UNIT 3 RELIABILITY AND VALIDITY

Structure

3.1 INTRODUCTION

Dear learners, in the first unit of this block, we discussed that measurement of social and psychological variables is a complex and demanding task. In urban development research, the common term for any type of measurement devise is ‘instrument’. Thus the instrument could be a test, scale, questionnaire, interview schedule etc. An important question that is often addressed is what is the reliability and validity of the measuring instrument? Therefore, the purpose of this unit is to make you understand the concept of reliability and validity and their interrelationship in urban development research.

After studying this unit you should be able to:

- discuss the meaning of reliability and methods of determining the reliability of measuring instruments.
- describe the meaning of validity, approaches and types of validating measuring instruments.
- differentiate the interrelationship between reliability and validity of measuring instruments.

3.2 RELIABILITY

In the context of development research, one of the most important criterions for the quality of measurement is reliability of the measuring instrument. A reliable person for instance, is one whose behavior is consistent, dependable and predictable – what (s)he will do tomorrow and next week will be consistent with what (s)he does today and what (s)he has done last week. An unreliable person is one whose behavior is much more variable and one can say (s)he is inconsistent.

The inherent aspects and synonyms of reliability are:

- dependability
- stability

- consistency
- predictability
- accuracy
- equivalence

3.2.1 What is Reliability of Measuring Instrument?

Reliability means consistency with which the instrument yields similar results. Reliability concerns the ability of different researchers to make the same observations of a given phenomenon if and when the observation is conducted using the same method(s) and procedure(s).

Stability and Equivalence Aspects of Reliability

Stability and equivalence deserves special attention among different aspects of reliability,

- The *stability* aspect is concerned with securing consistent results with repeated measurements of the same researcher and with the same instrument. We usually determine the degree of stability by comparing the results of repeated measurements.
- The *equivalence* aspect considers how much error may get introduced by different investigators or different samples of the items being studied. A good way to test for the equivalence of measurements by two investigators is to compare their observations of the same events.

3.2.2 How to Improve Reliability?

The reliability of measuring instruments can be improved by two ways.

i) By standardizing the conditions under which the measurement takes place i.e. we must ensure that external sources of variation such as boredom, fatigue etc., are minimized to the extent possible to improve the *stability* aspect.

ii) By carefully designing directions for measurement with no variation from group to group, by using trained and motivated persons to conduct the research and also by broadening the sample of items used to improve *equivalence* aspect.

Check Your Progress 1

Note: a) Use the spaces given below for your answers.

b) Check your answers with those given at the end of the unit.

1) What is the common name for any type of measurement device?

....

....

....

....

....

2) What do you mean by reliability?

....

....

....

....

....

3) Write the synonyms for reliability.

....

....

....

....

....

4) How can you improve the reliability of measuring instruments?

....

....

....

....

....

3.3 METHODS OF DETERMINING THE RELIABILITY

The three basic methods for establishing the reliability of empirical measurements are:

i) Test - Retest Method

ii) Alternative Form Method / Equivalent Form / Parallel Form

iii) Split-Half Method

3.3.1 Test - Retest Method

One of the easiest ways to estimate the reliability of empirical measurements is by the test - retest method in which the same test is given to the same people after a period of time. Two weeks to one month is commonly considered to be a suitable interval for many psychological tests. The reliability is equal to the correlation between the scores on the same test obtained at two points in time. If one obtains the same results on the two administrations of the test, then the test – retest reliability coefficient will be 1.00. But, invariably, the correlation of measurements across time will be less than perfect. This occurs because of the instability of measures taken at multiple points in time. For example, anxiety, motivation and interest may be lower during the second administration of the test simply because the individual is already familiar with it.

Advantages

- This method can be used when only one form of test is available.
- Test – retest correlation represent a naturally appealing procedure.

Limitations

- Researchers are often able to obtain only a measure of a phenomenon at a single point in time.
- Expensive to conduct test and retest and some time impractical as well.
- Memory effects lead to magnified reliability estimates. If the time interval between two measurements is short, the respondents will remember their early responses and will appear more consistent than they actually are.
- Require a great deal of participation by the respondents and sincerity, devotion by the research worker. Because, behaviour changes and personal characteristics may likely to influence the re-test as they are changing from day to day.
- The validity process of re-measurement may intensify difference in momentary factors such as anxiety, motivation etc.
- The interpretation of test-retest correlation is not necessary straightforward. A low correlation may not indicate low reliability, may instead signify that the underlying theoretical concept itself has changed.

  Example: The attitude of a person towards functioning of a public hospital may be very different before and after the person's visit. The true change in this example is interpreted as instability of attitude scale measurement.
- The longer the time interval between measurements, the more likely that the concept has changed.
- The process of measuring a phenomenon can induce change in the phenomenon itself. This process is called reactivity. In measuring a person's attitude at test, the person can be sensitized to the subject under investigation and demonstrate change during retest. Thus the test - retest correlation will be low.

3.3.2 Alternative Form Method/Equivalent Form/Parallel Form

The alternative form method which is also known as equivalent / parallel form is used extensively in education, extension and development research to estimate the reliability of all types of measuring instruments. It also requires two testing situations with the same people like test- retest method. But it differs from test – retest method on one very important regard i.e., the same test is not administered on the second testing, but an alternate form of the same test is administered. Thus two equivalent reading tests should contain reading passages and questions of the same difficulty. But the specific passages and questions should be different i.e., approach is different. It is recommended that the two forms be administered about two weeks apart, thus allowing for day –to- day fluctuations in the person to occur. The correlation between two forms will provide an appropriate reliability coefficient.

Advantages

- The use of two parallel tests forms provides a very sound basis for estimating the precision of a psychological or educational test
- Superior to test- retest method, because it reduces the memory related inflated reliability.

Limitations

- Basic limitation is the practical difficulty of constructing alternate forms of two tests that are parallel.
- Requires each person's time twice.
- To administer a secondary separate test is often likely to represent a somewhat burdensome demand upon available resources.

3.3.3 Split-Half Method

Split - half method is also a widely used method of testing reliability of measuring instrument for its internal consistency. In split-half method, a test is given and divided into halves and are scored separately, then the score of one half of test are compared to the score of the remaining half to test the reliability.

In split-half method, 1st-divide test into halves. The most commonly used way to do this would be to assign odd numbered items to one half of the test and even numbered items to the other, this is called, Odd-Even reliability. 2nd- Find the correlation of scores between the two halves by using the Pearson *r* formula. 3rd- Adjust or revaluate correlation using Spearman-Brown formula which increases the estimate reliability even more.

Spearman-Brown formula

$$r = \frac{2r}{1+r}$$

r = estimated correlation between two halves (Pearson r).

Advantages

- Both, the test – retest and alternative form methods require two test administrations with the same group of people. In contrast the split –half method can be conducted on one occasion.
- Split-half reliability is a useful measure when impractical or undesirable to assess reliability with two tests or to have two test administrations because of limited time or money.

Limitations

- Alternate ways of splitting the items results in different reliability estimates even though the same items are administered to the same individuals at the same time.

  Example: The correlation between the first and second halves of the test would be different from the correlation between odd and even items.

Major Limitations of Reliability Estimating Methods

Test-retest method: Experience in the first testing usually will influence responses in the second testing.

Alternative form method: It can be quite difficult to construct alternative forms of a test that are parallel.

Split-half method: The correlation between the halves will differ depending on how the total number of items is divided into halves.

Alternate form method provide excellent estimate of reliability in spite of its limitation of constructing two forms of a test. To over come this limitation, it is recommended that, randomly divide a large collection of items in half to have two test administrations.

Check Your Progress 2

Note: a) Use the spaces given below for your answers.

b) Check your answers with those given at the end of the unit.

1) Write the three basic methods of determining the reliability?

....

....

....

....

....

2) Write the major limitations in reliability determining methods?

....

....

....

....

....

3.4 VALIDITY

According to Goode and Hatt, a measuring instrument (scale, test etc) possesses validity when it actually measures what it claims to measure. The subject of validity is complex and very important in development research because it is in this more than anywhere else, that the nature of reality is questioned. It is possible to study reliability without inquiring into the nature and meaning of one's variable. While measuring certain physical characteristics and relatively simpler attributes of persons, validity is no great problem. For example, the anthropometrics measurements of a pre-school child i.e., head and chest circumference can be measured by a measuring instrument having standard number of centimeters or inches. The weight of the child can be measured in pounds and kilograms. On the other hand, if a child development extension professional wish to study the relation between malnutrition and intellectual development of pre-school children, there are neither any rule to measure the degree of malnutrition nor there any scales or clear cut physical attributes to measure intellectual development. It is necessary in such cases to invent indirect means to measure these characteristics. These means are often so indirect that the validity of the measurement and its product is doubtful.

Validity of Measuring Instrument or Measuring Phenomenon?

We defined validity as the extent to which any measuring instrument measures what it is intended to measure. But, strictly speaking, one validates not a measuring instrument, but an interpretation of data arising from a specified procedure. This distinction is central to validation, because it is quite possible for a measuring instrument to be relatively valid for measuring one kind of phenomenon but entirely invalid for assessing other phenomenon. Thus, one validates not the measuring instrument itself, but the measuring instrument in relation to the purpose for which it is being used.

3.4.1 Approaches to Validation of Measuring Instrument

Every measuring instrument, to be useful, must have some indication of validity. There are four approaches to validation of measuring instruments:

i) Logical validity / Face validity

ii) Jury opinion

iii) Known-group

iv) Independent criteria

i) Logical Validity

This is one of the most commonly used methods. It refers to either theoretical or commonsense analysis, which concludes simply that, the items, being what they, the nature of the continuum cannot be other than it is stated to be. Logical validation or face validity as it is sometimes called is almost always used because it automatically springs from the careful definition of the continuum and the selection of items. Such measure, which focuses directly on behavior of the kind in which the tester is interested, is said to have logic / face validity.

Example: The reading speed is measured by computing how much of a passage person reads with comprehension in a given time and the ability to solve arithmetic problems by success in solving a sample of such problems.

Limitation

- It is not wise to rely on logical and commonsense validation alone. Such claims for validity can at best be merely plausible and never definite. More than logical validity, it is required to render satisfactory use of a measuring instrument.

ii) Jury Opinion

This is an extension of the method of logical validation, except that in this case the confirmation of the logic is secured from a group of persons who would be considered experts in the field in which the measuring instrument is being used.

Example: If a scale to measure mental retardation of pre-school children is constructed, psychologists, psychiatrists, pediatrician, clinical psychologists, social worker and teachers might constitute the jury to determine the validity of the scale.

Limitation

- Experts too are human and nothing but logical validity can result from this approach. Therefore, jury validation can be considered only slightly superior to logical validation.

iii) Known-Group

This technique is a variant of the jury procedure. In this case, the validity is implied from the known attitudes and other characteristics of analytical groups, however, rather than from their specific expertness. Thus, if a scale were being devised for the purpose of measuring the attitudes of people towards the Church, the questions could be tested by administering them to one group known to attend Church, to be active in Church activities and otherwise to give evidence of a favorable attitude towards this institution. These answers would be compared with those from a group known not to attend Church and also known to oppose the Church. If the scale failed to discriminate between the two groups it could not be considered to measure this attitude with validity. The known group technique of validation is frequently used and should not be discarded for falling somewhat short of perfection.

Limitation

- There might be other differences between the groups in addition to their known behavior with regard to religion, which might account for the differences in the scale scores.

Example: Differences in age, socioeconomic status, ethnic background etc.

- Further perhaps the known behavior under the study might be associated with a differential inclination to agree or disagree on a question in general. Hence careful use of the known group technique should be made.

iv) Independent Criteria

This is an ideal technique abstractly speaking but its application is usually difficult. There are four qualities desired in a criterion measure. In order of their importance they are :

a) *Relevance:* We judge a criterion to be relevant the extent to that standing on the criterion measure corresponds to the scores on scale.

b) *Freedom from bias :* By this we mean that the measure should be one on which each person has the same opportunity to make a good score. Example of biasing factors are such things as variation in the quality of equipment or conditions of work for a factory worker, a variation in the quality of teaching received by studying in different classes.

c) *Reliability :* If the criterion score is one that jumps around from day to day, so that the person who shows high job performance one week may show low job performance the next or who receives a high rating from one supervisor gets a low rating from another, then there is no possibility of finding a test that will predict that score. A measure that is completely unstable by itself cannot be predicted by anything else.

d) *Availability:* Finally, in the choice of a criterion measure we always encounter practical problems of convenience and availability. How long will we have to wait to get a criterion score for each individual? How much is it going to cost? Any choice of a criterion measure must make a practical limit to account.

However, when the independent criteria are good validation, it becomes a powerful tool and is perhaps the most effective of all techniques of validation.

Check Your Progress 3

Note: a) Use the spaces given below for your answers.

b) Check your answers with those given at the end of the unit.

1) Do you agree that 'one validates not the measuring instrument, but the purpose for which it is being used'? Write your agreement or disagreement.

....

....

....

....

....

2) Name the four approaches to validation of measuring instrument.

....

....

....

....

....

3.5 TYPES OF VALIDITY

The most important classification of types of validity is that prepared by a Joint Committee of American Psychological Association, the American Educational Research Association and the National Council on measurements used in education. There are three types of validity:

i) Content validity

ii) Criterion validity (Predictive validity and Concurrent validity)

iii) Construct validity.

3.5.1 Content Validity

The term content validity is used, since the analysis is largely in terms of the content.

Content validity is the representative ness or sampling adequacy of the content. Consider a test that has been designed to measure competence in using the English language. How can we tell how well the test in fact measures that achievement? First we must reach at some agreement as to the skills and knowledge that comprise correct and effective use of English, and that have been the objectives of language instruction. Then we must examine the test to see what skills, knowledge and understanding it calls for. Finally, we must match the analysis the test content against of course content and instrumental objectives, and see how well the former represents the latter. If the test represents the objectives, which are the accepted goals for the course, then the test is valid for use.

3.5.2 Criterion Validity

The two types of criterion validities are predictive validity and concurrent validity. They are much alike and with some exceptions, they can be considered the same, because they differ only in the time dimension. They are characterized by prediction and by checking the measuring instrument either now or in future against some outcome.

Example: A test that help researcher / teacher to distinguish between students who can study by themselves after attending the class and those who are in need of extra and special coaching, is said to have concurrent validity. The test distinguishes individually who differ in their present status. On the other hand, the investigator may wish to predict the percentage of passes during the final examination for that particular period. The adequacy of the test for distinguishing individuals who differ in the future may be called as predictive validity.

Predictive Validity Vs. Concurrent Validity

Predictive validity concerns a future criterion which is correlated with the relevant measure.

Example: Tests used for selection purposes in different occupations are, by nature, concerned with predictive validity. Thus a test used to screen applications for the post of ‘health extension and development workers’ could be validated by correlating their test scores with future performance in fulfilling the duties associated with health extension work.

Concurrent criterion is assessed by correlating a measure and the criterion at the same point in time.

Example: A verbal report of voting behaviour could be correlated with participation in an election, as revealed by official voting records.

3.5.3 Construct Validity

Both content and criterion validities have limited usefulness for assessing the validity of empirical measures of theoretical concepts employed in extension and development studies. In this context, construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured. Examination of construct validity involves validation not only of the measuring instrument but of the theory underlying it. If the predictions are not supported, the investigator may have no clear guide as to whether the shortcoming is in the measuring instrument or in the theory.

Construct validation involves three distinct steps.

a) specify the theoretical relationship between the concepts themselves

b) examine the empirical relationship between the measures of the concepts

c) interpret the empirical evidence in terms of how it clarifies the construct validity of the particular measure.

Indeed strictly speaking, it is impossible to validate a measure of a concept in this sense unless there is a theoretical network that surrounds the concept.

Check Your Progress 4

Note: a) Use the spaces given below for your answers.

b) Check your answers with those given at the end of the unit.

1) Name the three types of validity.

....

....

....

....

....

2) Write the major difference between predictive and concurrent validities.

....

....

....

....

....

3.6 RELIABILITY OR VALIDITY - WHICH IS MORE IMPORTANT?

The real difference between reliability and validity is mostly a matter of definition. Reliability estimates the consistency of your measurement, or more simply the degree to which an instrument measures the same way each time it is used in under the same conditions with the same subjects. Validity, on the other hand, involves the degree to which you are measuring what you are supposed to, more simply, the accuracy of your measurement. Reliability refers to the consistency or stability of the test scores; validity refers to the accuracy of the inferences or interpretations you make from the test scores. Note also that reliability is a necessary but not sufficient condition for validity (i.e., you can have reliability without validity, but in order to obtain validity you must have reliability). In this context, validity is more important than reliability because if an instrument does not accurately measure what it is supposed to, there is no reason to use it even if it measures consistently (reliably).

Let us examine the following three principles to understand the relationship between reliability and validity and to answer the question which is more important.

a) A test with high reliability may have low validity.

b) In the evaluation of measuring instruments, validity is more important than reliability.

c) To be useful, a measuring instrument must be both reasonably valid and reasonably reliable.

Consider the following four figures to understand easily the complex relationship between reliability and validity (Source: Patten, 2005).

In Fig. 3.1, the gun is aimed in a valid direction towards the target, and all the shots are consistently directed, indicating that they are reliable.

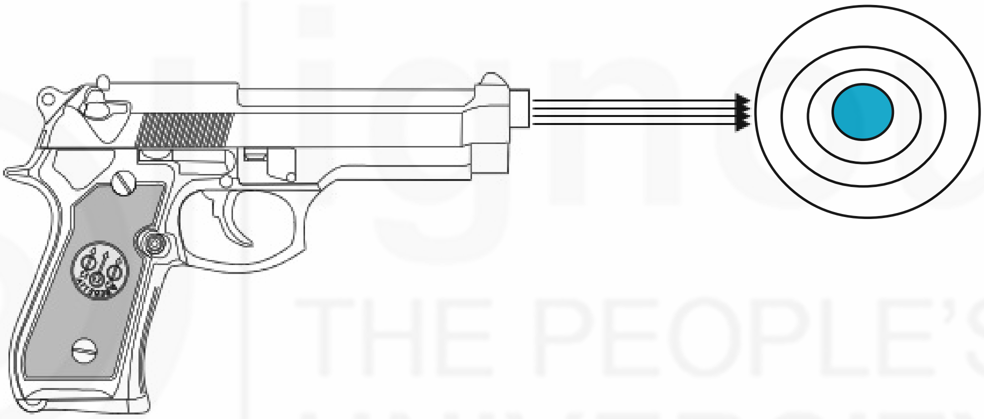

Fig. 3.1: Reliable and valid

In Fig. 3.2, the gun is also aimed in the direction of the target, but the shots are widely scattered, indicating low consistency or reliability. Thus the poor reliability undermines an attempt to achieve validity.

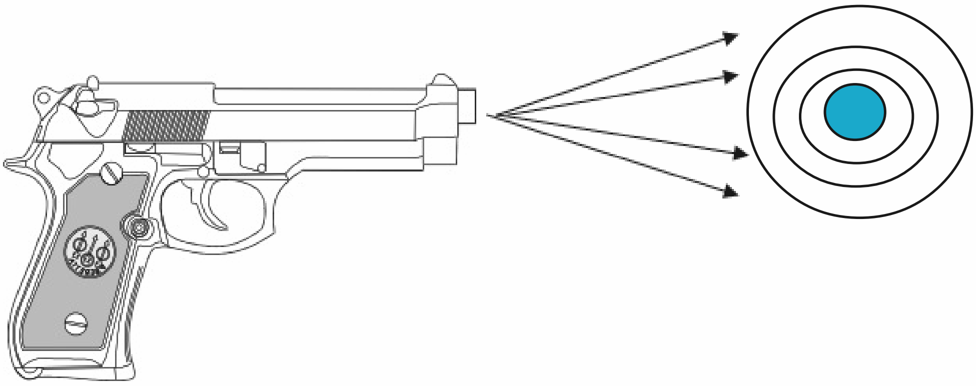

Fig. 3.2: Unreliable which undermines the valid aim of the gun – Not usefull

In Fig. 3.3, the gun is not pointed at the target, making it invalid, but there is great consistency in the shots in one direction, indicating that it is reliable (In a sense, it is very reliably invalid).

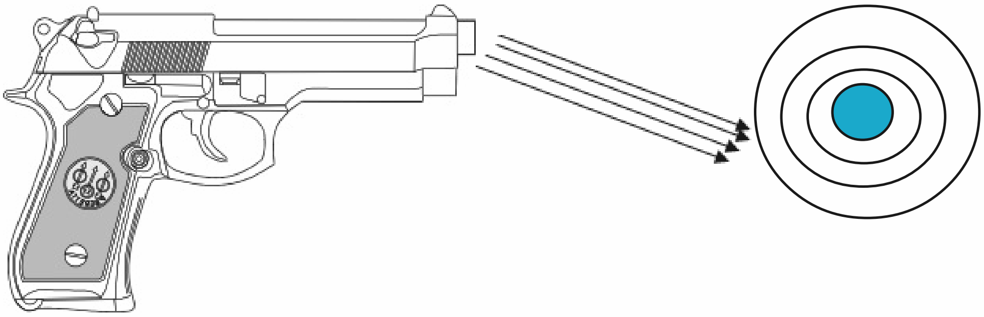

Fig. 3.3: Reliable but invalid – Not useful

In Fig. 3.4, the gun is not pointed at the target making it invalid, and the lack of consistency in the direction of the shots indicates its poor reliability.

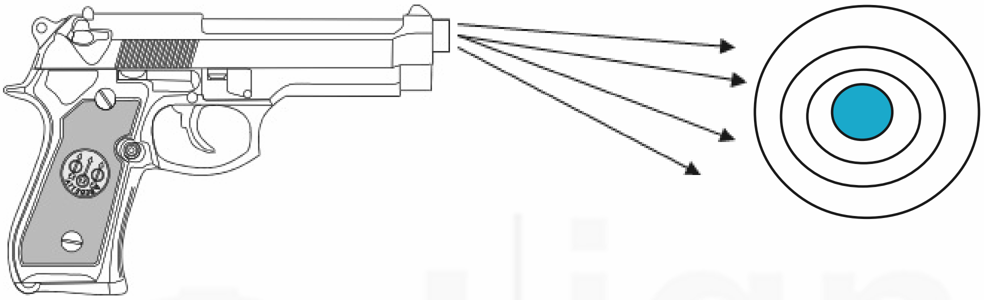

Fig. 3.4: Unreliable and invalid – Not useful

We may arrive at a conclusion that Fig. 3.1 represents the ideal in measurement. However, due to the limitations of measuring instruments in extension and development studies / social and behavioural sciences, we should not expect perfect reliability and validities. The direction of gun should be off at least a small amount - indicating a less than perfect validity. We also should expect some scatter in the shots, indicating less- than - perfect reliability. Clearly, our first priority should be to point the gun in the correct general direction, which promotes validity and then work on increasing reliability. This indicates that both reliability and validity are important in measurement, but among them validity is more important.

Check Your Progress 5

Note: a) Use the spaces given below for your answers.

b) Check your answers with those given at the end of the unit.

1) Among reliability and validity, which is more important and why?

....

....

....

....

....

....

....

3.7 LET US SUM UP

In this unit we started by discussing the meaning of reliability and understood that reliability means consistency with which the instrument yields similar results. Later we highlighted that, among different aspects of reliability, two aspects i.e. stability and equivalence deserves special attention. We discussed the three important methods for assessing the reliability of measuring instruments. For the limitations mentioned in the discussion, neither test- retest method nor split-half method is recommended for estimating reliability. In contrast, the alternative form method is excellent for estimating reliability.

In the second part of the unit we have discussed the concept of validity and understood a measuring instrument possesses validity when it actually measures what it claims to measure. We examined the four approaches of validation of measuring instruments: logical validity / face validity, jury opinion, known-group and independent criteria. We also discussed the three types of validities and found that both content and criterion validities have limited usefulness in assessing the quality of development measures. In contrast, construct validation has generalized applicability in the extension and development research by placing the measure in theoretical context.

In the third and final part of the unit, we discussed, the relationship between reliability and validity and concluded that both reliability and validity are important in measurement, but among them validity is more important.

3.8 KEYWORDS

Reliability : Reliability means consistency with which the instrument yields similar results.

Validity : Validity is the ability of a measuring instrument to actually measure what it claims to measure.

Logical Validity : It refers to either theoretical or commonsense analysis, which concludes simply that, the items, being what they, the nature of the continuum cannot be other than it is stated to be.

Jury Opinion : The confirmation of the logic is secured from a group of persons who would be considered experts in the field in which the measuring instrument is being used.

Known-Group : The validity is implied from the known attitudes and other characteristics of analytical groups, however, rather than from their specific expertness.

Content Validity : Content validity is the representativeness or sampling adequacy of the content.

Predictive Validity : It concerns a future criterion which is correlated with the relevant measure.

Concurrent Validity : It is assessed by correlating a measure and the criterion at the same point in time.

Construct Validity : Construct validity involves validation of not only the measuring instrument but of the theory underlying it.

3.9 REFERENCES AND SELECTED READINGS

Carmines, E.G., and Zeller, R.A., (1979). Reliability and Validity Assessment. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 17. Beverly Hills and London: Sage Publications.

Kothari, C.R. (2004). Research Methodology – Methods and Techniques, Second Revised Edition, New Age International Publishers, New Delhi.

Patten M.L. (2005). Understanding Research Methods – An Overview of the Essentials. Pyrczak Publishing, USA.

3.10 CHECK YOUR PROGRESS – POSSIBLE ANSWERS

Check Your Progress 1

1) The common name for any type of measurement device is 'instrument'.

2) Reliability estimates the consistency of our measurement, or more simply the degree to which an instrument measures the same way each time it is used in under the same conditions with the same subjects.

3) The synonyms for reliability are : dependability; stability; consistency; predictability; accuracy and equivalence .

4) The reliability of measuring instruments can be improved by (i) by standardizing the conditions under which the measurement takes place and (ii) by carefully designing directions for measurement with no variation from group to group, by using trained and motivated persons to conduct the research and also by broadening the sample of items.

Check Your Progress 2

1) The three basic methods of determining the reliability are : test – retest method; alternative form method and split-half method.

2) The major defect of test-retest method is that experience in the first testing usually will influence responses in the second testing. The practical limitation of alternative form method is that it can be quite difficult to construct alternative forms of a test that are parallel. The major problem with the split-half method approach is that the correlation between the halves will differ depending on how the total number of items is divided into halves.

Check Your Progress 3

1) Yes. I agree with the statement 'one validates not the measuring instrument, but the purpose for which it is being used' because it is quite possible for a measuring instrument to be relatively valid for measuring one kind of phenomenon, but entirely invalid for assessing other phenomenon.

2) The four approaches to validation of measuring instrument are: logical validity / face validity; jury opinion; known-group and; independent criteria.

Check Your Progress 4

1) The three types of validity are : Content validity; Criterion validity (Predictive validity and Concurrent validity) and Construct validity.

2) Predictive validity concerns a future criterion which is correlated with the relevant measure. Concurrent criterion is assessed by correlating a measure and the criterion at the same point in time.

Check Your Progress 5

1) Validity is more important than reliability because if an instrument does not accurately measure what it is supposed to, there is no reason to use it even if it measures consistently (reliably).