UNIT 2 RELIABILITY AND VALIDITY (EXTERNAL AND INTERNAL)

Structure

2.0 INTRODUCTION

Most research is designed to draw the conclusion about the cause and effect relationship among the variables. The goal of the research remains to develop a theory that explains the relationship found among variables. This unit mainly concerns about various problems that can threaten the reliability and validity of conclusions drawn by the researcher.

There are two goals of research design;

1) Obtain information relevant to the purposes of the study.

2) Collect this information with maximal reliability and validity.

How can a researcher be sure that the data gathering instrument being used will measure what it is supposed to measure and will do this in a consistent manner?

This is a question that can only be answered by examining the definitions for and methods of establishing the validity and reliability of a research instrument.

Reliability and validity are central issues in all measurement. Both concern connecting measures to constructs. Reliability and validity are salient because constructs are often ambiguous, diffused and not directly observable. Perfect reliability and validity are virtually very difficult to achieve. These two very important aspects of research design will be discussed in this unit. All researchers want their measures to be reliable and valid. Both ideas help to establish the truthfulness, credibility, or believability of findings. This unit will be discussed in two parts. First part covers the concept of reliability and the definitions of reliability. This is followed by various methods of establishing reliability of a research instrument of this unit. Second part of this unit discusses the concept of validity in research. You will familiarise with the various types of validity. Finally, some problems that constitute threats to validity are described.

2.1 OBJECTIVES

After reading this unit, you will be able to:

- Define reliability;
- Describe the various methods of calculating reliability;
- Explain how test retest reliability is accessed;
- Differentiate between tests of reliability;
- Define validity;
- Describe various methods of validity;
- Identify the problems that constitute threats to internal external validity; and
- Differentiate between internal and external validity.

2.2 RELIABILITY

Meaning of Reliability

The idea behind reliability is that any significant results must be repeatable. Other researchers must be able to perform exactly the same experiment, under same conditions and generate the same results. This will vindicate the findings and ensure that all researchers will accept the hypothesis. Without this replication of statistically significant results, experiment and research have not fulfilled all of the requirements of testability. This prerequisite is essential to a hypothesis establishing itself as an accepted scientific truth. For example, if you are performing a time critical experiment, you will be using some type of stopwatch. Generally, it is reasonable to assume that the instruments are reliable and will keep true and accurate time. However, scientists take measurements many times, to minimize the chances of malfunction and maintain validity and reliability. At the other extreme, any experiment that uses human judgment is always going to come under question. Human judgment can vary as individual observer may rate

things differently depending upon time of day and current mood. This means that such experiments are more difficult to repeat and are inherently less reliable. Reliability is a necessary ingredient for determining the overall validity of a scientific experiment and enhancing the strength of the results.

Reliability is the consistency of your measurement, or the degree to which an instrument measures the same way each time it is used under the same condition with the same subjects. In short, it is the repeatability of measurement. A measure is considered reliable if a person's score on the same test given twice is similar. It is important to remember that reliability is not measured, it is estimated. For instance, if a test is constructed to measure a particular trait; say, neuroticism, then each time it is administered, it should yield same results. A test is considered reliable if we get same result repeatedly.

According to Anastasi (1957), the reliability of test refers to the consistency of scores obtained by the individual on different occasions or with different sets of equivalent items.

According to Stodola and Stordahl (1972), the reliability of a test can be defined as the correlation between two or more sets of scores of equivalent tests from the same group of individuals.

According to Guilford (1954), reliability is the proportion of the true variance in obtained test scores.

The reliability of test is also defined from another angle. Whenever we measure something, measurement involves some kind of measure. Error of measurement is generally between true scores and the observed score. However, in psychological term, word error does not imply the mistake has been made. In other words, error in psychological testing implies that there is always some inaccuracy in measurement. Hence, goal of psychological measurement remains to find out the magnitude of such error and develop ways to minimize them.

2.3 METHODS OF ESTIMATING RELIABILITY

There are number of ways of estimating reliability of an instrument. Various procedures can be classified into two groups:

External consistency procedures

Internal consistency procedures

2.3.1 External Consistency Procedures

External consistency procedures compare findings from two independent process of data collection with each other as a means of verifying the reliability of the measure. Two methods are as beneath.

2.3.1.1 Test Re-test Reliability

The most frequently used method to find the reliability of a test is by repeating the same test on same sample, on two different time periods. The reliability coefficient in this case would be the correlation between the score obtained by the same person on two administrations of the test.

Test-Retest reliability is estimated, when same test is administered on same sample. Therefore, if refers to the consistency of a test among on two different time periods different administrations. The assumption behind this approach is that there will be no substantial changes in the measurement of the construct in question, upon administration on separate occasions. The time gap that is given between measures is of critical value, the shorter the time gap, higher the correlation value and vice versa. If the test is reliable, the scores that are attained on first administration should be more or less equal to those obtained on second time also. The relationship between the two administrations should be highly positive.

Limitations of this approach

There are a few limitations which include the following: (i) Memory Effect/ carry over Effect (ii) Practice effect, (iii) Absence. These are being discussed below:

i) **Memory effect /carry over effect:** One of the common problems with test-retest reliability is that of memory effect. This argument particularly holds true when, the two administrations takes place within short span of time, for example, when a memory related experiment including nonsense syllables is conducted whereby, the subjects are asked to remember a list in a serial wise order, and the next experiment is conducted within 15 minutes, most of the times, subject is bound to remember his/her responses, as a result of which there can be prevalence of artificial reliability coefficient since subjects give response from memory instead of the test. Same is the condition when pre-test and post-test for a particular experiment is being conducted.

ii) **Practice effect:** This happens when repeated tests are being taken for the improvement of test scores, as is typically seen in the case of classical IQ where there is improvement in the scores as we repeat these tests.

iii) **Absence:** People remaining absent for re-tests.

2.3.1.2 Parallel Forms Reliability

Parallel-Forms Reliability is known by the various names such as Alternate forms reliability, equivalent form reliability and comparable form reliability.

Parallel forms reliability compares two equivalent forms of a test that measure the same attribute. The two forms use different items. However, the rules used to select items of a particular difficulty level are the same. When two forms of the test are available, one can compare performance on one form versus the other. Sometimes the two forms are administered to the same group of people on the same day.

The Pearson product moment correlation coefficient is used as an estimate of the reliability. When both forms of the test are given on the same day, the only sources of variation are random error and the difference between the forms of the test. Sometimes the two forms of the test are given at different times. In these cases, error associated with time sampling is also included in the estimate of reliability.

The method of parallel forms provides one of the most rigorous assessments of reliability commonly in use. Unfortunately the use of parallel forms occurs in practice less often than is desirable. Often test developers find it burdensome to

develop two forms of the same test, and practical constraints make it difficult to retest the same group of individuals. Instead many test developers prefer to base their estimate or reliability on a single form of a test.

In practice, psychologists do not always have two forms of a test. More often they have only one test form and must estimate the reliability for this single group of items. You can asses the different sources of variation within a single test in many ways. One method is to evaluate the internal consistency of the test by dividing it into subcomponents.

2.3.2 Internal Consistency Procedures

The idea behind internal consistency procedures is that items measuring same phenomena should produce similar results. Following internal consistency procedures are commonly used for estimating reliability-

2.3.2.1 Split Half Reliability

In this method, as the name implies, we randomly divide all items that intends to measure same construct into two sets .The complete instrument is administered on sample of people and total scores are calculated for each randomly divided half; the split half reliability is then, the simply the correlation between these two scores.

Problem in this approach

A problem with this approach is that when the tests are shorter, they run the risk of losing reliability and it can most safely be used in case of long tests only. It is, hence, more useful in case of long tests as compared to shorter ones. However to rectify the defects of shortness, Spearman- Brown's formula can be employed, enabling correlation as if each part were full length:

$r = (2rhh)/(1 + rhh)$ (Where rhh is correlation between two halves)

2.3.2.2 Kudar-Richardson Estimate of Reliability

The coefficient of internal consistency could also be obtained with the help of Kudar-Richardson formula number 20. One of the techniques for item analysis is item difficulty index. Item difficulty is the proportion or percentage of those answering correctly to an item. For example – symbol 'p' is used to represent the difficulty index. Suppose an item 'X' has p=0.67.this means item 'X' was answered correctly by 74% of those who answered the item. To compute reliability with the help of Kuder-Richardson formula number 20, the following formula is used:

$$KR\text{-}20 = \frac{N}{N-1}\left(1 - \frac{\Sigma pq}{\sigma^2}\right)$$

Where

N = the number of items on the test,

σ^2 = the variance of scores on the total test,

p = the proportion of examinees getting each item correct,

q = the proportion of examinees getting each item wrong.

Kuder-Richardson formula 20 is an index of reliability that is relevant to the special case where each test item is scored 0 or 1 (e.g., right or wrong).

2.3.2.3 Cronbach's Alpha (α)

As proposed by Cronbach (1951) and subsequently elaborated by others (Novick & Lewis, 1967; Kaiser & Michael, 1975), coefficient alpha may be thought of as the mean of all possible split-half coefficients, corrected by the Spearman-Brown formula .The formula for coefficient alpha is

$$r_\alpha = \left(\frac{N}{N-1}\right)\left(1 - \frac{\Sigma \sigma^2_j}{\sigma^2}\right)$$

Where $r\alpha$ is coefficient alpha,

N is the no. of items,

σ^2_j is the variance of one item,

$\Sigma\sigma^2_j$ is the sum of variances of all items, and

$\sigma^2$is the variance of the total test scores.

As with all reliability estimates, coefficient alpha can vary between 0.00 and 1.00.

Coefficient alpha extends the Kuder-Richard-son method to types of tests with items that are not scored as 0 or 1. For example, coefficient alpha could be used with an attitude scale in which examinees indicate on each item whether they strongly agree, agree, disagree, or strongly disagree.

2.4 COMPARISON OF RELIABILITY ESTIMATORS

All of the reliability estimators listed above have certain pros and cons, like for example: inter-rater is best suited when the measure involves observation, it however requires multiple observers as an alternative one can look at of rating of a single observer repeated on single occasion. It can also be used if the examiner is interested in using a team of raters.

In a situation that involves use of two forms as alternate measure of the same thing, parallel forms estimator is best suited. However, this and the internal consistency measures of reliability have constrains, i.e. one has to have multiple items engineered to measure same construct.

Cronbach's Alpha is useful in case, where lots of items are present. The test-retest reliability is mostly employed in case of experimental and quasi-experimental designs. This also depends upon string of availability of a control group that is measured on two different occasions and until post-test is done, one does not have information about reliability. Accordingly, each one of the above mentioned estimators will give a different value for reliability. Generally, test-retest and inter-rater reliability estimates will be lower in value as compared to parallel forms and internal consistency due to involvement in measurement at different times or with different raters.

Self Assessment Questions

1) Internal Consistency Concerns whether the various items on a test are measure the same thing. T / F

2) Memory effect / carry over effect is possible in parallel form method. T / F

3) K.R. Formula is applied in which each test item is scored 0 or 1. T / F

4) Scores from the two halves of a test are correlated with one another in split half reliability. T / F

5) Spearman Brown formula is used for adjusting split half correlation T / F

Answer: 1) T, 2) F, 3) T, 4) T, 5) T

2.5 VALIDITY

As you know that the merit of the psychological test is determine first by its reliability but then ultimately by its validity. Validity refers to the degree to which a test measures, what it claims to measure. It is very necessary for a test to be valid for its proper administration and interpretation.

According to Standard for Educational and Psychological testing (AERA, APA & NCME 1985, 1999); a test is valid to the extent that inferences drawn from it are appropriate, meaningful and useful.

According to Cronbach (1951) validity is the extent to which a test measures what it purports to measure.

According to Freeman (1971) an index of validity shows the degree to which a test measures what it purports to measure when compared with accepted criteria.

According to Anastasi (1988) the validity of a test concerns what the test measures and how well it does so.

The above definitions pointed out that for determining the validity of the test, the test must be compared with some ideal independent measures or criteria. The correlation coefficients computed between the test and an ideal criterion is known as the validity coefficients. Independent criteria refer to some measure of the trait or group of the traits (out side the test) that the test itself claims to measure.

2.6 TYPES OF VALIDITY

There are six types of validity, viz., (i) Content validity (ii) Criterion-related validity (iii) Con current validity (iv) Predictive validity (v) Construct validity (vi) Convergent validity (vii) Discriminate validity and (viii) Face validity. These are being discussed below:

2.6.1 Content Validity

According to Mc Burney and White (2007); content validity is the notion that a test should sample range of behaviour that is represented by the theoretical concept being measured.

It is a non-statistical type of validity with involvement of assessment of the content of the test to ascertain whether it includes the sample representative of the behaviour that is intended to be measured. When a test has content validity, the items on the test represent the entire range of possible items the test should cover. For instance, if researcher wants to develop an achievement test of spelling for the third grade children then a researcher could identify nearly all the possible words that third grade children should know. Individual test items may be taken from a huge group of items that include a broad range of items.

A test has content validity inbuilt in it. Items are selected in accordance with their compliance with the requirements of the test after a careful examination of the subject area.

In certain cases, where a test measures a trait which is difficult to define, an expert can rate the relevance of items. Since, each judge have their own opinion on their rating, two independent judges will rate the test separately. Items which are rated as highly relevant by both judges would be included in the final test.

2.6.2 Criterion-related Validity

Criterion related validity is the idea that a valid test should relate closely to other measure of the same theoretical concept. A valid test of intelligence should correlate highly with other intelligence test. If a test demonstrates effective predicting criterion or indicators of the construct, it is said to possess criterion – related validity. There are two different types of criterion validity-

2.6.2.1 Concurrent Validity

Its occurrence is found when criterion measures are achieved at the same time as the test scores. It reflects the degree to which the test scores estimate the individual's present status with regards to criterion. For instance, if a test measures anxiety, it would be said to have concurrent validity if it rightly reflects the current level of anxiety experienced by an individual. Concurrent evidence of test validity is usually desirable for achievement tests and diagnostic clinical test.

2.6.2.2 Predictive Validity

Predictive validity occurs when criterion measures are obtained at a time after the test. For example, aptitude tests are useful in identifying who will be more likely to succeed or fail in a particular subject. Predictive validity is part curly relevant for entrance examination and occupational test.

2.6.3 Construct Validity

Construct validity approach is complex than other forms of validity. Mc Burney and White (2007) defined construct validity as the property of a test that the measurement actually measures the constructs they are designed to measure. There are several ways to determine whether a test generate data that have construct validity.

i) The test should actually measure whatever theoretical construct it supposedly tests, and not something else. For example a test of leadership ability should not actually test extraversion.

ii) A test that has construct validity should measure what it intends to measure but not measure theoretically unrelated constructs. For example, a test of musical aptitude should not require too much reading ability.

iii) A test should prove useful in predicting results related to the theoretical concepts it is measuring. For example, a test of musical ability should predict who will benefit from taking music lessons, should differentiate groups who have chosen music as a career from those who haven't should relate to other tests of musical ability and so on.

There are two types of construct validity— 'convergent validity' and 'divergent validity' (or discriminant validity).

2.6.3.1 Convergent Validity

It means the extent to which a measure is correlated with other measure which is theoretically predicted to correlate with.

2.6.3.2 Discriminant Validity

This explains the extent to which the operationalisation is not correlated with other operationalisations that it theoretically should not be correlated with.

2.6.4 Face Validity

Face validity refers to what appears to measure superficially. It depends on the judgment of the researcher. Each question is scrutinised and modified until the researcher is satisfied that it is an accurate measure of the desired construct. The determination of face validity is based on the subjective opinion of the researcher.

Self Assessment Questions

Fill in the blanks

1) If a test measures what it purports to measure it is caleed …….

2) If a test is correlated against a criterion to be made available at the present time it is a type of validity known as……validity.

3) The property of a test that measurement actually measure the constructs they are design to measure are known as…….validity

4) A test should sample the range of behaviour represented by the theoretical concept being tested, is known as …….. validity.

5) ……. refers to what appears to measure superficially.

Answers: (1) Validity (2) Criterion Validity (3) Construct (4) Content (5) Face Validity

2.6.5 Internal Validity

Internal validity is the most fundamental type of validity because it concerns the logic of the relationships between the independent variable and dependent variable. This type of validity is an estimate of the degree to which inferences about causal relationship can be drawn, based on the measures employed and research design. Properly suited experimental techniques, where the effect of an independent variable upon the dependent one is observed under highly controlled conditions makes possible higher degree of internal validity.

2.6.5.1 Threats to Internal Validity

These include (i) confounding, (ii) selection bias, (iii) history, (iv) maturation, (v) repeated testing ,(vi) instrument change, (vii) regression toward the mean, (viii) mortality, (ix) diffusion, (x) compensatory rivalry, (xi) experimenter bias.

i) *Confounding:* Confounding error that occurs when the effects of two variables in an experiment cannot be separated, resulting in a confused interpretation of the results. Confounding is one of the biggest threat to validity in experimentation. The problem of confounding is particurly acute in research in which the experimenter cannot control the independent variable. When participants are selected according to presence or absence of a condition, subject variable can affect the results. Where a false relationship cannot be avoided, a rival hypothesis may be developed to the original cause and inference hypotheses.

ii) *Selection bias:* Any bias in selecting a group can undermine internal validity. Selection bias indicates the problem that occurs as a result of its existence at the pre-test differences between groups, may interact with the independent variable and thus influence the observed outcome and creates problems; examples would be gender, personality, mental capabilities, and physical abilities, motivation level and willingness to participate.

If at the time of selection, an uneven number of subjects to be tested have similar subject-related variables, there could be a threat to the internal validity, for instance, if two groups are formed i.e. experimental and control group, the subjects in the two groups are different with regards to independent variable but alike in one or more subject related variables. It would then be difficult for the researcher to identify if the difference between in the groups is the result of independent variable or subject related variable as well as randomisation of group assignment. It is not possible always as some significant variables may go unnoticed.

iii) *History*: Events outside the experiment or between repeated measures of dependent variables may influence participants' responses, attitudes and behaviour during process of experiment, like; natural disasters, political changes etc. In this condition, it becomes impossible to determine whether change in dependent variable is caused by independent variable or historical event.

iv) *Maturation:* Usually, it happens that subjects change during the course of an experiment or between measurements. For instance, in longitudinal studies young kids might grow up as a result of their experience, abilities or attitudes which are intended to be measured. Permanent changes [such as physical growth] and temporary changes [like fatigue and illness] may alter the way a subject would react to the independent variable. Thus, researcher may have trouble in ascertaining if the difference is caused by time or other variables.

v) *Repeated testing*: Participants may be driven to bias owing to repeated testing. Participants may remember correct answers or may be conditioned as a result of incessant administration of the test. Moreover, it also causes possibility of threat to internal validity.

vi) *Instrument change*: If any instrument is replaced/changed during process of experiment, then it may affect the internal validity as alternative explanation easily available.

vii) *Regression toward the mean*: During the experiment, if subjects are selected on the basis of extreme scores, then there are chances of occurrence of such an error. For example, when subjects with minimum mathematical abilities are chosen, at the end of the study if there is any improvement chances are that it would be due to regression towards the mean and not due to effectiveness of the course.

viii) *Mortality*: It should be kept in mind that there may be some participants who may have dropped out of the study before its completion. If dropping out of participants leads to relevant bias between groups, alternative explanation is possible that account for the observed differences.

ix) *Diffusion*: It might be observed that there will be a lack of differences between experimental and control groups if treatment effects spread from treatment groups to control groups. This, however, does not mean that, independent variable will have no effect or that there would not be a no relationship between dependent and independent variable.

x) *Compensatory rivalry/resentful demoralisation*: There will be a change in the behaviour of the subject if the control groups alter as a result of the study. For instance, control group participants may work extra hard to see that expected superiority of the experimental group is not demonstrated. Again, this does not imply that the independent variable created no effect or that there would be no relationship between dependent and independent variable. Vice-versa, changes in the dependent variable may only be effected due to a demoralised control group, working less hard or demotivated.

xi) *Experimenter bias:* Experimenter bias happens while experimenters, without any intention or reluctance, behave differently to the participants of control and experimental groups, that in turn, affect the results of the experiment. Experimental bias can be reduced by keeping the experimenter from knowing the condition in the experiment or its purpose and by standardising the procedure as much as possible.

2.6.6 External Validity

According to McBurney and White(2007), external validity concerns whether results of the research can be generalised to another situation, different subjects, settings, times and so on.

External validity lacks from the fact that experiments using human participants often employ small samples collected from a particular geographic location or with idiosyncratic features (e.g. volunteers). Because of this, it cannot be made sure that the conclusions drawn about cause-effect-relationships are actually applicable to the people in other geographic locations or in the absence of these features.

2.6.6.1 Threat to External Validity

How one may go wrong in making generalisations, is one of the major threats to external validity. Usually, generalisations are limited when the cause (i.e. independent variable) is dependent upon other factors; as a result, all the threats to external validity interact with the independent variable

a) *Aptitude-Treatment-Interaction*: The sample might have some features that may interact with the independent variable causing to limit generalisability,

for instance, conclusions drawn from comparative psychotherapy studies mostly use specific samples (example; volunteers, highly depressed, hardcore criminals).

b) *Situations:* All the situational factors, for example, treatment conditions, light, noise, location, experimenter, timing, scope and degree of measurement etc may limit generalisations.

c) *Pre-Test Effects*: When the cause-effect relationships can only be found out after the pre-tests are carried out, then, this also tends to limit the generality of the findings.

d) *Post-Test Effects:* When cause-effect relationships can only be explored after the post-tests are carried out, then this can also be a cause for limiting the generalisations of the findings.

e) *Rosenthal Effects*: When derivations drawn from the cause-consequence relationships cannot be generalised to other investigators or researchers.

Self Assessment Questions

1) Results can not be generalised to another situation or population in external Validity. T / F

2) Dropping out of some subjects before an experiment is completed causing a threat to internal validity. T / F

3) Any bias in selecting the groups can enhance the internal validity. T / F

4) Internal Validity concern the logic of relationship between the independent variable and dependent variable. T / F

5) Confounding error occurs when the effects of to variable in an experiment can not be separated. T / F

Answers: (1) F, (2) T, (3) F, (4) T, (5) T

2.7 LET US SUM UP

In psychological testing, reliability refers to the attribute of consistency of measurement. There are various types of reliability. The Pearson product-moment correlation coefficient can be used to gauge the consistency of psychological test scores. This form of reliability is referred to as test-retest reliability. Alternate-forms reliability is computed by correlating scores on two equivalent forms, administered in counterbalanced fashion to a large group of heterogeneous subjects. Internal consistency approaches to reliability include split-half reliability, in which scores on half tests are correlated with each other, and coefficient alpha, which can be thought of as the mean of all possible split-half coefficients. For tests that require examiner judgment for assignment of scores, inter scorer reliability is needed. Computing inters corer reliability is straightforward: A sample of tests is independently score by two or more examiners and scores for pairs of examiners are then correlated.

The validity of a test is the degree to which it measures what it claims to measure. A test is valid to the extent that inferences made from it are appropriate, meaningful, and useful. There are various kinds of validity – content validity

determine by the degree to which the question, task or items on a test are representative of the universe of behaviour the test was designed to sample. A test has face validity if it looks valid to test users, examiners, and especially the examinees. Criterion-related validity is demonstrated when a test is effective in predicting performance on an appropriate outcome measure. An investigation has internal validity is a cause-effect relationship actually exists between the independent and dependent variables. Confounding occurs when the effects of two independent variables in an experiment cannot be separately evaluated. External validity concerns whether the results of the research can be generalised to another situation: different subjects, settings, times, and so forth. Treats to the internal validity of an experiment include events outside the laboratory, maturation, effects of testing, regression effect, selection and mortality. Threats to external validity include problems arising from generalising to other subjects, other times, or other settings. Experimenter bias can be reduced by keeping the experiment from knowing he conditions in the experiment or its purpose and by standardising procedure as much as possible.

2.8 UNIT END QUESTIONS

1) Define reliability. Discuss any two methods of estimating reliability of test scores.
2) What is meant by internal consistency reliability. Discuss any two methods of assessing internal consistency reliability.
3) What are some problems associated with reliability assessed via the test-retest.
4) State the strengths and drawbacks of parallel forms reliability.
5) Write short notes on:
   K-R formula 20
   Spearman Brown formula
   Cronback alfa
6) Define validity and distinguish between reliability and validity.
7) Explain construct validity. How does it differ from content validity.
8) What is internal validity? Discuss various threats of internal validity.
9) What is external validity? Discuss various threats of external validity.
10) Write short notes on :
    Convergent and divergent validity
    Concurrent and predictive validity

2.9 GLOSSARY

Concurrent validity : a type of criterion-related validity in which the criterion measures are obtained at approximately the same time as the test scores.

Confounding : error that occurs when the effects of two variables in an experiment cannot be separated, resulting in a confused interpretation of he results.

Construct : a theoretical, tangible quality or trait in which individuals differ.

Construct validity : the property of a test that the measurements actually measure the constructs they are designed to measure, but no others.

Content validity : idea that a test should sample the range of behaviour represented by the theoretical concept being tested.

Criterion validity : idea that a test should correlate with other measures of the same theoretical construct.

Cronback alpha : an index of reliability that may be thought of as the mean of all possible split-half co-efficient, corrected by the Spearman-Brown formula.

External validity : how well the findings of an experiment generalise to other situations or populations.

Inter observer reliability : the typical degree of agreement between scores. Internal consistency: the degree to which the various items on a test are measures of the same thing.

Internal validity : extent to which a study provides evidence of a cause-effect relationship betwen the independent and dependent variables.

Kuder-Richardson formula 20 : an index of reliability that is relevant to the special case where each test item is scored 0 or 1 (example, right or wrong)

Maturation : a source of error in an experiment related to the amount of time between measurements.

Regression effect : regression effect tendency of subjects with extreme score on a first measure to score closer to the mean on a second testing.

Reliability : the property of consistency of a measurement that gives the same result on different occasions.

Spearman-Brown formula: a formula for adjusting split-half correlations so that they reflect the full length of a scale.

Split-half reliability : a form of reliability in which scores from the two halves of a test (e.g. even items versus odd items) are correlated with one another; the correlation is then adjusted for test length.

Test – retest reliability : the degree to which the same test score would be obtained on another occasion.

Validity : of a measurement the property of a measurement that tests what it is supposed to test.

2.10 SUGGESTED READINGS AND REFERENCES

Anastasi, Anne. (1988). *Psychological Testing* (6th edition.) London: Mac-Millan.

Freeman, F. S. (1971). *Theory and Practice of Psychological Testing*. New Delhi: Oxford (India).

References

Guilford, J.P. (1954). *Psychometric Methods.* New Delhi: Tata McGraw Hill.

Cronbach, L.(1951). *Coefficient Alpha and the Internal Structure of Tests*. Psychometrika, 16, 297-334.

Kaiser, H.F., & Michael, W.B. (1975). *Domain Validity and Gernalisability. Educational and Psychological Measurement,* 35, 31-35.

McBurney, D.H. & White, T. L. (2007) *Research Methods,* New Delhi; Akash Press.

Novick, M.R., & Lewis, C. (1967). *Coefficient Alpha and the Reliability of Composite Measurements.* Psychometrika, 32, 1-13.

Stodola, Q. and Stordahl, K. (1972*). Basic Educational Tests and Measurement*. New Delhi: Thomson (India).