UNIT 9  INTRODUCTION TO TEST CONSTRUCTION*

Structure
9.0 Objectives
9.1 Introduction
9.2 Psychological Test or Measure
  9.2.1 Purpose of Psychological Test
9.3 Validity
  9.3.1 Types of Validity
9.4 Reliability
  9.4.1 Measuring Reliability
9.5 Norms
  9.5.1 Types of Norms
9.6 Test Construction
  9.6.1 Standardization of Psychological Tests
  9.6.2 Classification of Standardized Testing
  9.6.3 Steps for constructing Standardized Tests
9.7 Let Us Sum Up
9.8 Unit End Questions
9.9 Answers to Self Assessment Questions
9.10 Glossary
9.11 Suggested Readings and References

9.0 OBJECTIVES

After reading this unit, you will be able to:
- Define Reliability, Validity, Norms and Standardized Tests;
- Describe the process of test construction;
- Illustrate the different types of reliability and validity; and
- Explain the steps of constructing a standardized test.

9.1 INTRODUCTION

The present unit of this block deals with the meaning and ways of test construction with reference to psychological research. You will also be introduced with the concept of reliability, validity, norms and standardization of psychological tests. At the end of this unit, you will be explained about the various steps of constructing a standardized test.

9.2 PSYCHOLOGICAL TEST OR MEASURE

It basically refers to a self-report study where the answers or the responses are measured and combined to get a total score. Here we can interchange the usage

* Dr. Sarika Boora, Psychologist, Sambharti Foundations, New Delhi.
Methods of Data Collection

of the words “test” and “measures” although the term “test” in common jargon refers to educational test or an exam with right or wrong responses, though, in most cases, there is no right and wrong answer in a psychological “measure”. Also, we can use the terms “scale” and “questionnaire” interchangeably and the both would imply to a set of question whose answers are combined to form a total score. The most important aspects therefore are (a) a set of questions for an individual to answer, and (b) a combined score that comes from measuring their answers. This set of questions in conjunction is known as a “scale,” “test,” or “measure.”

9.2.1 Purpose of a Psychological Test

There are two objectives of a test whether in education or psychology:

- An attempt to compare the same person on two or more aspects or variables of characteristics or trait.
- A comparison to be made between two individuals on the basis of a common trait. This test can either be qualitative or quantitative.

9.3 VALIDITY

It is apparent from the description of statistical inferences that all tests do not have a common degree of validity, it being completely dependent upon the extent to which it has been used to measure the same dimension; along with its norms in detail judgment based upon the test’s result is objective. For example, validity for intelligence tests may be attributed only to those tests that actually succeed in testing the individual’s level of intelligence. A test can be accepted as valid only to that degree to which it can correctly gauge the mentioned dimension of the participant, which it claims to measure. In this way, validity of a test is that quality on the basis of which the correctness or incorrectness of judgments based upon it is evaluated. For example, the validity of interest tests is less than the case of intelligence tests. Here there is a slight difficulty. Suppose for the moment, that the intelligence of some students was measured by one particular method. Now the validity of the test will depend upon whether the students tested do, in fact, possess the intelligence that they are indicated as possessing. The problem that arises here is how can one ascertain whether the students do or do not possess the dimension of intelligence as indicated by the above test. Evidently, there must be some independent criterion for deciding upon the validity of the particular test in question or tests in general. In this case, of the intelligence level of students, the examination results can be the basis for measuring the validity of the test. Generally speaking, it can be said that if there is a correlation between marks obtained at an examination and the result of the test, then the test is valid.

9.3.1 Types of Validity

As it is evident from the foregoing description, validity is a relative term, as no test can have complete validity. Hence, whenever a particular test is termed valid, or whenever the lack of validity of a test is in question, it is necessary to indicate the sense in which it is considered to be valid or invalid.

Apparenty, validity is of many kinds. Psychologists have roughly accepted the following kinds of validity:

1) Face validity: It only focuses on the form of the test. Such validity is attributed only to the test which provides an item or subject that just appears to be valid.
2) **Content validity**: Another kind of validity is content validity in which the validity of the content forms the basis of the validity of the test. In order to obtain this kind of validity in a particular test, it becomes imperative that the items of the test achieve the objective for which they are originally designed. For example, content validity in the case of an intelligence test will be attributed only in the event of its succeeding in discovering all the factors that are concerned with intelligence.

3) **Factorial validity**: This is inclusive of the validity of the factors in the test and in order to judge whether a test has factorial validity, it is examined by the method of factor analysis, and a correlation between this result and the evident factor resultant of tests is established.

4) **Predictive validity**: This is the most popular form of validity. In this, results are obtained on the basis of a particular criterion, and the correlation between the scores and the criterion is established. In this, the choice of a criterion requires much care and attention. The coefficient obtained by this correlation between scores and criterion is called the validity coefficient. The validity coefficient varies between 0.5 and 0.8. A lower coefficient makes the test inapplicable and lacking in utility, while a higher coefficient is not normally obtained.

5) **Concurrent validity**: It resembles the predictive validity since in it, also, a correlation between the test and some definite standard is established. But, despite these common features, there are also some definite variations.

From the above analysis of the various kinds of validity, it is evident that validity exists in a particular context, or in other words, every test is valid for a particular objective and for a specific age group among individuals. It can just as well be invalid for a specific age group in individuals and in a particular context. Hence, to attribute validity to a test without qualification is completely unjustified and inaccurate. For the sentiment to have any value or meaning, it is essential to state the context and conditions in which it is applicable.

### 9.4 RELIABILITY

In addition to validity, it is essential that every test should possess definite element of reliability. It is only then that the conclusions of the test can be considered reliable and worthy of trust. The term basically refers to the extent to which a test can be relied upon, i.e. it gives consistency in scores even if it is tested on the same group after frequent intervals/ time gap.

Reliability of a test refers to the quality of the test that may inspire confidence and trust for the measurement. And this quality can be attributed to only that test which provides the same score every time it is performed on the same individual. Now, if some intelligence test yields one score for an individual at one time, and another at the same individual if it is applied to the same individual at a different time, it is too evident that such a test cannot be considered reliable and the reliability of a test is not a part of it, but is in its wholeness or completeness. Its reliability will considerably be weakened and decreased if even one part of it is injured in some respect. Hence, it is essential that the internal parts of a test possess internal consistency and uniformity. It is only on the basis of such a reliable test that guidance can be given.
9.4.1 Measuring Reliability

Reliability can be measured in the following four ways:

1) **Test retest method:** One method of gauging reliability is to perform the same test on the same group of individuals at two different occasions, and then the scores or results obtained are compared. For example, a group of individuals can be subjected to the Binet intelligence test. Then later on the same group of individuals can again be subjected to the Binet intelligence test. If the results obtained in each case do not tally, then the tests cannot be considered reliable.

2) **Parallel form method:** In the parallel form of reliability, same group is provided two different tests measuring the same dimension or construct. Finally, the results scores of the two can be compared or corrected to judge the reliability of the test. Gulliksen has suggested that more than one parallel method be devised for greater accuracy. It is also known as equivalent form of reliability.

3) **Split half method:** The reliability of a test can also be judged by dividing the components of the test into even and odd times whose results can be individually obtained. Now the results can be compared between the groups to check the reliability of the test.

4) **Inter item consistency:** In this method of measuring the reliability only one method is applied at one time. The mutual relation between the scores obtained for each specific item in the test is observed. At the same time the relation between the marks obtained for one specific question and the marks obtained for the whole test is also ascertained. This method of measuring reliability involves considerable statistical skill in correlation. Psychologists Kuder and Richardson have devised some formulae for application in this method.

As has been indicated previously, the implication and meaning of reliability also changed the method of judging reliability. Hence, it is not sufficient to remark that a particular test is reliable. It is equally essential that the sense in which reliability is judged also be mentioned.

Of the above mentioned methods of judging reliability of psychological tests, the third one is the most prevalent and useful, since it is the most easy. In this method, the necessity of collecting the same group of individuals more than once is obviated. Reliability is known from the coefficient of reliability and this coefficient is known as the reliability coefficient.

In this manner, both reliability and validity are important qualities of tests. Validity is related to the scale or structure of the test while reliability is an attribute of its ability of costing.

**Self Assessment Questions I**

Describe the following in two-three lines:

1) **Validity**

........................................................................................................................................................................
........................................................................................................................................................................
Norm refers to the typical performance level for a certain group of individuals. Any psychological test with just the raw score is meaningless until it is supplemented by additional data to interpret it further. Therefore, the cumulative total of a psychological test is generally inferred through referring to the norms that depict the score of the standardized sample. Norms are factually demonstrated by establishing the performance of individuals from a specific group in a test. To determine accurately a subject’s (individual’s) position with respect to the standard sample, the raw score is transformed into a relative measure. There are two purposes of this derived score:

1) They provide an indication to the individuals standing in relation to the normative sample and help in evaluating the performance.
2) To give measures that can be compared and allow gauging of individuals performance on various tests.

9.5.1 Types of Norms

Fundamentally, norms are expressed in two ways, developmental norms and within group norms.

1) Developmental Norms

These depict the normal developmental path for an individual’s progression. They can be very useful in providing description but are not well suited for accurate statistical purpose. Developmental norms can be classified as mental age norms, grade equivalent norms and ordinal scale norms.

2) Within Group Norms

This type of norm is used for comparison of an individual’s performance to the most closely related groups’ performance. They carry a clear and well defined quantitative meaning which can be applied to most statistical analysis.

a) Percentiles ($P(n)$ and $PR$): They refer to the percentage of people in a standardized sample that are below a certain set of score. They depict an individual’s position with respect to the sample. Here the counting begins from bottom, so the higher the percentile the better the rank. For example if a person gets 97 percentile in a competitive exam, it means 97% of the participants have scored less than him/her.
Methods of Data Collection

b) Standard Score: It signifies the gap between the individuals score and the mean depicted as standard deviation of the distribution. It can be derived by linear or nonlinear transformation of the original raw scores. They are also known as T and Z scores.

c) Age Norms: To obtain this, we take the mean raw score gathered from all in the common age group inside a standardized sample. Hence, the 15 year norm would be represented and be applicable by the mean raw score of students aged 15 years.

d) Grade Norms: It is calculated by finding the mean raw score earned by students in a specific grade.

9.6 TEST CONSTRUCTION

Attention must be given to the below mentioned points while constructing a potent, constructive and relevant questionnaire/schedule:

• The researcher must first define the problem that s/he wants to examine, as it will lay the foundation of the questionnaire. There must be a complete clarity about the various facets of the research problem that will be encountered as the research progresses.

• The correct formulation of questions is dependent on the kind of information the researcher seeks, the objective of analysis and the respondents of the schedule/questionnaire. Whether to use open ended or close ended questions should be decided by the researcher. They should be uncomplicated and made with such a view that there will be an objective part of a calculated tabulation plan.

• A researcher must prepare a rough draft of the schedule while giving ample thought to the sequence in which s/he wants to place the questions. Previous examples of such questionnaires can also be observed at this stage.

• A researcher by default should recheck and if required make changes in the rough draft to improve the same. Technical discrepancies should be examined in detail and changed accordingly.

• There should be a pre-testing done through a pilot study and changes should be made to the questionnaire if required.

• The questions should be easy to understand the directions to fill up the questionnaire clearly mentioned; this should be done to avoid any confusion.

The primary objective of developing a tool is obtaining a set of data that is accurate, trustworthy and authentic so as to enable the researcher in gauging the current situation correctly and reaching conclusions that can provide executable suggestions. But, no tool is absolutely accurate and valid, thus, it should carry a declaration that clearly mentions its reliability and validity. Next, we will discuss how to develop a standardised tool/test.

9.6.1 Standardization of Psychological Tests

Standardization refers to the consistency of processes and procedures that are used for conducting and scoring of a test. To compare the scores of different individuals the conditions should be the same.

In case of a new step the first and major step in standardization is formulating the directions. This also includes the type of materials to be used, verbal instructions,
Establishing the norms is also a key step for standardization. Norm refers to the average performance. To standardize a test, we administer it to a big, representative sample of the kind of individuals it was designed for. The aforementioned group sets the norms and is called the standardization sample.

The norms for personality tests are set in the same way as those set for aptitude tests. For both, the norm would refer to the performance of average individuals.

To construct and administer a test, standardization is a very important. The test is administered on a large set number of the people (the conditions and guidelines need to be the same for all). After which the scores are modified using Percentile rank, Z-score, T-score and Stanine, etc. The standardization of a test can be established from this modified score. Hence, “standardization is a process of ensuring that a test is standardized, (Osadebe, 2001)”. There are lots of advantages when a test is standardized. A standard test is usually produced by experts and it is better than teacher made test. The standardized test is highly valid, reliable and normalized with Percentile rank, Z-score, T-score among scores derived from others to produce age norm, sex norm, location norm and school-type norm.

Generally, a standardized test could be used to assess, and compare students in the same norming group.

The normal process for administering standardization includes:

1) A calm, quiet and disturbance free setting
2) Accurately understanding the written instructions, and
3) Provisioning of required stimuli.

This makes the normative data applicable to the individuals being evaluated.

### 9.6.2 Classification of Standardized Testing

**Norm-referenced Testing:** It is used to measure the result or performance in relation to all other individuals being administered the same test. It can be used to compare an individual to the others.

**Criterion referenced Testing:** It is used for measuring the real knowledge of a certain topic.

For example: Multiple choice questions in a geography quiz.

### 9.6.3 Steps for Constructing Standardized Tests

A carefully constructed test where the scoring, administration and interpretation of result follows a uniform process can be termed as a standardized test. Following are the steps that can be followed to construct a standardised test:

**Steps**

1) Plan for the test.
2) Preparation of the test.
3) Trial run of the test.
4) Checking the Reliability and Validity of the test.
5) Prepare the norms for the test.
6) Prepare the manual of the test and reproducing the test.
Methods of Data Collection

1) **Planning** – There needs to be a systematic planning in order to formulate a standardized test. Its objectives should be carefully defined. The type of content should be determined for example using short/long/very short answers or using multiple type questions, etc. A blueprint must be ready with instructions to the method to be used for sampling, making the necessary requirements for preliminary and final administration. The length, time for completing the test and number of questions should be fixed. Detailed and precise instructions should be given for administration of the test and also its scoring.

2) **Writing the items of the test** – This requires a lot of creativity and is dependent on the imagination, expertise and knowledge. Its requirements are:

   - In-Depth knowledge of the subject
   - Awareness about the aptitude and ability of the individuals to be tested.
   - Large vocabulary to avoid confusion in writing. Words should be simple and descriptive enough for everybody to understand.
   - Assembly and arrangement of items in a test must be proper, generally done in ascending order of difficulty.
   - Detailed instructions of the objective, time limit and the steps of recording the answers must be given.
   - Help from experts should be taken to crosscheck for subject and language errors.

3) **Preliminary Administration** – After modifying the items as per the advise of the experts the test can be tried out on experimental basis, which is done to prune out any inadequacy or weakness of the item. It highlights ambiguous items, irrelevant choices in multiple choice questions, items that are very difficult or easy to answer. Also the time duration of the test and number of items that are to be kept in the final test can be ascertained, this avoids repetition and vagueness in the instructions.

   This is done in following three stages:

   a) **Preliminary try-out** – This is performed individually and it helps in improving and modifying the linguistic difficulty and vagueness of items. It is administered to around hundred people and modifications are done after observing the workability of the items.

   b) **The proper try-out** – It is administered to approximately four hundred people wherein the sample is kept same as the final intended participants of the test. This test is done to remove the poor or less significant items and choose the good items and includes two activities:

      - **Item analysis** – The difficulty of the test should be moderate with each item discriminating the validity between high and low achievers. Item analysis is the process to judge the quality of an item.

      - **Post item analysis**: The final test is framed by retaining good items that have a balanced level of difficulty and satisfactory discrimination. The blueprint is used to guide in selection of number of items and then arranging them as per difficulty. Time limit is set.
c) **Final try-out** – It is administered on a large sample in order to estimate the reliability and validity. It provides an indication to the effectiveness of the test when the intended sample is subjected to it.

4) **Reliability and Validity of the test** – When test is finally composed, the final test is again administered on a fresh sample in order to compute the reliability coefficient. This time also sample should not be less than 100. Reliability is calculated through test-retest method, split-half method and the equivalent -form method. Reliability shows the consistency of test scores. Validity refers to what the test measures and how well it measures. If a test measures a trait that it intends to measure well then the test can be said to be a valid one. It is correlation of test with some outside independent criterion.

5) **Norms of the final test** – Test constructor also prepares norms of the test. Norms are defined as average performance scores. They are prepared to meaningfully interpret the scores obtained on the test. The obtained scores on test themselves convey no meaning regarding the ability or trait being measured. But when these are compared with norms, a meaningful inference can be immediately drawn.

The norms may be age norms, grade norms etc. as discussed earlier. Similar norms cannot be used for all tests.

6) **Preparation of manual and reproduction of the test** – The manual is prepared as the last step and the psychometric properties of the test norms and references are reported. It provides in detail the process to administer the test, its duration and scoring technique. It also contains all instructions for the test.

**Self Assessment Questions II**

Fill in the following blanks:

1) A blue print must be ready with instructions to the method to be used for sampling, making the necessary requirements for preliminary and final administration.

2) .................................. signifies the gap between the individuals score and the Mean depicted as standard deviation of the distribution.

3) .................................. refers to the typical performance level for a certain group of individuals.

4) .................................. is calculated by finding the mean row score earned by students in a specific grade.

**9.7 LET US SUM UP**

It can be summed up from the above discussion that psychological tests needs to be prepared in a standardized way. A test can be said to be a standardized one, if it is reliable, valid and has standardized norms. Different types of reliability and ways of measuring reliability were also discussed in the unit. At the end of the unit, you were also informed about the various steps of constructing a standardized test.

**9.8 UNIT END QUESTIONS**

1) What is a Psychological test? Explain it’s purpose.
2) Write down the steps of test construction.

3) Explain the concept and ways of measuring reliability.

4) Describe the different types of validity.

5) Explain the concept and types of norms.

9.9 ANSWERS TO SELF ASSESSMENT QUESTIONS

Self Assessment Questions I

1) It refers to the degree to which it can correctly gauge the mentioned dimension of the participant, which it claims to measure.

2) It refers to the consistency in scores even if it is tested on the same group after frequent intervals/time gap.

3) It refers to a self-report study where the answers are measured and combined to get a total score.

Self Assessment Questions II

1) Planning

2) Standard Score

3) Norm

4) Grade Norms

9.10 GLOSSARY

Validity: It refers to the degree to which it can correctly gauge the mentioned dimension of the participant, which it claims to measure.

Reliability: It refers to the consistency in scores even if it is tested on the same group after frequent intervals/time gap.

Psychological Test: It refers to a self-report study where the answers are measured and combined to get a total score.

Norms: Norm of a psychological test refers to the typical performance level for a certain group of individuals.

9.11 SUGGESTED READINGS AND REFERENCES


P. J. A. Rulon, (1939), Simplified procedure for determining the reliability of a test by split halves theory, Edu. Pr. 9, 99-103

