**This article throws light upon the four important characteristics of a good test. The four characteristics are: 1. Reliability 2. Validity 3. Objectivity 4. Usability.**

**Characteristic # 1. Reliability:**

The dictionary meaning of reliability is consistency, depend­ence or trust. So in measurement reliability is the consistency with which a test yields the same result in measuring whatever it does measure. A test score is called reliable when we have reason for believing the score to be stable and trust-worthy. Stability and trust-worthiness depend upon the degree to which the score is an index of time-reliability’ is free from chance error. Therefore reliability can be defined as the degree of consistency between two measurements of the same thing.

For example we administered an achievement test on Group-A and found a mean score of 55. Again after 3 days we ad­ministered the same test on Group-A and found a mean score of 55. It indicates that the measuring instrument (Achievement test) is providing a stable or dependable result. On the other hand if in the second measurement the test provides a mean score around 77 then we can say that the test scores are not consistent.

In the words of Gronlund and Linn (1995) **“reliability refers to the consistency of measurement—that is, how consistent test scores or other evaluation results are from one measurement to other.”**

C.V. Good (1973) has defined reliability as the “wor­thiness with which a measuring device measures something; the degree to which a test or other instrument of evaluation measures consistently whatever it does in fact measure.”

According to Ebel and Frisbie (1991) “the term reliability means the consistency with which a set of test scores measure whatever they do measure.”

Theoretically, reliability is defined as the ratio of the true score and observed score variance.

According to Davis (1946) “the degree of relative precisions of measurement of a set of test score is defined as reliability.”

**Thus reliability answers to the following questions:**

Gronlund and Linn (1995)

How similar the test scores are if the lost is administered twice?

How similar the test scores are if two equivalent forms of tests are administered?

To what extent the scores of any essay test. Differ when it is scored by different teachers?

It is not always possible to obtain perfectly consistent results. Because there are several factors like physical health, memory, guessing, fatigue, forgetting etc. which may affect the results from one measurement to other. These extraneous variables may introduce some error to our test scores. This error is called as measurement errors. So while determining reliability of a test we must take into consideration the amount of error present in measurement.

**Nature of Reliability:**

1. Reliability refers to consistency of the results obtained with an instrument but not the instrument itself

2. Reliability refers to a particular interpretation of test scores. For example a test score which is reliable over a period of time may not be reliable from one test to another equivalent test. So that reliability cannot be treated as general characteristics.

3. Reliability is a statistical concept to determine reliability we administer a test to a group once or more than once. Then the consistency is determined in terms of shifts in the relative position of a person in the group or amount of variation expected in an individual’s score. Shifting of relative position of an individual is related by means of a coefficient of correlation called ‘Reliability Coefficient’ and the amount of variation is reported by ‘Standard error of measurement’. Both these processes are statistical.

4. Reliability is necessary but not a sufficient condition for validity. A test which is not reliable cannot be valid. But it is not that a test with high reliability will possess high validity. Because a highly consistent test may measure something other than that what we intend to measure.

**Methods of Determining Reliability:**

For most educational tests the reliability coefficient provides the most revealing statistical index of quality that is ordinarily available. Estimates of the reliability of test provide essential information for judging their technical quality and motivating efforts to improve them. The consistency of a test score is expressed either in terms of shifts of an individual’s relative position in the group or in terms of amount of variation in an individual’s score.

**On the basis of this estimation of reliability fall in to two general clas­sifications**:

**(i)Relative Reliability or Reliability Coefficient:**

In this method the reliability is stated in terms of a coefficient of cor­relation known as reliability coefficient. Hence we determine the shifting of relative position of an individual’s score by coefficient of correlation.

**(ii) Absolute Reliability or Standard error of Measure­ment:**

In this method, the reliability is stated in terms of the standard error of measurement. It indicates the amount of varia­tion of an individual’s score.

**Methods of Determining Relative Reliability or Reliability Coefficient:**

In order to determine reliability coefficient we have to obtain two sets of measurements in identical condition and then compare the two sets. But it is only a theoretical condition, because it is impossible on our part to get two measurements on exactly two identical conditions. So that several methods have been developed to determine the relative reliability.

**They are as follows (Gronlund and Linn—1995):**

(i) The same form of test may be administered twice to the same group of individuals.

(ii)Two separate but equivalent forms of the test may be administered to the same individuals.

(iii) The test items of a single test are divided into two separate sets and the scores of two sets are correlated.

The methods are similar in that all of them involve correlating two sets of data, obtained either from the same evaluation in­strument or from equivalent forms of the same procedure. This reliability coefficient must be interpreted in terms of the types of consistency being investigated.

**Different types of consistency are determined by different methods. These are as follows:**

1. Consistency over a period of time.

2. Consistency over different forms of instrument.

3. Consistency within the instrument itself

**There are four methods of determining reliability coefficient, such as:**

(a) Test-Retest method.

(b) Equivalent forms/Parallel forms method.

(c) Split-half method.

(d)Rational Equivalence/Kuder-Richardson method.

**(а) Test-Retest Method:**

This is the simplest method of determining the test reliability. To determine reliability in this method the test is given and repeated on same group. Then the correlation between the first set of scores and second set of scores is obtained.

A high coefficient of correlation indicates high stability of test scores. In the words of Gronlund, Measures of stability in the .80’s and .90’s are com­monly reported for standardized tests over occasions within the same year. But this method suffers from some serious drawbacks. First of all what should be the interval between two administrations.

If it is administered within a short interval say a day or two, then the pupil will recall their first answers and spend their time on new material. It will tend to increase their score in second administrations. If interval is too long say one year, then the maturation effect will affect the retest scores and it will tend to increase the retest scores.

In both the cases it will tend to lower the reliability. So what should be the time gap between two administrations depends largely on the use and interpretation of test scores. Due to its difficulties in controlling conditions which influence the scores of retest, reduces the use of test-retest method in estimating, reliability coefficient.

**(b) Equivalent Forms/Parallel Forms Method:**

Reliability of test scores can be estimated by equivalent forms method. It is also otherwise known as Alternate forms or parallel forms method. When two equivalent forms of tests can be con­structed the correlation between the two may be taken as measures of the self correlation of the test. In this process two parallel forms of tests are administered to the same group of pupils in short interval of time, then the scores of both the tests are cor­related. This correlation provides the index of equivalence. Usually in case of standardized psychological and achievement tests the equivalent forms are available.

Both the tests selected for administration should be parallel in terms of content, difficulty, format and length. When time gap between the administrations of two forms of tests are provided the coefficient of test scores provide a measure of reliability and equivalence. But the major drawback with this method is to get two parallel forms of tests. When the tests are not exactly equal in terms of content, difficulty, length and comparison between the scores ob­tained from these tests may lead to erroneous decisions.

**(c) Split-Half Method:**

There are also methods by which reliability can be determined by a single administration of a single test. One of such method is split-half method. In this method a test is administered to a group of pupils in usual manner. Then the test is divided into two equivalent values and correlation for these half-tests are found.

The common procedure of splitting the test is to take all odd numbered items i.e. 1, 3, 5, etc. in one half and all even-numbered items i.e. 2, 4, 6, 8 etc. in the other half Then scores of both the halves are correlated by using the Spearman- Brown formula.

For example by correlating both the halves we found a coef­ficient of .70.

**By using formula (5.1) we can get the reliability coefficient on full test as:**

The reliability coefficient .82 when the coefficient of correlation between half test is .70. It indicates to what extent the sample of test items are dependable sample of the content being measured—internal consistency.

Gronlund (1995) is of the opinion that **“split half reliabilities tend to be higher than equivalent form reliabilities because the split half method is based on the administration of a single test form.”** This method over-comes the problem of equivalent forms method introduced due to differences from form to form, in attention, speed of work, effort, fatigue and test content etc.

**(d) Rational Equivalent/Kuder Richardson Method:**

Rational equivalence is another method of determining reliability by using the formula developed by Kuder and Richardson. Like split-half method this method also provides a measure of internal consistency. It neither requires administration of two equivalent forms of tests nor it requires to split the tests into two equal halves. Reliability coefficient is determined by using the Kuder-Richardson formula-20 which reads like this.



This method provides information about the degree to which the items in the test measure similar characteristics. Though the simplicity of applying this method has made it wide spread still it has some limitations.

**1.** Kuder-Richardson method and split-half method are not appropriate for speed tests.

**2.** Both Kuder-Richardson and split half method do not measure the consistency of pupil response from day to day.

**3.** Kuder-Richardson method is cumbersome to calculate unless information is already available concerning the proportion of passing.

**Methods of Determining Absolute Reliability or Standard Errors of Measurement:**

If we shall administer a test again and again we shall found some variation in scores. Because the obtained score is an index of the examinee’s true score plus: errors of measurement. H.E. Garrett (1985) has defined a true score as **“a measure which would be obtained by taking the mean of an infinity large number of measurement of a given individual on similar tests under similar conditions. A true score cannot, of course, be deter­mined experimentally”**.

If the test scores include a large component of error its reliability is low and if it includes a little of errors its reliability is high. Thus the extent to which a true score exceeds, error in the obtained scores can be indicated by reliability coefficient.

**This relationship between true score, obtained scores and the error can be expressed mathematically as follows:**



We can find out standard error of measurement (SE) when the reliability coefficient and standard deviation of the distribution is given.

**The formula (Garrett—1985) to calculate standard error of measurement is as follows:**



For example in a group of 200 High School students, reliability coefficient of an achievement test in mathematics is .70, Mean = 65 and o = 20. Lipu achieves a score of 60. What is the S.E. of this score.

**By putting the value in formula (5.3):**



So the true score of Lipu is 60 ± 10.95 i.e. 70.50 to 49.05.

No obtained score tells us what the true score is, but the knowledge of the SE indicates the difference between the obtained score and true score. When the SE is small, it indicates that the true score is closer to the obtained score and it also indicates whether the difference between scores of two individuals is real difference or difference due to errors of measurement.

**Factors Affecting Reliability:**

There are a number of factors which affect the measures of reliability. So that when we interpret and use the scores we must be cautious and manipulate those factors through test preparation and administration.

**The major factors which affect the reliability of test, scores can be categorized in to three headings:**

1. Factors related to test.

2. Factors related to testee.

3. Factors related to testing procedure.

**1. Factors related to test:**

**(а) Length of the test:**

SpearmanBrown formula in­dicates the longer the test is, the higher the reliability will be. Because a longer test will provide adequate sample of the behaviour. Another cause is that guessing factor is apt to be neutralized in a longer test.

For example if we shall give one computation to measure the numerical ability of the students. Those who have computed cor­rectly are perfect in numerical ability those who failed are complete failures. If the computation is a difficult one, then most of the students will fail. If it is easy, then most of the students will compute it correctly. So that single item score never gives a reliable result.

**(b) Content of the test:**

According to Guilford homogeneity of test content also increases the reliability of test scores. A test of 50 items on Vedic Civilization will provide more reliable scores than a test of 50 items on Indian history. According to Ebel (1991) **“subject matter in some courses, such as mathematics and foreign language, is more tightly organized, with greater inter­dependence of facts, principles abilities and achievements, than in the subject matter literature or history.”** So this content homogeneity is also a factor which results is high reliability.

**(c) Characteristics of items:**

The difficulty level and clarity of expression of a test item also affect the reliability of test scores. If test items are too easy or difficult for the group members it will tend to produce scores of low reliability. Because both the tests have a restricted spread of scores.

**(d) Spread of Scores:**

According to Gronlund and Minn (1995) “other things being equal, larger the spread of scores is the higher the estimate of reliability will be.” When t he spread of scores are large there is greater chance of an individual to stay in the same relative position in a group from one testing to another. We can say that errors of measurement affect less to the relative position of the individual when the spread of scores are large.

For example in Group A students have secured marks ranging from 30 to 80 and in Group B student have secured marks ranging from 65 to 75. If we shall administer the tests second time in Group A the test scores of individuals could vary by several points, with very little shifting in the relative position of the group mem­bers. It is because the spread of scores in Group A is large.

On the other hand the scores in Group B are more likely to shift positions on a second administration of the test. As the spread of scores is just 10 points from highest score to lowest score, so change of few points can bring radical shifts in relative position of individuals. Thus greater the spread more is the reliability.

**2. Factors related to testee:**

Variability in achievement, test wiseness of the individuals, and motivation of the students also influence the reliability of test scores.

**Following are some of the important factors with the testee which affect the test reliability:**

**(a) Heterogeneity of the group:**

When the group is a homogeneous group the spread of the test scores is likely to be less and when the group tested is a heterogeneous group the spread of scores is likely to be more. Therefore reliability coefficient for a heterogeneous group will be more than homogeneous group.

**(b) Test wiseness of the students:**

Experience of test taking also affect the reliability of test scores. Practice of the students in taking sophisticated tests increases the test reliability. But when in a group all the students do not have same level of test wiseness, it leads to greater measurement errors.

**(c) Motivation of the students:**

When the students are not motivated to take the test, they will not represent their best achievement. This depresses the test scores.

**3. Factors related to testing procedure:**

As the test related factors and testee related factors affect the reliability of test scores, so also the factors related to testing procedure also affect the test scores. If the test users can control these factors, then they can increase the consistency of the test scores.

**(a) Time Limit of test:**

According to Ebel and Frisbie (1991) “scores from a test given under highly speeded conditions will ordinarily show a higher internal consistency reliability coefficient than would be obtained for scores from the same test given to the same group under more generous time limits.” Thus when the students get more time to take the test they can make more guessing, which may increase the test scores. Therefore by speeding up a test we can increase the test reliability.

**(b) Cheating opportunity given to the students:**

Cheat­ing by the students during the test administration leads to meas­urement errors. Some students may provide correct answer by copying it from cheat sheets or listening from other students without knowing the correct answer. This will cause a higher score of those students than they actually deserve. This will make the observed score of cheaters higher than their true score.

**How Higher Should Reliability be?**

Obviously evaluation devices are never perfectly reliable. How unreliable a test may be and still useful depends mainly on the fineness of discrimination desired from the test scores. (Rem-mers. 1967) The degree of reliability coefficient depends upon the nature of the test, the size and variability of the group, the purpose for which the test was administered and the method used for the estimation of reliability. A test with low reliability may have higher validity and so may be used. But in the words of Remmers (1967) ‘Most standardized test published for school use have reliability coefficients of at least .80 in the population for which they are designed.

When one is choosing a standardized test on interpreting its results, it is not sufficient to just look at the numerical value of reliability estimate, one must also take into account how that estimate was obtained. Gronlund (1976) has remarked about the significance of methods of estimating reliability.

According to him “the split-half method gives the largest numerical values to the reliability coefficient. Equivalent forms method and test retest tend to give lower numerical value to the reliability coefficient. Typically these two methods provide medium to large reliability coefficient. Equivalent forms method typically provides smallest reliability coefficient for a given test.”

Therefore it may be said that the teacher should seek a standardized test whose reliability is as high as possible. But he must interpret this reliability coefficient in the light of the groups of pupils on which it is based, the variability of this group and methods of estimating reliability.

**Characteristic # 2. Validity:**

**“In selecting or constructing an evaluation instrument, the most important question is; To what extent will the results serve the particular uses for which they are intended ? This is the essence of validity.”** —GRONLUND

Validity is the most important characteristic of an evaluation programme, for unless a test is valid it serves no useful function. Psychologists, educators, guidance counselors use test results for a variety of purposes. Obviously, no purpose can be fulfilled, even partially, if the tests do not have a sufficiently high degree of validity. Validity means truth-fullness of a test. It means to what extent the test measures that, what the test maker intends to measure.

**It includes two aspects:**

What is measured and how consistently it is measured. It is not a test characteristic, but it refers to the meaning of the test scores and the ways we use the scores to make decisions. Following definitions given by experts will give a clear picture of validity.

Gronlund and Linn (1995)—”Validity refers to the ap­propriateness of the interpretation made from test scores and other evaluation results with regard to a particular use.”

Ebel and Frisbie (1991)—”The term validity, when applied to a set of test scores, refers to the consistency (accuracy) with which the scores measure a particular cognitive ability of interest.”

C.V. Good (1973)—In the dictionary of education defines validity as the “extent to which a test or other measuring instru­ment fulfils the purpose for which it is used.”

Anne Anastasi (1969) writes “the validity of a test concerns what the test measures and how well it does so.”

According to Davis (1964) validity is the extent of which the rank order of the scores of examinees for whom a test is appropriate is the same as the rank order of the same examinees in the property or characteristic that the test is being used to measure. This property or characteristic is called the criterion. Since any test may be used for many different purposes, it follows that it may have many validities one corresponding to each criterion.”

Freeman (1962) defines, “an index of validity shows the degree to which a test measures what it purports to measure, when compared with accepted criteria.”

Lindquist (1942) has said, “validity of a test may be defined as the accuracy with which it measures that which it is intended to measure, or as the degree to which it approaches infallibility in measuring what it purports to measure.”

From the above definitions it is clear that validity of an evaluation device is the degree to which it measures what it is intended to measure. Validity is always concerned with the specific use of the results and the soundness of our proposed interpretation.

It is not also necessary that a test which is reliable may also be valid. For example suppose a clock is set forward ten minutes. If the clock is a good time piece, the time it tells us will be reliable. Because it gives a constant result. But it will not be valid as judged by ‘Standard time’. This indicates “the concept that reliability is a necessary but not a sufficient condition for validity.”

**Nature of Validity:**

1. Validity refers to the appropriateness of the test results but not to the instrument itself.

2. Validity does not exist on an all-or-none basis but it is a matter of degree.

3. Tests are not valid for all purposes. Validity is always specific to particular interpretation. For example the results of a vocabulary test may be highly valid to test vocabulary but may not be that much valid to test composition ability of the student.

4. Validity is not of different types. It is a unitary concept. It is based on various types of evidence.

**Factors Affecting Validity:**

Like reliability there are also several factors which affect the validity of test scores. There are some factors about which we are alert and can avoid easily. But there are some factors about which we are ignorant and it makes the test results invalid, for their intended use.

**Some of these factors are as following:**

**1. Factors in the test:**

(i) Unclear directions to the students to respond the test.

(ii) Difficulty of the reading vocabulary and sentence structure.

(iii) Too easy or too difficult test items.

(iv) Ambiguous statements in the test items.

(v) Inappropriate test items for measuring a particular outcome.

(vi)Inadequate time provided to take the test.

(vii) Length of the test is too short.

(viii) Test items not arranged in order of difficulty.

(ix) Identifiable pattern of answers.

**Factors in Test Administration and Scoring:**

(i) Unfair aid to individual students, who ask for help,

(ii) Cheating by the pupils during testing.

(iii) Unreliable scoring of essay type answers.

(iv) Insufficient time to complete the test.

(v) Adverse physical and psychological condition at the time of testing.

**Factors related to Testee:**

(i) Test anxiety of the students.

(ii) Physical and Psychological state of the pupil,

(iii) Response set—a consistent tendency to follow a certain pattern in responding the items.

**Characteristic # 3. Objectivity:**

Objectivity is an important characteristic of a good test. It affects both validity and reliability of test scores. Objectivity of a measuring instrument moans the degree to which different per­sons scoring the answer receipt arrives of at the same result. C.V. Good (1973) defines objectivity in testing is “the extent to which the instrument is free from personal error (personal bias), that is subjectivity on the part of the scorer”.

Gronlund and Linn (1995) states “Objectivity of a test refers to the degree to which equally competent scores obtain the same results. So a test is considered objective when it makes for the elimination of the scorer’s personal opinion and bias judgement. In this con­text there are two aspects of objectivity which should be kept in mind while constructing a test.”

(i) Objectivity in scoring.

(ii) Objectivity in interpretation of test items by the testee.

**(i) Objectivity of Scoring:**

Objectivity of scoring means same person or different persons scoring the test at any time arrives at the same result without may chance error. A test to be objective must necessarily so worded that only correct answer can be given to it. In other words the personal judgement of the individual who score the answer script should not be a factor affecting the test scores. So that the result of a test can be obtained in a simple and precise manner if the scoring procedure is objective. The scoring procedure should be such that there should be no doubt as to whether an item is right or wrong or partly right or partly wrong.

**(ii) Objectivity of Test Items:**

By item objectivity we mean that the item must call for a definite single answer. Well-con­structed test items should lead themselves to one and only one interpretation by students who know the material involved. It means the test items should be free from ambiguity. A given test item should mean the same thing to all the students that the test maker intends to ask. Dual meaning sentences, items having more than one correct answer should not be included in the test as it makes the test subjective.

**Characteristic # 4. Usability:**

Usability is another important characteristic of measuring instruments. Because practical considerations of the evaluation instruments cannot be neglected. The test must have practical value from time, economy, and administration point of view. This may be termed as usability.

**So while constructing or selecting a test the following practical aspects must be taken into account:**

**(i) Ease of Administration:**

It means the test should be easy to administer so that the general class-room teachers can use it. Therefore simple and clear directions should be given. The test should posses very few subtests. The timing of the test should not be too difficult.

**(ii) Time required for administration:**

Appropriate time limit to take the test should be provided. If in order to provide ample time to take the test we shall make the test shorter than the reliability of the test will be reduced. Gronlund and Linn (1995) are of the opinion that “Somewhere between 20 and 60 minutes of testing time for each individual score yielded by a published test is probably a fairly good guide”.

**(iii) Ease of Interpretation and Application:**

Another im­portant aspect of test scores are interpretation of test scores and application of test results. If the results are misinterpreted, it is harmful on the other hand if it is not applied, then it is useless.

**(iv) Availability of Equivalent Forms:**

Equivalent forms tests helps to verify the questionable test scores. It also helps to eliminate the factor of memory while retesting pupils on same domain of learning. Therefore equivalent forms of the same test in terms of content, level of difficulty and other characteristics should be available.

**(v) Cost of Testing:**

A test should be economical from preparation, administration and scoring point of view.