UNIT 7: RELIABILITY

UNIT STRUCTURE

7.1 Learning Objectives

7.2 Introduction

7.3 Meaning and definition of Reliability

7.4 Factors affecting Reliability of a test

7.5 Methods of estimating Reliability of a test

7.6 Relation between Reliability and Validity

7.7 Let Us Sum Up

7.8 Further Reading

7.9 Answer to Check Your Progress

7.10 Model Questions

7.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define Reliability of a test
- know about the different factors that affect Reliability of a test
- explain the process of estimating Reliability of a test
- understand the relation between Reliability and Validity.

7.2 INTRODUCTION

Reliability is one of the most important technical criteria of a good test. Any measuring tool used to collect data for measurement and evaluation in education or in any other field, must be reliable besides fulfilling other criteria such as validity, objectivity and the like. Reliability, like validity is a very important criterion of any measuring instrument. A good test or other measuring device must be reliable. If the tools used for data collection are erroneous, the data used will also be erroneous. In order for any data to be of value and of use, they must be reliable as well as valid. Here in this unit we will discuss the meaning and other aspects of reliability in a detailed way.

7.3 MEANING AND DEFINITION OF RELIABILITY

Reliability refers to consistency of results or findings. When an investigation is repeated it must yield same results if it were to be reliable. When more than one investigator or evaluator is making the measurement, all of them must arrive at similar findings in order to claim that the data are reliable. Similarly, when an individual takes the same test twice, if the scores on the two occasions are very similar, then the test can be described as reliable. This means, a reliable test must give the same result when given twice to the same person within a short period of time and when different evaluators make the measurement. On the contrary if a test or other measuring device gives different results on different occasions or when used by different evaluators on the same individuals, then it should be understood that the test is not reliable.

Reliability of a test refers to the consistency of scores obtained by students on alternate forms or different administration of the same test. However, no two tests with exact content can ever give identical results due to differences in the environmental conditions including psycho-physical state of the test-takers. Even the same test gives different results when administered to the same group of students on two different occasions. Nevertheless, whenever we administer two different tests with similar material or the same test on different occasions, we wish that the scores obtained by students in the two different administrations are similar. In this situation, the test is reliable to the extent to which the scores are comparable.

Reliability of a test is an indicator of validity of the test as well. If a test cannot give consistent results, it is obvious that the test is not accurately measuring the trait or variable of interest. A test to be valid must be reliable. Therefore, estimation of reliability is considered to be the first-step in the process of validating a test. If the test is not reliable, then there is no need of examining the validity and it can be easily concluded that the test is not valid. Several methods are used to estimate reliability of a test depending on the nature of the measurement.

In the views of Anastasi, "Reliability refers to the consistency of

scores obtained by the same individuals when re-examined with the same test on different occasions or with different sets of equivalent items or under variable examining conditions."

According to Freeman, "The term reliability refers to the extent to which it gives consistent result on testing and retesting."

Henry E. Garrett says, "The reliability of a test or of any measuring instrument depends upon the consistency with which it gauges the ability to whom it is applied."

According to Stodola & Stordahl, "Reliability is the correlation between two or more sets of scores on equivalent tests from the group of individuals."

In the views of Mehrens and Lehmann, "Reliability can be defined as the degree of consistency between two measures of the same thing."

After an analysis of the definitions mentioned above, it can said that reliability refers to the consistency of test scores. It represents the extent of accuracy or precision of a measuring tool. If a test or other tool gives consistent results on testing and retesting in variable conditions then the tool is accepted to be reliable. Reliability is a precondition to the validity of a test as well. However, reliability is the characteristic of the scores only and it is not the property of the test itself.

7.4 FACTORS AFFECTING RELIABILITY OF A TEST

While constructing or using a test, it is very important to understand that reliability of a test may be affected by various factors. Some of the important factors that affect reliability of a test may be discussed as under:

- **Length of the test:** Test reliability depends on the length of the test. According to Psychometricians, increasing the length of a test by adding items of acceptable quality and representing the content increases the reliability of the test. On the contrary if the length of test is reduced by deleting items then the reliability of the test will also be reduced.

- **Methods used to estimate reliability:** There are different methods that may be used to estimate reliability of tests. Different indices of reliability of the same test may be obtained when different methods to estimate reliability are used.
- **Nature of the group:** Reliability of a test also depends on the nature of the group upon which the test is administered to estimate reliability. The test is more reliable if the group of students on which the test is administered is more heterogeneous in ability; conversely, if the test is administered to a homogeneous group then the reliability of the test is low.
- **Difficulty level of items:** Difficulty level of test items may also affect reliability of a test. Too difficult or too easy items in a test reduce reliability of the test.
- **Objectivity of scoring:** reliability of a test is high when the scoring procedure is objective. Subjective scoring leads to variation in scores in the test resulting in lower consistency of scores and thus affect the reliability index of the test adversely.
- **Clarity of instruction:** When the instructions to candidates who take a test are clear and they understand what to do, the scores seem to be consistent on repeated testing. If the instructions are unclear then reliability of the test goes down.
- **Interval between testing and retesting:** In order to estimate reliability coefficient of a test or other measuring device, the instrument has to be administered twice to get two sets of scores. The interval of time between the two administrations of the test affects the reliability index of the test. Normally, when the interval between the two administrations of the test is short then the reliability index of the test is high and when the interval between the two administrations of the test is long then the reliability index of the test is low.
- **Environment of test administration:** the physical condition under which the test is administered for estimating reliability, affects the test. Factors such as noise, temperature, light, seating arrangement

etc. influence the test scores which ultimately affect the reliability of the test.

- **Psycho-physical condition of the test-takers:** The psychological and physical conditions of the respondents while appearing the test affect their test scores. Factors such as physical wellbeing, emotional stability, motivation, anxiety, etc. affect test scores and thereby affect the reliability of a test.
- **Guessing:** When the items provide scope for guessing while answering the items then that leads to inconsistent scores by same individuals on the same test. Thus, the factor of guessing may significantly affect the reliability of the test.

These are some of the important factors affecting the reliability of a test.



CHECK YOUR PROGRESS

Q 1: What is reliability?

.....

Q 2: Mention two factors that affect reliability of a test.

.....

Q 3: What are the two types of reliability?

.....

7.5 METHODS OF ESTIMATING RELIABILITY OF A TEST

There are various methods of estimating reliability of test. However, the different methods of estimating reliability may be categorized under two different types. They are: (i) methods of estimating relative reliability, and (ii) methods of estimating absolute reliability. Relative reliability is generally stated in terms of a coefficient or correlation, which is known as reliability coefficient. In this method the correlation between the scores obtained from

two different administration of the test is estimated and labelled as reliability coefficient. The second type of methods estimating absolute reliability is known as Standard Error of Measurement. Standard error of measurement is actually an estimate of the deviation of a set of obtained scores from their true scores.

Several methods are used to estimate relative reliability or reliability coefficient of a test. The most popular methods are:

- The test-retest method of reliability
- The equivalent or parallel form method of reliability
- The split-half or the odd-even method of reliability
- The method of rational equivalence or the Kuder-Richardson method of reliability

The test-retest method: This is a very simple method of estimating reliability of a test. In this method, the same test or the same form of the test is administered twice on the same group of individuals with a reasonable interval of time between the two administrations. Thus, the two sets of scores are obtained from the two administrations of the test. Now, in order to estimate the reliability of the test, coefficient of correlation is calculated between the two sets of scores. This coefficient of correlation is the test-retest reliability of the test which is sometimes called "temporal stability" or "coefficient of stability" as well. This reliability coefficient indicates the extent of consistency of scores in the administration of the test. The greater the index of coefficient of correlation, the more reliable is the test.

The advantages of the test-retest method of reliability are:

- it is a simple method,
- it is an economic method, since only one administration of the test is required to estimate this reliability,
- this method does not require the construction of a second test or form,
- this method is applicable to almost all types of tests,
- this method does not involve complex statistical procedures, simple coefficient of correlation helps.

The disadvantages of the test-retest method of reliability are:

- it is difficult to decide the interval between the two administrations of the test;
- if the interval between the two administration is short, immediate memory, habit, familiarity with the test items may raise the reliability coefficient to a very high level and on the other hand if the interval between the two administration is long, changes in behaviour and maturity of students may make the reliability coefficient low,
- if the conditions under which the two different administrations of the test held are different, then that results in low reliability of the test,
- this method is time consuming.

The equivalent or parallel form method: This method is also known as alternative form method of reliability. In order to estimate reliability using this method, two equivalent or parallel form of the test is prepared. These two forms of the test should be similar in terms of objectives, content, difficulty level of items, form and type of items, length of the test, and other such characteristics. Efforts should be made to make the two forms of the test as much similar as possible but not the same. Freeman holds that both forms (of the test) should meet all of the test's specifications as follows:

The number of items should be the same.

- The kinds of items in both should be uniform in respect to content, operations or traits involved, levels and range of difficulty, and adequacy of sampling.
- The items should be similarly distributed as to difficulty.
- Both test forms should have the same degree of item homogeneity in the operations or traits being measured. The degree of homogeneity may be shown by inter-correlations of each item with subtest scores, or with total-test scores.
- The means and standard deviations of both forms should correspond closely.

- The mechanics of administering and scoring should be uniform.

Though these ideal conditions are desirable, achieving all these is very difficult. After preparing the two equivalent forms of the test, they are administered to the same group of individuals one after the other. The two sets of scores obtained from administering the two forms of the test are correlated to calculate the coefficient of correlation. This coefficient of correlation is the reliability coefficient of the test. According to Guilford, the alternate form method indicates both equivalence of content and stability of performance.

The advantages of the equivalent form method are as follows:

- In this method there is no need to administer the test twice,
- There is no problem of time interval between two administrations,
- The problem of familiarity with test or immediate memory affecting reliability is absent here,
- This method is an advancement over the test-retest method.

The limitations of this method are:

- It is very difficult to prepare two identical forms of a single test,
- This method requires the construction and standardization of a second form of the test which is time consuming and difficult,
- Since items in the two forms are similar in content, difficulty and form, the practice and familiarity effect cannot be controlled completely,
- There are many tests that cannot be divided into two forms.

The split-half or the odd-even method: The split half method is the most commonly used method of estimating reliability of a test. This method divides the test into two equal halves and the scores on one half of the items are correlated with the scores on the other half. However, the test can be split into two halves by using different methods. Commonly, the odd numbered items are pulled together to form one half and the even number items are pulled together to form another half of the test. That is why this method is also known as odd-even method. This method is applied by administering the test only once to the group and then scores on the two halves is drawn separately. Then the scores on the two halves are correlated

using the Spearman-Brown formula to find out the reliability coefficient of the test. Thus, this method measures the internal reliability of the test. If the results show that the two halves do not correlate highly it implies that the two halves are not measuring the same variable. When preparing parallel forms of a test are not advisable and re-testing is not possible this method of estimating reliability is used. Usually this method is used to estimate reliability of attitude scale, personality, inventories, questionnaire, intelligence, performance test, etc. This method of estimating reliability is most popular since only one administration of the test is required. This method thereby saves time, money and effort required to construct and administer a second test.

The major advantages of this method are:

- The problem of preparing equivalent halves of same test is eliminated,
- Issues of memory, familiarity, practice effects, fatigue, maturity etc are removed,
- The variations in testing situations are eliminated since only one administration is required.

The main disadvantages of this method are:

- A test can be divided into two halves in a number of ways; and hence the correlation between the scores on the two halves may not have a unique value,
- It is very difficult to divide a test into two halves in such a way that both halves are equivalent,
- Splitting a test into two equivalent forms is difficult also because items of a test measure same trait or ability.
- This method cannot be used in power tests and heterogeneous tests.

The method of rational equivalence: The rational equivalence method is another method of estimating reliability based on a single administration of the test. This method uses the formula developed by Kuder-Richardson, and hence this method is also known as Kuder-Richardson method. This method can be appropriately used when all the

items in the test are psychologically homogeneous. This means that all the items in the test must measure the same variable or variables in the same proportion. This is known as 'inter-item consistency'. Inter-item consistency means that every item in the test is highly correlated with all other items. If the items in the test are not highly homogeneous, then this method would give a very low index of reliability coefficient.

Kuder-Richardson developed different formulas for computing reliability of a test. Following is the most popularly used formula:

$$r_{11} = \left(\frac{n}{n-1}\right)\left(\frac{SD_t^z - \sum pq}{SD_t^z}\right)$$

Where,

r_{11}= reliability coefficient of the whole test

n = total number of items in the test

SD = Standard Deviation of the whole test

SD^2= the variance of the total test

P = the percent of individuals passing each item

Q = the percent of individuals not passing each item

$\sum pq$ = the sum of the products of pq

The chief advantages of using this method are:

The difficulties of the other methods can be overcome by using this method,

- This method requires single administration of the test,
- The method is simple and economic,
- This method gives accurate results when items are homogeneous,
- This method is scientific.

The limitations of this method are:

- This method cannot be used successfully when the items of the test are not homogeneous,
- Different formulas given by Kuder-Richardson give different estimates of reliability of the same test
- This method does not work with power test or tests with heterogeneous items.

Method of estimating absolute reliability or Standard Error of Measurement:

There is hardly any measurement tool that is perfectly reliable. All instruments have some errors and individuals who measure also make mistakes. Hence inconsistency of tools of measurement and those who use them is not very uncommon to find. Hence, scores obtained by individuals in a test (X) can be thought of as a function of two components, i.e. a true score (T) and an error component (E). Thus, an obtained score(X) = T ± E

The variation between the true score and the observed score is called the measurement error. In statistics, 'error' is any source of variability that cannot be explained. A true score is free from errors of all kinds and it remains the same. But the obtained score may vary from trial to trial because obtained score is influenced by errors. Theoretically, true score is represented by the average (Mean) of the obtained scores on an infinite number of measurements. Each obtained score is either bigger or smaller than the true score. This variation of obtained score from true score is called the errors of measurement. When the error of measurement is smaller, the reliability coefficient will be higher. The errors of measurement or the variation of obtained score from true score are normally distributed and the standard deviation of these variations or errors of measurement is called the standard error of measurement (SEM). The standard error of measurement is a reflection of the reliability of the test.

The formula for determining SEM is written as below:

$$SE_M == SD_x\sqrt{1-r_{xx}}$$

Where,

SE_M = Standard Error of Measurement

SD_x = Standard deviation of the obtained scores

r_{xx} = Reliability Coefficient of the test

Since all measurement contains some error, it is highly unlikely that any test will yield the same scores for a given person each time they are retested. The standard error of measurement (SE_M) estimates how repeated

measures of an individual on the same test are distributed around his "true" score. The true score is always unknown because no test can give a perfect reflection of the true score. The SE_M is a measure of absolute reliability of a test, the smaller the SEM, the greater the reliability and the larger the SE_M, the lower the reliability of the test.



CHECK YOUR PROGRESS

Q 4: Mention any two limitations of the test-retest method of estimating reliability.

.....

Q 5: What is inter-item consistency?

.....

7.6 RELATION BETWEEN RELIABILITY AND VALIDITY

The relation between validity and reliability is very close. Both are indispensable criteria for a good test or measuring instrument. Reliability and validity are the two different but interrelated dimensions of the same thing called test efficiency. Explaining the relation between reliability and validity, F.S. Freeman wrote, "The first necessary condition of a valid test is that it has an adequate degree of reliability. If the reliability coefficient of a test is 'zero', it cannot correlate with anything. A test that correlates poorly even with itself cannot correlate well with a measure of another variable."

The relationship between validity and reliability may be seen through the following points:

- Validity means truthfulness of a test; it indicates how well the purpose of the test is being served by the test. It is a measure of the extent to which a test measures what it purports to measure. On the other hand, reliability refers to the dependability or consistency or stability of the test scores. It is not directly concerned about the purpose of

the test.

- While reliability is concerned with the consistency of the test scores or self-correlation of the test, validity is the measure of the correlation of the test with some external criterion.
- A valid test is always reliable, but a reliable test is not necessarily valid. A test with high correlation with itself (reliability) may not have any correlation with external criterion (validity).
- A test which is not reliable cannot be valid. If a test is not reliable, it implies that, it is not measuring the same thing every time it is used; if it is not measuring the same thing then it cannot be valid. From this perspective, reliability is the precursor of validity. However, reliability alone cannot ensure validity of a test. For example, a clock may consistently (reliable) give wrong time (invalid). But a valid test is always reliable. A test to be valid must consistently measure what it intends to measure.
- A test has maximum reliability when items in the test are homogeneous; on the other hand, a test is more valid when items are heterogeneous.
- Reliability of a test requires inter item consistency with uniform difficulty level; on the other hand validity requires items with varied difficulty level.

Though there are some differences between validity and reliability, significant relation between the two exists and both are equally important for a good test.



7.7 LET US SUM UP

After going depth into this unit, it can be summarized as-

- Any measuring tool used to collect data for measurement and evaluation in education must be reliable. Reliability refers to the consistency of test scores; it represents the extent of accuracy or precision of a measuring tool. If a test or other tool gives consistent results on testing and retesting in variable conditions then the tool is

accepted to be reliable. Reliability is a precondition to the validity of a test as well. However, reliability is the characteristic of the scores only and it is not the property of the test itself.

- Reliability of a test may be affected by various factors. Some of the important factors that affect reliability of a test include: length of the test, methods used to estimate reliability, nature of the group, too difficult or too easy items, lack of objectivity in scoring, lack of clarity of instruction to candidates, interval between testing and retesting, environmental conditions under which the test is administered, psycho-physical condition of the test-takers at the time of taking the test, guessing of correct answer by the respondents, etc.
- There are various methods of estimating reliability of test. The different methods of estimating reliability may be categorized under two different types. They are: (i) methods of estimating relative reliability, and (ii) methods of estimating absolute reliability. Relative reliability is a coefficient or correlation, which is known as reliability coefficient. Several methods are used to estimate relative reliability or reliability coefficient of a test. The most popular methods are:
  - the test-retest method of reliability
  - the equivalent or parallel form method of reliability
  - the split-half or the odd-even method of reliability
  - the method of rational equivalence or the Kuder-Richardson method of reliability
- The second type of methods estimating absolute reliability is known as Standard Error of Measurement. Standard error of measurement is actually an estimate of the deviation of a set of obtained scores from their true scores.
- Validity and reliability are two indispensable criteria for a good test or measuring instrument. The relation between validity and reliability is very close. They are the two different but interrelated dimensions of the same thing called test efficiency.

7.8 FURTHER READING

1) Anastasi, Anne (1982): Psychological testing, Edition 5, illustrated. Macmillan, The University of Michigan.
2) Ebel, R.L. and Frisbel, D.A. (1979): Essential of Educational Measurement, 3rd Edition, Prentice Hall of India, New Delhi.
3) Freeman, F.S. (1962): Theory and Practice of Psychological Testing, Holt, Rirehant & Winston, New York
4) Gupta, S.P. (2009): Modern Measurement and Evaluations with Statistics, Sharda Pustak Bhavan, Allahabad
5) Kelley, T. L. (1927): Interpretation of educational measurements, Macmillan, New York
6) Mehrens, W.A. and Lehmann, I.J. (1962): Standardized tests in Education, Holt, Rinehart and Winston, New York
7) Newton, Paul & Shaw, Stuart (2014): Validity in Educational and Psychological Assessment, Sage Publications, London.
8) Pradhan Hemant (1996): Essentials of Educational Measurement, Evaluation and Statistics, Saha Publishers and Distributers, Sambalpur
9) Stanley & Hopkins (1966): Education and Psychological Measurement and Evaluation, Prentice Hall Pvt. Ltd. Eaglewood Cliff, New Jersey

7.9 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: Reliability refers to the consistency of test scores; it represents the extent of accuracy or precision of a measuring tool. If a test or other tool gives consistent results on testing and retesting in variable testing conditions then the tool is accepted to be reliable.

Ans to Q No 2: Two factors affecting the reliability of a test are: (i) environmental conditions under which the test is administered and (ii) psycho-physical condition of the test-takers at the time of taking the test.

Ans to Q No 3: The two types of reliability are: (i) relative reliability, and (ii) absolute reliability. Relative reliability is generally stated in terms of a coefficient or correlation, which is known as reliability coefficient. In this

method the correlation between the scores obtained from two different administration of the test is estimated and labelled as reliability coefficient. The second type of methods estimating absolute reliability is known as Standard Error of Measurement. Standard error of measurement is actually an estimate of the deviation of a set of obtained scores from their true scores.

Ans to Q No 4: The two limitations of the test-retest method of estimating reliability are: (i) test retest method requires two administration of the same test; if the interval between the two administration is short, immediate memory, habit, familiarity with the test items may raise the reliability coefficient to a very high level and on the other hand if the interval between the two administration is long, changes in behaviour and maturity of students may make the reliability coefficient low, (ii) if the conditions under which the two different administrations of the test held are different then that results in low reliability of the test.

Ans to Q No 5: Inter-item consistency means that every item in the test is highly correlated with all other items. This means that all the items in the test must measure the same variable or variables in the same proportion. This inter-item consistency is very important for calculating reliability of a test using the kuder-Richardson formula.



7.10 MODEL QUESTIONS

Short Questions (Answer each question in around 150 words)

Q 1: Define reliability.

Q 2: What is standard error of measurement?

Q 3: Explain the relation between reliability and validity.

Long Questions. (Answer each question in around 300-500 words)

Q 1: Discuss the factors that affect reliability of a test.

Q 2: Explain any two methods of estimating relative reliability of a test.

Q 3: Discuss the relative advantages and disadvantages of the different methods of estimating relative reliability of a test.

Q 4: Explain the process of estimating absolute reliability of a test.

*** ***** ***