ASSESSMENT FOR LEARNING

UNIT 6: CHARACTERISTICS OF INSTRUMENTS OF EVALUATION

Validity - different methods of finding validity – Reliability - different methods of finding reliability – Objectivity – Interdependence of validity, reliability and objectivity – Usability – Norms.

6.1 VALIDITY

Validity is a measure what it intends to measure. Validity refers to the extent to which the results of an evaluation procedure serve the particular uses for which they are intended. (Gronlund, 1981)

A tool is valid if it serves the purpose for which it is designed. The Validity of a measure is how well it fulfills the function for which it is being used, that is, the degree to which it is capable of achieving certain aims.

The concept of validity of a test is chiefly a concern for the basic honesty of the test, honesty in the sense of doing what one promises to do. It is a basic concern for the relationship between the purpose set to achieve and the efforts taken, the means employed and what these efforts and means really achieve.

Nature of Validity in Evaluation: (i) Validity is a matter of degree. It does not exist on an all-or-none basis. An instrument designed for measuring a particular ability cannot be said to be either perfectly valid or not valid at all. It is generally more or less valid. (ii) Validity is a relative term. A tool is valid for a particular purpose or in a particular situation; it is not generally valid for all situations. (iii) Validity refers to the results of a test or evaluation tool for a given group of individuals, not the tool itself.

6.1.1 DIFFERENT METHODS OF FINDING VALIDITY (TYPES OF VALIDITY)

There are five different methods (types) of measuring Validity. They are described below: (i) Content Validity, (ii) Criterion - related Validity-(a) Concurrent Validity and (b) Predictive Validity, (iii) Construct Validity, (iv) Face Validity and (v) Factorial Validity.

(i) Content Validity: Content validity is a process of matching the test items with the instructional objectives. Content validity is the most important criterion for the usefulness of a test, especially of an achievement test. It is a measure of the match between the content of a test and the content of 'teaching' that preceded it. The measure is represented subjectively after a careful process of inspection comparing the content of the test with the objectives of the course of instruction. In other words, the teacher has to match his/her test items with the content. The teacher has to check whether all the specific instructional objectives are represented in the test. This way, content validity refers to the extent to which a test contains items representing the behaviour that we are going to measure. In order to find out content validity, it is convenient to prepare a two-way table of content and objectives as in the Specification Table (Blue Print of an Achievement Test)

CONTENT	KNOWLEDGE	UNDERSTANDING	SKILL	TOTAL
UNIT – 1	8	4	5	17
UNIT – 2	12	6	4	22
UNIT – 3	10	8	6	24
UNIT -4	8	4	5	17
UNIT – 5	12	3	5	20
TOTAL	50	25	25	100

The Table reflects the sample of learning tasks to be measured. The closer the test items correspond to the specified sample, the greater the possibility of having satisfactory Content Validity. It is desirable that the items in a test are screened by a team of experts; they should check whether the placement of the various items in the cells of the Table is appropriate and whether all the cells of the Table have an adequate number of items. The adequacy is to be judged in terms of the weight age given to the different content-by-objective Table according to the team of experts who have designed the curriculum.

(ii) Criterion - related Validity: Criterion - related Validity refers to the extent to which test performance is related to some other valued measure of performance. Unlike the 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 as its yardstick of measurement. The 'external' criterion may be data of 'concurrent' information or of a future performance.

Two of its aspects namely, (a) Concurrent Validity and (b) Predictive Validity are explained below:

(a) Concurrent Validity: Concurrent Validity of a test is correlating the test scores with another set of criterion scores. The 'Concurrent' criterion is provided by a data-base of learner performance obtained on a test whose validity has been pre-established. The term 'concurrent' here implies the following characteristics: (i) The two tests - the one whose validity is being examined and the one with proven validity -are supposed to cover the same content area at a given level and the same objectives; (ii) The population for both the tests remains the same and the two tests are administered in almost similar environments; and (iii) The performance data on both the tests are obtainable almost simultaneously.

(b) Predictive Validity: Predictive Validity of a test is the extent to which test predicts the future performance of students. The 'Predictive' criterion is provided by the performance data of a group obtained on a course or career subsequent to the test which is administered to the group and whose validity is under scrutiny.

Validity established on correlation with 'concurrent' criterion yields concurrent validity and similarly, validity established against the scale of 'predictive' criterion is called 'predictive' validity. The former resolves the validity of tests serving the purpose of measuring proficiency, the latter resolves the validity of tests meant for predictive function. In all cases of criterion-

related validity is an index of the degree of correspondence between the tests being examined can be obtained. This index of agreement is known as Validity coefficient.

(c) Construct Validity: Construct Validity is the extent test results are interpreted in terms of known psychological concepts and principles. The word 'Construct' refers to the psychological quality that we assume, exists in order to explain some aspects of behaviour. Construct Validity is defined as the extent to which test performance can be interpreted in terms of certain psychological constructs. Usually Factor Analysis is done to determine the construct validity.

(d) Face Validity: Face Validity is the extent the test appears to measure what is to be measured. Face validity refers not to what the test measures, but what the test 'appears to measure'. Face validity is generally determined when a test is to be constructed quickly and there is no time or scope to determine the validity by other methods or when there is an urgent need of a test.

(e) Factorial Validity: Factorial Validity is the extent correlation of the different factors with the whole test. It is determined by a statistical technique known as Factor Analysis. It uses methods of explanation of inter-correlations to identify factors (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 tells us about the Factor Loadings. This relationship of the different factors with the whole test is called the Factorial Validity.

6.1.1.1 FACTORS AFFECTING VALIDITY

A large number of factors influence the validity of an evaluation tool. Gronlund (1981) has suggested the following factors that affect the Validity of a tool:

(i) Factors in the Test itself: Each test contains items and a close scrutiny of test items will indicate whether the test appears to measure the subject matter content and the mental functions that the teacher wishes to test. The following factors in the test itself can prevent the test items from functioning as desired and thereby lower the validity.

(a) 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.

(b) Reading vocabulary and sentence structures which are too difficult: The complicated vocabulary and sentence structure meant for the students appearing the test may fail in measuring the aspects of student performances; thus lowering the validity.

(c) 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.

(d) Poorly constructed test items: The test items which provide unintentional clues to the answer will tend to measure the Students' alertness in detecting clues as well as the aspects of pupil performance which ultimately affect the validity.

(e) Ambiguity: Ambiguity in statements in the test items leads to misinterpretation, differing interpretations and confusion. It may confuse the better students more than the poorer ones

resulting in the discrimination of items in a negative direction. As a consequence, the validity of the test is lowered.

(ii) Test items inappropriate for the outcomes being measured: The researcher tries to measure certain complex types of achievement, understanding, thinking and skills with test forms that are appropriate only for measuring factual knowledge. This affects the results and leads to a distortion of the validity.

(iii) Test too short: If the test is too short to become a representative one, then validity will be affected accordingly.

(iv) Improper arrangement of items: Items in the test are generally arranged in order of difficulty with the easiest items fist. 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.

(v) Identifiable pattern of answers: When the students identify the systematic pattern of correct answer (e.g. TTFF or ABCD), they can cleverly guess the answers and this will affect the validity.

(vi) Functioning Content and Teaching Procedure: 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 instrument for measuring the more complex mental processes and they thus, affect the validity.

(vii) Factors in Test Administration and Scoring: The test administration and scoring procedure may also affect the validity of the interpretation from the results.

(viii) Factors in Pupils' Response (Emotion, Motivation and Test Situation and Response set): The emotionally disturbed students, lack of students' motivation and students' being afraid of test situation may not respond normally and this may ultimately affect the validity. Response set also influences the test results. It is the test taking habit which affects the pupil's score. A response set is a consistent tendency to follow a certain pattern in responding to test items.

(ix) Nature of the Group and the Criterion: The Validity is always specific to a particular group. There are certain factors like age, sex, ability level, educational background and cultural background which influence the test measures. Therefore, the nature of the validation group should find a mention in the test manuals. The nature of the criterion used is another important consideration while evaluating validity coefficient.

6.2 RELIABILITY

Reliability is the trust worthiness of the test. The concept of reliability relates to the question of 'accuracy' with which the 'what' is measured. A test is said to be reliable to the extent the scores obtained through it are consistent over time and over different samples of the test items.

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.

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).

Reliability is a necessary but not a sufficient condition for Validity. Low reliability can restrict the degree of validity that is obtained, but high reliability provides no assurance for a satisfactory degree of validity.

Reliability is primarily statistical in nature in the sense that the scores obtained on two successive occasions are correlated with each other. This coefficient of correlation is known as self-correlation and its value is called the Reliability Coefficient. It may be expressed in terms of the shifts in relative standing of persons in the group or in terms of the amount of variation to be expected in a specific individual's score. In the former case it is reported through correlation coefficient called a Reliability Coefficient and in the latter case it is reported by means of the Standard Error of Measurement.

6.2.1 DIFFERENT METHODS OF FINDING RELIABILITY

There are common approaches to estimate the Reliability. They are Test-Retest, Equivalent Form and Internal Consistency.

(i) Test - Retest Reliability: Checking the test reliability by giving the test again.

To estimate reliability by means of the test-retest method, the same test is administered twice to the same group of pupils with a given time interval between the two administrations of the test. The resulting test scores are correlated and this correlation coefficient provides a measure of stability, that is, it indicates how stable the test results are over a given period of time. So it is otherwise known as a measure of stability. The estimate of reliability in this case will vary according to the length of time-interval allowed between the two administrations.

(ii) Equivalent - Forms of Reliability: Checking test reliability through two equivalent forms of the same test.

Estimating reliability by means of the equivalent form method involves the use of two different but equivalent forms of the test (also called parallel or alternate forms). The two forms of the test are equivalent so far as the content, objectives, format, difficulty level and Discriminating value of items and length of the test are concerned. Equivalent tests have equal inter-correlations among items. That is, two equivalent forms must be homogeneous in all respects, but not a duplication of test items. The two forms of test are administered to the same group of pupils in close succession and the resulting test scores are correlated. This correlation coefficient provides a measure of equivalence.

(iii) Internal Consistency: Internal consistency scores are obtained through a single administration of the test. The two types of measures of Internal Consistency are discussed below: (a) Split-half method: Checking test reliability by splitting a test into two halves.

Since a test consists of many questions, all the questions in a test together try to measure learning related to a particular aspect. When the test is divided / split into two halves, they would represent two equivalent forms, and each of them would still measure the same aspect. It is a general expectation that the total scores of students on each half should be consistent. Here, the reliability to the test is estimated in terms of the consistency of the scores over the two halves of the test. The usual procedure to split the test into halves that are most equivalent is to score the even-numbered items and the odd-numbered items separately. This provides two scores for each pupil, the correlation between which provides a measure of internal consistency. This gives the reliability estimate of half the length of the test. To estimate the reliability of the scores on the full length test, the following formula is used: Reliability on full test = 2 x Reliability on first half of the test / 1 + Reliability on second half of the test.

(b) Kuder -Richardson estimates: Kuder and Richardson developed two formulae KR-20 and KR-21 which could ***avoid the question of how to split a test into halves***. These formulae provide the means of all possible split-half estimates of reliability of a test. R = 2r/1+r.

6.2.1.1 FACTORS AFFECTING RELIABILITY

Factors affecting the Reliability test scores are Extrinsic and Intrinsic.

(i) Extrinsic Factors: Certain factors remain outside the test itself influence the Reliability test scores. They are described as follows:

(a) 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.

(b) Guessing and Chance Errors: Guessing in test gives rise to increased error variance and as such reduces reliability.

(c) Environmental conditions: As far as practicable, testing environment should be uniform. Arrangement should be such that light, sound, and other comforts should be equal to all tests; otherwise it will affect the reliability of the test scores.

(d) Momentary fluctuations: Momentary fluctuations may raise or lower the reliability of the test scores.

(ii) Intrinsic Factors: Certain factors lie within the test itself influences the Reliability test scores. They are described as follows:

(a) 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

(b) Homogeneity of 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 correlations of items are 'zero' or near to it, then the reliability is 'zero' or very low and vice-versa.

(c) Difficulty value of items: Broadly, items having indices of difficulty at 0.5 or close to it yield higher reliability than items of extreme indices of difficulty.

(d) Discriminative value: When items can discriminate well between superior and inferior, the item-total correlation is high and then the reliability is also likely to be high and vice-versa.

(e) 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 scorers agree in scoring the same set of response. The reliability is likely to be lowered if they do not agree.

6.3 OBJECTIVITY

Objectivity is the consistency in scoring. A tool is objective if it gives the same score even when different scorers score the item. Objectivity in scoring may be considered as consistency in scoring by different scorers.

6.3.1 FACTORS AFFECTING OBJECTIVITY

The following factors affect Objectivity of a test:

(i) Teachers' influence: Teachers may differ to some order in their opinion about the correctness of the answer and therefore, scoring is not likely to be objective. When the scoring is based on students' answers, not on teachers' opinions, then it will be more objective.

(ii) Nature of questions: Objectivity is also based on the nature of questions. When the question is stated differently, difference in scoring will occur. For example: Explain the concept of guidance in about fifty words? Here the scores given by the teachers will not vary to a large extent because the rewarded question clearly indicates the nature of the correct answer that is expected.

Consistency among different scorers: In the objective type test (having multiple choice items the chances of giving different scores to the same student by the different examiners is very little and thus objectivity is more. / In a good objective-type item, answers are specific and have one and only one correct response answer. Therefore, an objective type test is more reliable, since reliability tends to mean consistency of scores. Objectivity now can be viewed as consistency among different scorers in giving scores to answers on a test. Hence, objectivity is considered as inter-scorer reliability.

Subjectivity in scoring: In actual situations, examiner's prejudices influence marking. The evaluation mode Questions, asked in certain topics for which the examiner has an inclination may fetch more marks than the other questions. This type of irrational temperament towards scoring system is a kind of his/her subjective treatment of the syllabus which, in turn, affects the evaluation process. Therefore, objectivity in evaluation is to be ensured for accurate evaluation. At the same time, subjectivity need not be condemned and entirely excluded, as that is how most evaluations in reality are made. Subjective assessment based on careful observation, unprejudiced and unbiased thinking and logical analysis of situations and phenomena may also give accurate evaluation.

6.4 INTERDEPENDENCE OF VALIDITY, RELIABILITY AND OBJECTIVITY

There is a close relationship between the Validity and Reliability. They are the two dimensions of the test efficiency. Reliability is concerned with the stability of test scores-self correlation of the test. Validity is the correlation of the test with some outside independent criteria. A test which possesses poor reliability is not expected to yield high validity.

Reliability is a prerequisite of validity. To be valid a test must be reliable. A highly reliable tool is always a valid measure of some function. Thus, reliability controls validity.

A test maybe theoretically valid, but may be practically invalid as judged against its correlation with different independent criteria. Let us consider a highly valid and reliable Achievement test. When it is used for diagnostic purposes it would be invalid. Thus, a highly reliable test may be highly valid for one purpose and may be invalid for other purposes.

Objective judgements are accurate and hence tend to the reliable. Objectivity is a pre – requisite of reliability and therefore of validity.

6.5 USABILITY

Usability - degree to which the tool of evaluation can be successfully used by the educational practitioner.

The important characteristic of a tool is its usability or practicability. While selecting evaluation tools one must look for certain practical considerations like ***ease of administration and scoring, ease of interpretation, availability of comparable forms and cost of testing.***

(A)Ease of Administrability:

- Definite provision should be made for the preparation, distribution and collection of test materials.
- Instruction to the pupils should be simple, clear and concise.
- Sample items should be illustrated by practice exercises.
- The test format should be such that pupils will have no difficulty in reading the items, in recording their answers, in moving from one page to the next, etc.

(B)Ease of Scoring

- The results of a test possessing scorability should he obtainable in as simple, rapid and routine a manner as in proportion to their importance.
- The test should be subjected to accurate scoring even by persons not conversant with their content.
- No algebraic manipulations should be required to get the s.-ores.

(C)Ease of Interpretation

- The raw scores of a test should be easily converted into meaningful derived scores.
- It should be feasible to interpret the results with the competence of classroom teachers. No specially trained personnel should be required in order that the results may be interpreted validly.

(D) Economy

- The economy of a testing programme should be computed in terms of the validity of the tests per unit of cost.
- Economy refers to the cost as well as the time required for administering and scoring a test. Any test of a duration which does not exceed a period of 45 minutes is preferred by teachers.

6.6 NORMS

Norms are the scores earned by pupils in clearly-defined reference groups. These groups are known as norm groups and are generally representative of the pupil population. There are different types of norms, such as age norms, grade norms, percentiles, standard scores and quotients. Norms are not standards; they go on changing form population to population.

Age Norms: Age norms indicate the average performance of a particular age group.

Grade Norms: Grade scores compare individuals' performance with that of the average student in various grades or classes. The procedure of item selection, test construction and scoring is similar to that we follow for age scale. The only difference is that in grade scores we use grade levels in the place of age levels.

Percentile Norms: Percentile rank of a score is defined as the number of persons in a group who obtain lower scores. Percentile rank is an individual's rank in a norm group expressed in terms of percentage of persons.

Standard Scores: A standard score is defined as the deviation of a raw score from the mean, expressed in standard deviation units.