the field. The norms, such as age, gender, percentile, habitation, etc. are also developed in a standardized tool. A test prepared for classes 10 or 12 Board Examination by Central Board of Secondary Education (CBSE) or by any State Board can be an example of standardized test. Now you might have understood that the standardized tool is used on a large population, say entire country, entire State or a University. The entrance examination conducted by the CBSE for admitting students in medical or engineering programmes is also an example of a standardized test.

6.3.2 Difference between Self-made and Standardized Tool

We hope that you have understood the concept of self-made and standardized tool. Now we shall clarify the difference between the self-made and standardized tool. Read the points given in Table 6.1.

Table 6.1: Difference between self-made and standardized tool

Self or Teacher-made Tool	Standardized Tool
The purpose of teacher-made tool is to know the progress of the students in a particular subject, especially at the school stage.	The purpose of standardized tool is to admit the students in various courses and also to provide them placement or employment.
Teacher-made tool is made by the teacher for the students to whom he/she teaches.	Standardized tools are prepared by the test specialists and the experts in the field.
Teacher-made tool is prepared for the small group students in a school or class.	Standardized tools are prepared for the large group population.
The use of teacher-made tool is very limited.	The use of standardized tool is very vast.
The rigorous procedures of tool construction are not followed in teacher-made tool.	The complete procedure of test standardization is followed in standardized test.
Norms are not developed for developing teacher-made tool.	Developing norms are compulsory in standardized tool.
No statistical techniques are used to determine the reliability and validity of the teacher-made tool.	Rigorous statistical methods are used to determine the reliability and validity of the test.
Item analysis and item discrimination are not calculated in teacher-made tool.	Item-wise analysis is made to know the item difficulty and discrimination before including the item in the final test.
The difficulty level of the item in teacher-made tool depends upon the standard of teaching of the teacher to the group. It may be easy, average or difficult.	The difficulty value of the item in standardized tools is generally average in standard.
Subjectivity in preparing and applying the teacher-made tool is always there.	Standardized tools are always objective in nature.
Teacher-made tool supports the principles and spirit of continuous comprehensive evaluation, specifically the formative assessment.	Standardized tools support the principles and spirit of summative evaluation.
The Unit tool, quarterly and half-yearly examinations conducted in the schools are examples of teacher-made tool.	The tests prepared for the Board Examination may be central or state level and the test prepared for admission, placement and recruitment for various vocations/jobs are also the examples of standardized tool.

The following learning points emerge to summarize the above table :

- Self-made tool is used for limited purpose that is in particular class whereas standardized tool is used for the large group of students that may even cover students of a state or country.
- Self-made tool is prepared by the class teacher whereas standardized tool is prepared by test specialists.
- All steps of tool construction including development of norms is followed in standardized tool whereas teacher-made tool follows only few steps upto getting the content validity.

- The difficulty level of items in teacher-made tool depends upon the standard of teaching by the teacher, that is it may be easy or difficult or average whereas the difficulty level of a standardized tool is always average in standard.

Activity 1

Analyse the Table 6.1 and summarize it on the following aspects :

Purpose of both type of tools :

.....

.....

.....

Process followed to prepare the tools :

.....

.....

.....

Main beneficiaries of the tools and the jurisdiction of its use :

.....

.....

.....

Examples of the tools :

.....

.....

.....

Check Your Progress 1

Note : a) Write your answers in the space given below.

b) Compare your answers with the ones given at the end of the unit.

1. Define self or teacher-made tool in your own words.

.....

.....

.....

.....

.....

2. Define standardized tool in your own words.

.....

.....

.....

.....

.....

6.4 ESSENTIAL CRITERIA OF AN EFFECTIVE TOOL OF EVALUATION

Developing an effective tool is a challenging task. You have to follow scientific and systematic procedure to develop or standardize a tool. To get accurate results, you have to prepare or select a proper tool. Before selecting a tool for certain purpose, you have to look into the criteria or qualities of the tool. The essential qualities or criteria of an effective tool may be as follows:

- Reliability
- Validity
- Usability
- Objectivity
- Norm

Let us now discuss each one of these criteria in detail.

6.4.1 Reliability

Reliability is the important criteria of a good test/tool. Reliability refers to consistency. A test which shows a consistent result in its frequent uses in different situations and places is called reliability of the test. The other synonyms that can be used for getting reliability of the test are: dependability, stability, consistency, predictability, accuracy, etc. It implies that the reliable test always provides a stable, dependable, accurate and consistent result in its subsequent uses. Before discussing the methods or techniques of determining reliability, it is worthwhile to observe here some points cited by Gronlund (1981) :

i. Reliability refers to the results obtained with an evaluation instrument and not to the instrument itself. An evaluation tool may have a large number of different reliabilities depending on the groups of subjects and situations of use.

ii. Test scores are not reliable in general. An estimate of reliability always relates to a particular type of consistency – say consistency of scores over a period of time (stability) or consistency of scores over different samples of questions (equivalence), etc.

iii. Reliability is a necessary but not a sufficient condition for validity. A valid test is always reliable, but a reliable test may or may not be valid.

iv. Reliability is primarily statistical in nature in the sense that the scores obtained on two successive occasions are correlated with each other.

v. The coefficient of correlation is known as self-correlation and its value is called the 'reliability coefficient'.

*(**Source :** ES – 333; Educational Evaluation, IGNOU, 2010)*

Methods or techniques of reliability : As per Singh (2002), there are three common methods of estimating the reliability coefficient of test scores. These methods are :

(i) Test-retest reliability.

(ii) Parallel-forms reliability.

(iii) Internal consistency reliability.

Let us discuss each method of reliability in detail.

(i) Test-retest reliability : Test-retest reliability means the same test is administered twice on the same group of sample within a given time interval and correlation is calculated between the two sets of scores (first and second administration). If the coefficient of correlation is positive and high, it is considered that the test is reliable. Let us discuss the procedures of using test-retest reliability.

Table 6.2 : Precedures of using test-retest reliability

Conditions	Use
Test	Single form
Administer	Twice
Group	Single
Time interval	Ideally 15 days to 6 Months
Determining reliability	By employing correlation statistics
Decision	Reliable if correlation coefficient is positive and high. (r = +0.5).
Threat of error	Error of 'carry over effect', 'memory effect', 'practice effect', etc.
Suggestion	Highly used method as it is easy to determine reliability of the test, but because of the error factor, test–retest method should be the last option for the user.

After analysis of Table 6.2, we can say that test-retest method is one of the easiest method to determine reliability of the test. In this method, the same test is administered twice within a gap of time limit to minimum 15 days to maximum six months. After getting the scores of test and re-test results, any method of correlation either Spearman's Rank Difference or Pearson's Product Moment Correlation method are used to get the coefficient of correlation (please read Unit-16, Block-4, BES-127). As this has already mentioned in the Table, if the coefficient of correlation (r) is = +0.5, we can say that the test is reliable. The following methods can be employed to calculate correlation of the testing and re-testing scores.

Spearman's Rank Difference Method (ρ)

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

ρ = Is called as 'rho', means correlation in test-retest method.

d = Deviations ($R_1 - R_2$) [Rank of the candidates in first test score (R_1) and Rank of the candidates in second test score (R_2)]

n = Is total number of candidates

Karl Pearson's Product Moment Method (r)

$$r = \frac{\sum xy}{N\sigma x\sigma y}$$

r = Correlation coefficient

x = Deviations taken from mean of the distribution (X-M) [X – Individual Score and M – Mean of the first distribution]

y = Deviations taken from mean of the distribution (Y-M) [Y – Individual Score and M – Mean of the second distribution]

N = Total number of the candidates

σx = Standard deviation of first distribution

σy = Standard deviation of second distribution

Limitations of using the method : The following are the limitations of the test-retest method :

- As the same test is administered twice on the same group, there will be the threat of carry over effect, it means, during the second administration, the candidates may remember many items from the first administration.
- The scoring of second administration is usually high than the first one.
- Maintaining a gap of time between test and re-test is also again one of the important aspects to determining exact value of reliability. If time gap is very less, then carry over effect will be high and on the other side, if time gap is very high, maturity effects of the candidates may hamper the test results.
- This method is not free from errors. Memory, carry over, practice and maturity effects are high in this technique.

Activity 2

Prepare a small test and find out reliability of the test by using test-retest method of reliability.

.....

.....

.....

.....

.....

(ii) Parallel-form Reliability : Because of the error factors in test-retest method, parallel-form method is one of the alternate methods of the test-retest method and it can minimize many of the errors occurred in the earlier method. In the parallel form method, two parallel tests are prepared keeping in consideration equivalence in all aspects such as similarities in content, objectives, types and number of items, time allowed in both the tests, level of difficulty, discrimination value, conditions of use, etc. The main effort by doing these is to make two equivalent forms of test.

Let us discuss the procedures and conditions of using parallel-form method for determining reliability.

Table 6.3 : Precedures of using parallel-form method for reliability

Conditions	Use
Test	Two equivalent forms
Administer	Twice (First test for first administration and second test for second administration)
Group	Single
Time interval	This is advisable to maintain a gap of minimum 15 days to maximum six months for first and second testing.
Determining reliability	Use of correlation statistics by the Spearman or Pearson formula of correlation.
Decision	Reliable if correlation coefficient is positive and high. (r = +0.5).
Threat of error	Error of 'carry over effect', 'memory effect', 'practice effect', etc. are not totally minimized, but in comparison to test-retest method, the occurrences of error is comparatively less. Here there is no question of splitting the items into two equal halves.
Suggestion	As it is an advanced method of test-retest method, this can be used in many situations, but the main challenge here is to develop exactly two equivalent forms of test.

In Table 6.3 we have stated the processes adopted in the parallel-form method to determine the reliability of a test. In this method, two parallel tests are developed which are equivalent in all respects, such as content, objectives, types of items, time given, difficulty level of the item, etc. When two parallel tests are used to get the reliability of the test, it is quite natural that the carry over, memory, and practice effects are highly minimized. For example, when items are different, there is very less chance to remember the earlier questions. Just like the test-retest method, the first form of the test is administered on the group and after a gap of some time period say, from a minimum of 15 days to a maximum of 6 Months, the second form of the test may be used. Spearman's or Karl Pearson's method can be used to get the coefficient of correlation of the two administrations. If correlation coefficient is, $r \geq +0.5$, then the test is said to be reliable.

Limitation of parallel form method : Parallel form method is also not completely free from errors. There are possibilities of making errors in this method also:

- Practice and carry over effect is not totally minimized, as both the tests are equivalent in nature in many respects except only the items are different and a time interval of 15 days to 6 months is given for testing the second form of the test. During this period, there is a chance that

the students may practice the similar content and items and hence chances for getting better scores in second testing are generally more.

- Preparing two parallel forms of the tests is also a complex task.
- This method is comparatively time taking to get the reliability.

(iii) Internal consistency reliability : Internal consistency reliability indicates the homogeneity of the test. If all the items of the test measure the same function or trait, the test is said to be a homogeneous one and its internal consistency reliability would be pretty high. The most common methods of estimating internal consistency reliability are the (a) Split-half method and (b) Rational equivalence method. Let us discuss split-half method first.

(a) Split-half method : This method is also called as 'odd-even method'. Let us discuss the procedures and conditions of using split-half method for determining reliability.

Table 6.4 : Precedures of using split-half reliability method

Conditions	Use
Test	Single form
Administer	Once
Group	Single
Time interval	Not necessary as the test will administer only once.
Determining reliability	By employing correlation statistics as well as split-half reliability formula.
Decision	Reliable if reliability index is positive and high. (r = +0.5).
Threat of error	Error of 'carry over effect', 'memory effect', 'practice effect', etc. are minimized, but still splitting the full test into two halves is also difficult because splitting the full test into two-halves can be made in many different ways such as odd and even, first fifty percent and last fifty percent, and so on.
Suggestion	Is one of the highly used method as most errors are minimized.

Table 6.4 reveals the procedures of using split-half reliability method. Most of the errors that occur in test-retest method are minimized in split-half method as the test is not conducted twice on the same group. In this method, the test is to be administered only once to a single group and after getting the scores of the students, it has to be split into two equal halves like in odd and even items, or any other technique to divide it into two halves. The scores of the candidates for odd items and for even items will be separated and suitable statistical techniques are used to get the reliability of the scores of two halves. The Spearman-Brown Prophecy formula is generally used for calculating reliability of the full test. The formula is as follows:

$$r_{tt} = \frac{n\sigma 2 - M(n-M)}{\sigma 2\,(n-1)}$$

r_{tt} = Reliability of the whole test

$r_{1/2}$ = Correlation of the two halves of the test

Just like, test-retest method, correlation of the two halves of the test ($r_{1/2}$) will be calculated by using the method of Spearman or Pearson's correlation. Splitting of the item can be done as shown in Table 6.4.

Table 6.5: Worksheet for the Odd-Even Reliability

Examinee	Total No. of correct scores	Number of correct scores on odd-numberred items	Number of correct score on even-numbered items
1	50	29	21
2	46	20	26
3	55	28	27
4	39	19	20
5	55	25	30
6	49	27	22
7	60	32	28
8	42	20	22
9	45	24	21
10	53	28	25

Table 6.5 represents the total number of correct scores of the examinees and also the number of correct scores on odd numbered items and on even numbered items. For example, the examinee at serial number 1 scored a total of 50 out of which 29 scored from the odd items and 21 from the even items. Accordingly, all examinee's scores are grouped for odd and even items and made into two groups for getting the reliability of the whole test.

Limitations of this method : The important error in this method is to split the full test into two equal halves. This can be done in many ways like odd-even, first 50 percent items and next 50 percent items, etc. The correlation of two-haves of the test and the reliability of the full test will also be different based on different ways of splitting the test items. Besides this type of errors, the other type of errors such as 'carry over error', 'maturity error', 'memory and practice effect' etc. are mostly minimized in this method.

Activity 3

With the example presented in Table 6.5, calculate reliability coefficient of whole test by using Spearman-Brown Prophecy formula.

.....

.....

.....

(b) Rational equivalence method : This is purely a statistical process of determining reliability of a test. The formula for calculating reliability in rational equivalence method is developed by *Kuder and Richardson.* They developed two formulae popularly known as Kuder and Richardson (KR) formula number 20 and 21 (KR 20 and KR 21). Lee J. Cronbach has called it *Coefficient of Internal Consistency.* This method of reliability is an improvement of all the earlier methods as it minimises errors in calculating reliability. You should also understand that rational equivalence method is the improvement of split-half method. The assumption in *Kuder and Richardson* formula is that all items have the same or equal difficulty value, but not necessarily the same persons solve each item correctly. The difficulty value of an item depends on knowing aspect of a person. If the person knows the item, it is easy for him/her and if he/she does not know the item, it is difficult for him/her. The assumption is that the test is homogeneous in nature.

Table 6.6: Precedures of using rational equivalence method

Conditions	**Use**
Test	One
Administer	Once
Group	Single
Time interval	No need
Determining reliability	By employing KR-20 and KR-21 formullae.
Decision	Reliable if reliability index is positive and high. ($r \geq +0.5$).
Threat of error	Errors are mostly minimized
Suggestion	Can be used effectively in research

Table 6.6 reveals the conditions of using rational equivalence method in determining reliability of a test. This method statistically eliminates most of the errors of determining reliability of a test. It requires certain statistical applications to determine the reliability, that's why the skill of applying statistics is required to get the reliability. As indicated, all most all errors are minimized in this method besides the human and calculation errors. This is highly used in tool construction for research. The formula of *KR-20* and *KR-21* are as follows :

KR-20 formula is the more accurate and is given as follows :

$$r = \left(\frac{n}{n-1}\right)\left(\frac{\sigma^2 - \sum pq}{\sigma^2}\right) \qquad \text{—— } KR\text{-}20$$

r = Reliability Index

σ^2 = Variance of the total test (square of the standard deviation)

n = Number of items in the test

p = Proportion of right responses

q = (1-p) proportion of wrong responses

KR-21 formula is given as follows :

r = —— *KR-21*

r = Reliability Index

σ^2 = Variance of the total test (square of the standard deviation)

n = Number of items in the test

M = Mean of the test scores

Limitations of rational equivalence Method : The following are the main limitations of this method :

- The item variation of item difficulty is not considered in calculating reliability index.
- The reliability coefficient is approximate and lower than the accurate reliability coefficient.
- It can not be used for power test or heterogeneous test.
- The different KR formula yields different reliability coefficient.

Factors influencing Reliability Coefficient : Factors influencing reliability of a test can be classified as extrinsic factors and intrinsic factors.

Extrinsic factors : The important extrinsic factors (i.e. the factors which remain outside the test itself) influencing the reliability are :

i. **Group variability :** When the group of pupils being tested is homogeneous in ability, the reliability of the test scores is likely to be lowered and vice-versa. It is therefore, the method of randomization is used to get the groups having a representative of all abilities.

ii. **Guessing and chance errors :** Guessing in test results in to increased error variances and thus reduces reliability. For example, in two-alternative response options there is a 50% chance of answering the items correctly in terms of guessing. It is therefore difficult to determine the true scores of the examinees.

iii. **Environmental conditions :** As far as practicable, testing environment should be uniform like the proper facilities of light, air, seating arrangement, audibility of the instructions, time of testing, qualities of materials used for the test, etc.

iv. **Momentary fluctuations :** Momentary fluctuations may raise or lower the reliability of the test scores. A broken pencil, momentary distraction by the sudden sound of a train running outside, anxiety regarding non-completion of homework, mistake in giving the answer and knowing no way to change it are the factors which may affect the reliability of the test scores.

Intrinsic factors : The principal intrinsic factors (i.e. those factors which lie within the test itself) which affect the reliability are :

i. **Length of the test :** Reliability has a definite relation with the length of the test. The more the number of items the test contains, the greater will

be its reliability and vice-versa. Therefore adequate number of items needs to be included in the test.

ii. **Homogeneity of the items :** Homogeneity of items has two aspects : item reliability and the homogeneity of traits measured from one item to another. If the items measure different functions and the inter-correlation of items is 'zero' or near to it, then the reliability is 'zero' or very low or vice-versa.

iii. **Difficulty value of the items :** Broadly, items having indices of difficulty at 0.5 or close to it yield higher reliability than items of extreme indices of difficulty. It emphasizes that the items should be of the average difficulty values. Too much difficulty and too much easy items are usually avoided in the test.

iv. **Discriminative value :** When items can discriminate well between superior and inferior, the item total correlation is high, then the reliability is also likely to be high and vice-versa. It implies that the item has the quality to discriminate the lower and the higher group.

v. **Scorer reliability :** Scorer reliability, otherwise known as reader reliability, also affects the reliability of a test. Scorer reliability speaks of how closely two or more scores agree in scoring the same set of responses. The reliability is likely to be lowered if they do not agree.

*(**Source :** Factors Influencing Reliability, ES-333, IGNOU, 2010)*

Check Your Progress 2

Note : a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the unit.

3. Define reliability in your own words.

.....

.....

.....

4. Which is the best method for determining reliability and why?

.....

.....

.....

5. List any five factors which decrease the reliability of a test.

.....

.....

.....

6.4.2 Validity

In the preceding sub-section you have studied about the reliability. Let us now discuss about validity. Validity tells us about the accuracy and truthfulness of a test. The accuracy of a test can be said if and when the test measures the

the purpose it was constructed. Cronbach (1970) defines validity as 'validity is the extent to which a test measures what it purports to measure'. Further Freeman (1965) defines validity as 'an index of validity shows the degree to which a test measures what it purports to measure when compared with accepted criterion.'

From the above three definitions given we can say that validity talks not only about consistency but also accuracy and truthfulness of the test results. This can also be said that, validity is above reliability and it includes reliability also. See Figure 6.1 given below :

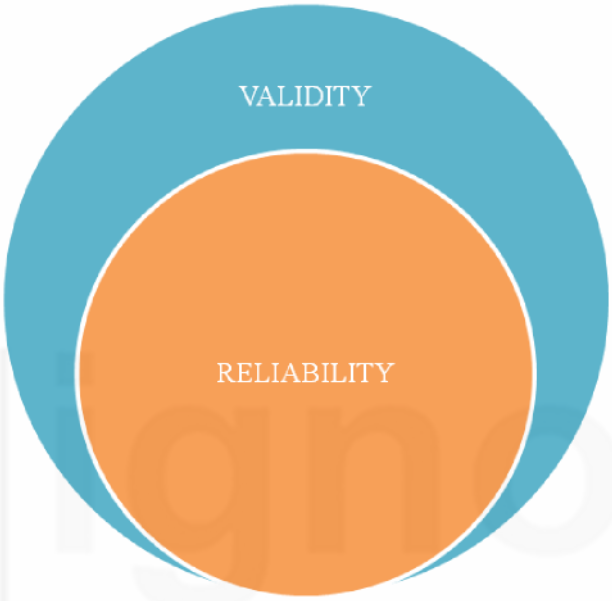

Figure 6.1: Validity and Reliability

From the definitions of validity, let us try to understand the characteristics of validity.

Characteristics of validity : The following are the main characteristics of validity of a test :

i. Validity is an index of external correlation. The test scores are correlated with external criterion scores such as the test scores will be correlated with an earlier developed valid test prepared for measuring the same aspect.

ii. The criterion may be a set of operation, purpose or predictor for future course of performance.

iii. It deals with the psychological construct of a variable which is indirectly measured with the help of behaviours.

iv. No test in education and psychology is perfectly valid because measurement is indirect.

v. Validity endorses the reliability of a test. If a test is valid, it must be reliable, but a reliable test may or may not be valid.

vi. It refers to the truthfulness or purposiveness of test scores.

vii. It indicates the degree to which the test is capable to achieve the aims for which it is developed.

viii. Validity is best considered in terms of matter of degrees, such as high, moderate and low validity.

Types of validity

Commonly, five types of validity are used in preparing tools. These are :

(i) Face Validity

(ii) Content Validity

(iii) Criterion-related Validity

- Predictive validity
- Concurrent validity

(iv) Construct Validity

(v) Factorial Validity

Let us understand the concept of the above types of validity.

(i) Face validity : Face validity is the first step to know validity of a test. This is also called validation by face. This method is not widely used because it never analyses the entire test and its items to determine the validity of the test. In the face validity, the appearance of the test, purpose of its construction, objectives it covers, dimensions it measures, language used, etc. are taken into consideration for determining the face validity of the test. This is the lowest level of determining validity of the test. This method can only be used in case of shortage of time for using other methods of validity. Further, before using other methods of determining validity, a judgment is taken, whether the test is validated by face or not. If the test lacks face validity then usually other methods of validity are not determined.

(ii) Content validity : Content validity is the second level of validity of the test. In this method, the format as well as the content of the test is examined and decision is taken for its validity. Anastasi (1968) defines content validity as, 'it involves essentially the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured.' Content validity basically matches the test items with the instructional objectives. Content validity is very important for achievement test. To know the content validity of a test, usually the items are examined as per the blue-print/table of specification on which the test is prepared. Let us try to understand, what is a blue-print or table of specification with the help of an example :

Example :

Test : Achievement Test of Social Science for Class-IX

Total Marks – 100

Total Item – 100 (Each item carries 1 mark)

Types of Items – Objective Type Item (multiple choice type with four options)

Weightage to Objective Areas – Knowledge (40%); Understanding (20%); Application (20%); and Skill (20%)

Weightage to Content Areas – History (25%); Political Science (25%); Economics (25%); and Civics (25%).

Blue-print/table of specification : A blue-print is a three dimensional chart, where weightage are given to Content, Objectives and Form of Questions in terms of marks. Here the first dimension is 'content', second dimension is 'objectives' and the third dimension is 'form of questions in terms of marks'.

Table 6.7: Blue-print of an Achievement Test

Content/Objectives	Knowledge	Understanding	Application	Skill	Total
History	10×1 = 10	5×1 = 5	5×1 = 5	5×1 = 5	25
Political Science	10×1 = 10	5×1 = 5	5×1 = 5	5×1 = 5	25
Economics	10×1 = 10	5×1 = 5	5×1 = 5	5×1 = 5	25
Civics	10×1 = 10	5×1 = 5	5×1 = 5	5×1 = 5	25
Total	**40**	**20**	**20**	**20**	**100**

As presented in Table 6.7, the content validity of a test depends upon as to how far the items are written as per the table of specification. The closer the test items correspond to the specified sample, the greater the possibility of satisfactory content validity. Therefore, it is desirable that the items in a test are screened by a team of experts; they should check whether the placement of various items in the cells of the table is appropriate and whether all the cells of the table have an adequate number of items. For determining content validity of the test, the experts require to examine the items with the textbook, syllabi, etc.

(iii) Criterion-related validity : Unlike content validity, criterion-related validity can be objectively measured and declared in terms of numerical indices. The concept of criterion-related validity focuses on a set of 'external' criterion. The external criterion may be data of 'concurrent' information or of a future performance. The criterion related to concurrent information is called as 'Concurrent Validity' and criterion related to future performance is called as 'Predictive Validity'.

- **Predictive validity :** Predictive validity refers to the predictive capacity of a test. It indicates the effectiveness of the test in forecasting or predicting future outcomes in a specific area. In short, predictive validity determines how far the test is able to predict future result. This can be better understood by an example. Suppose we have to prepare a medical entrance examination test to admit students in medical courses. The predictive validity of the test can be determined only when those qualified students who took admission in medical course performed well in the medical final examination. Predictive validity is a time consuming method. To get predictive validity of a test, you have to wait till the completion of the course. Sometimes, prediction can also take much time to correlate it with further criterion such as those students who performed well in the entrance examination, whether they successfully completed the course or not, and further, those students who successfully completed the course did get a job/ placement or not. So, to determine the predictive validity of a test, there is a need to establish correlation of the scores between entrance

examination result and the course completion result. If the correlation coefficient is positive and high, you can say that the test is valid.

This type of validity is sometimes referred to as 'empirical validity' or 'statistical validity' as our evaluation is primarily empirical and statistical. You can test the validity empirically.

- **Concurrent validity :** Concurrent validity refers to the extent to which the test scores correspond to already accepted measures of performance. For example, suppose you have prepared a test of 'intelligence' and you want to know the concurrent validity of the test, you have to correlate the scores of the test administration with the scores of another established standardized test. Let us understand it with the help of another example. The Intelligence test, which you have prepared and the intelligence test prepared by Stanford-Binet can be administered among the same group of students and correlation coefficient of two sets of scores can be determined. If the coefficient of correlation is high, we can say that the test has concurrent validity.

(iv) Construct validity : Construct validity is also called 'psychological or trait validity'. Construct validity means that the test scores are examined in terms of a construct. For example, the construct for achievement of a student may be his/her intelligence, practice, aptitude, interest, attitude, etc. Construct validity can be defined as the extent to which the test may be said to measure a theoretical construct or trait or psychological variable. In construct validity the variables related to the test which contributes to that aspect are correlated and examined.

For example, this is a theoretical fact that intelligence and achievement are positively correlated with each other. Suppose you have to prepare an intelligence test and you want to know the construct validity of the test. For that, you have to correlate the intelligence test scores of the students with their achievement test scores. The assumption here is that those students who have done well in intelligence test will naturally do well in achievement test, because as per the theory, both are positively correlated with each other. In case the correlation is negative, it can be said that the intelligence test is lacking construct validity. This can also be correlated with other theoretical and psychological construct of an intelligence test with the assumptions as follows :

- Intelligence and achievement are positively correlated with each other.
- Intelligence and aptitude are positively correlated with each other.

Construct validity is to the extent test results are interpreted in terms of known psychological concepts and principles. Certain common examples of theoretical constructs of most psychological tests are intelligence, scientific attitude, critical thinking, reading, comprehension, study skills and mathematical aptitude, etc.

(v) Factorial validity : Factorial validity determines the correlation of the different factors/components with the whole test. Factorial validity is determined by a statistical technique known as factor analysis. It uses methods of expansion of inter-correlations to identify factors (which may be verbalized as abilities) constituting the test. The correlation of the test

with each factor is calculated to determine the weight contributed by each such factor to the total performance of the test. This validity tells us about the factor loading. The factors responsible for achievement of students are called factor loading. This relationship of the different factors with the whole test is called the factorial validity. Factorial validity is the clearest description of what a test measures and by all means should be given preference over other types of validity.

Factors affecting validity : A large number of factors influence the validity of the test. Gronlund (1981) has suggested the following factors :

i. Factors in the test itself :

The following factors that affect validity of a test are included in the test itself. These are also called as intrinsic factors.

- **Unclear direction :** If directions regarding how to respond to the items, whether it is permissible to guess and how to record the answers, are not clear to the pupil, then the validity will tend to reduce. Hence, clear direction should be given in the test.
- **Reading difficult vocabulary and sentence structures :** The complicated vocabulary and sentence structure meant for the student taking the test may fail in measuring the aspects of pupil performances; thus it results in lowering the validity.
- **Inappropriate level of difficulty of the test items :** When the test items have an inappropriate level of difficulty, it will affect the validity of the tool. For example, in criterion referenced test, failure to match the difficulty specified by the learning outcome will lower the validity.
- **Poorly constructed test items :** The test items which provide unintentional clues to the answer will tend to measure the pupils' alertness in detecting clues as well as the aspects of pupil performance which ultimately affect the validity.
- **Ambiguity :** Ambiguity in statements in the test items leads to misinterpretation, multi-interpretations and confusion. Sometimes, it may confuse the good students more than the poor ones resulting in the discrimination of items in a negative direction. As a consequence, the validity of the test is lowered.
- **Test items inappropriate for the outcomes being measured :** Many a times we try to measure certain complex types of achievement, understanding, thinking, skills, etc. with test forms that are appropriate only for measuring factual knowledge. This affects the results and leads to a distortion of the validity.
- **Test too short :** A test usually represents a sample of many questions. If the test is too short to become a representative one, then validity will be affected accordingly.
- **Improper arrangement of items :** Items in the test are usually arranged in terms of difficulty with the easiest items first. If the difficult

items are placed early in the test, it may make the students spend too much of their time on these and fail to reach other items which they could answer easily. Also, such an improper arrangement may influence the validity by having a negative effect on pupil motivation.

- **Identifiable pattern of answer :** When the students identify systematic pattern of correct answer (e.g. T, T, F, F, T, T, F, F or A,B,C,D,A,B,C,D, etc.) they can cleverly guess the answers and this will affect the validity. It is therefore, in objective items, the order of answers should not follow any pattern.

ii. **Functioning content and teaching procedure :** In achievement testing, the functioning content of test items can not be determined only by examining the form and content of the test. The teacher has to teach fully how to solve a particular problem before including it in the test. Tests of complex learning outcomes seem to be valid if the test items function as intended. If the students have previous experience of the solution of the problem included in the test, then such tests are no more a valid measurement for measuring the more complex mental processes and thus affect the validity.

iii. **Factors in test administration and scoring :** The test administration and scoring procedure may also affect the validity of the interpretations from the results. For instance, in teacher-made tests factors like insufficient time to complete the test, unfair help to individual students, cheating during the examination, and the unreliable scoring of essay answers might lead to lower the validity. Similarly, in standardized tests the lack of standard directions and time limits, unauthorized help to students and errors in scoring, would tend to lower the validity. Whether it is a teacher-made test or standardized-test, adverse physical and psychological conditions during testing time may affect the validity.

iv. **Factors in students' response :** There are certain personal factors which influence the students' response to the test situation and invalidate the test interpretation. Students' emotional disturbance, lack of motivation, test anxiety, etc. may affect the validity.

v. **Nature of the group and the criterion :** Factors such as age, sex, ability level, educational and cultural background of the students influence the test measures. Therefore, the nature of the validation group should find a mention in the test manual. The nature of the criterion used is another important consideration while evaluating validity coefficient. For example, scores on a scientific aptitude test are likely to provide a more accurate prediction of achievement in an 'environmental studies' course. Other things being equal, the greater the similarity between the performance measured by the test and the performance presented in the criterion, the larger the validity coefficient.

*(**Source :** Factors affecting Validity, ES-333, IGNOU, 2010)*

Check Your Progress 3

Note : a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the unit.

6. Define validity in your own words.

.....

.....

.....

7. Explain predictive validity with an example.

.....

.....

.....

8. State any four important factors that influence validity of a test.

.....

.....

.....

6.4.3 Usability

Usability refers how successfully you, as a teacher, use the test in a classroom situation. It has been observed that, many highly valid tests lack the quality of usability. The user fails to understand or feels it difficult to use the test. Therefore, a good test should have the quality of usability. While selecting an evaluation tool, you should look for certain practical considerations like easy for administration and scoring, easy for interpretation, availability of comparable forms and cost of testing. All these considerations induce a teacher to use tools of evaluation and such practical considerations are referred to as the 'usability' of a tool of evaluation. In other words, usability means the degree to which the tool of evaluation can be successfully used by the teacher and school administrators. So usability of a test includes comprehensibility, easy for administration and scoring, easy for interpretation, appearance of the test, economy and availability of the test for use.

6.4.4 Objectivity

Objectivity is another important feature for a good test. Objectivity of a test refers to two aspects of the test, viz;

- Item-objectivity, and
- Scoring-objectivity.

Item-objectivity means the item is having only one right answer. Many a times, you might have observed that a single item has two or more related answers. That affects the validity of the test. Apart from these, ambiguous questions, lack of proper direction, double-barreled questions, questions with double negatives, essay type questions, etc. do not have objectivity. So much care has to be exercised while framing the questions.

Scoring-objectivity refers the test paper would fetch the same score. Scoring objectivity can be ensured by carefully maintaining the item-objectivity. In objective-type items, it is easy to ensure scoring-objectivity whereas in subjective item, certain precautions needs to be taken to ensure scoring-objectivity such as carefully phrasing the essay items, making proper directions of scoring, making the items short-answer type instead of essay type item, etc.

6.4.5 Norm

Generally norm is considered as a standard, but technically, there is difference between the concepts of norm and standard. Norm can be defined as the average or standard score on a particular test made by a specified population.' Thorndike and Hegen (1977) defines norm as 'average performance on a particular test made by a standardized sample.' Determining norm is one of the important criteria of a good test. Most standardized tests determine norm. Norm can be characterized as follows :

- It acts as a basis for interpreting test scores and minimize interpretive error of the test.
- It helps to transform the raw scores into standard scores or derived scores and put meaning to it.
- Norm suggests a level and therefore the individual departure from the level can be evaluated in quantitative term.
- Norms are necessary for the purpose of promotion, gradation, selection and classification of examinees.
- It refers to the average performance on a particular test made by standardized sample or specified population.

The procedure of deterring the norm is a challenging task. Without determining the norm of a test we can not say that the test is standardized. It is therefore, in a standardized test, usually the age, level, habitation, etc. are mentioned on which the test can be used and it also equally reflects in the interpretation of the test scores. For interpretation of scores of educational and psychological tests, different norms like age, grade, percentile, standard score, etc. are broadly employed.

Check Your Progress 4

Note : a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the unit.

9. Explain the concept of item-objectivity.

.....

.....

.....

10. What are the different aspects of determining usability of a test?

.....

.....

.....

6.5 LET US SUM UP

In this Unit, we presented the essential criteria of a good tool. In this context, we discussed first the concept of a teacher-made test or self-made test and standardized test. We also differentiated a self-made test and standardized test and also the purpose and context of its use. For clarifying criteria of a good tool, we discussed reliability, validity, usability, objectivity and norm. You were acquainted with the methods of determining reliability and validity of a test and you might have understood how to calculate the reliability and validity of a test. We have also discussed the other criteria of a tool such as objectivity, usability and norm. In this Unit, the concept of usability, objectivity and norm have been presented with examples.

6.6 REFERENCES AND SUGGESTED READINGS

Anastasi, Anne (1976). Psychological Testing, 4th ed., New York : Macmillan Publishing Co. Inc.

Cronback, L.J. (1970). Essentials of Psychological Testing, 3rd ed., New York : Harper and Row.

Ebel, Robert, L. (1996). Measuring Educational Achievement. New Delhi: Prentice-Hall of India.

Ebel, Robert, L. and Fristic, David, A. (1991). Essentials of Educational Achievement. New Delhi: Prentice-Hall of India.

Gronlund, N.E. (1981). Measurement and Evaluation in Teaching. 4th ed. New York : Macmillan Publishing Co. Inc.

Freeman, F.S. (1965). Theory and Practice of Psychological Testing (3rd ed.). Oxford and IBH Publishing Co. Pvt. Ltd., New Delhi.

IGNOU (2010). Criteria of a Good Tool (Unit 6; Block 2), Educational Evaluation (ES-333, B.Ed.), New Delhi: IGNOU.

Kerlinger, F.N. (1973). Foundations of Behavioral Research, 1st ed., New York : Holt, Rinchart and Winston Inc.

Nayak, B.K. and Rath, R.K. (2010). Measurement, Evaluation, Statistics and Guidance Services in Education, New Delhi: Axis Publications.

Nually, J.C. (1972). Educational Measurement and Evaluation. 2nd ed., New York : McGraw Hill Book Company.

Sharma, R.A. (2005). Mental Measurement and Evaluation. 2nd ed., Meerut : R. Lal Book Depot.

Singh, A.K. (2002). Tests, Measurements and Research Methods in Behavioural Sciences. 3rd ed., Patna : Bharti Bhawan.

Thorndike R. L. and E.P. Hagen (1977). *Measurement and Evaluation in Psychology and Education*. 4th ed, New York : John Wiley and Sons.

6.7 ANSWERS TO CHECK YOUR PROGRESS

1. Test prepared by the class-teacher for formative assessment of the students to whom he/she teaches which is not standardize is called as self-made or teacher-made test.

2. Standardized tool is prepared for the large group. The complete procedures for test construction/standardization are followed to standardize a test. The reliability, validity, norm, etc. are usually determined to standardize a tool.

3. Reliability refers to consistency. A test which shows a consistent result in its frequent uses in different situations and places is called reliability of the test.

4. Rational equivalence method is the best method for editermining reliability of a test as statistifically most of the errors are minimised in this method.

5. (a) In case number of items in the test is less; (b) If items are not homogeneous; (c) If memory and carry-over effect works; (d) Error in scoring the test, and (e) In case items fail to discriminate the higher and lower group students.

6. Validity means accuracy and truthfulness of the test and which satisfies the very purpose of the test.

7. Predictive validity indicates the effectiveness of the test in forecasting or predicting future outcomes in a specific area.

8. Self exercise.

9. Item-objectivity means the item is having only one right answer and it does not carry ambiguous questions, lack of proper direction, double-barreled questions, questions with double negatives, broad essay type questions, etc.

10. Comprehensibility, easy for administration, easy for scoring, easy for interpretation, appearance of the test, economy and availability of the test.