CHAPTER IV

RELIABILITY AND VALIDITY OF THE TEST

CONTENTS

4.0 INTRODUCTION

4.1 RELIABILITY

4.1.1 The Concept
4.1.2 Methods of Estimating Reliability
4.1.3 Reliability of the Present Test
4.1.4 The Index of Reliability

4.2 VALIDITY

4.2.1 The Concept
4.2.2 Types of Validity
4.2.3 Validity of the Present Test
4.0 INTRODUCTION

Different types of tests are constructed and administered to serve a variety of purposes. Each type differs from the other type of tests as it has a specific use. Accordingly, achievement tests are used to measure progress towards instructional objectives, aptitude tests are used to predict success in future activities, appraisals of personal-social development may be obtained to better understanding the individuals and so on. Thus different tests may evaluate different areas of behaviour and hence, they may differ in some of their special features. Yet, it is necessary for all the various tests to possess certain common characteristics. The most essential of the common characteristics are reliability and validity.

4.1 RELIABILITY

4.1.1 The Concept

Reliability is concerned with the accuracy and precision of the measurement results. Reliability suggests how consistent the evaluation results are from one measurement to another. The measurement results are expected to
be reasonably consistent over different occasions or over different samples of the same behaviour. Variation in the test scores is supposed to occur due to the individual differences of the tastees and to the instrument itself along with the factors other than the quality being measured. Reliability shows the extent to which the obtained test scores are free from internal defects inherent in the test items, that may produce error of measurement.

Reliability refers to the extent to which the test is internally consistent. All the items of the test should measure the same quality and hence yield consistency of results throughout the test when administered once. Secondly it refers to the extent to which the measuring device yields consistent results when the measurement procedure is repeated under similar conditions i.e. stability of the test scores.

Reliability is statistical in nature. The test has to be administered one or more times, to the group appropriate for that and the consistency of results has to be determined. The consistency of measurement results can be expressed in terms of shift in the relative standing of individuals in the group, which is reported by a correlation coefficient, called a reliability coefficient. It can also be expressed in terms of the amount of variation in
any individual's score that may be expected, this is expressed by the standard error of measurement.

In determining reliability of a test, two sets of measures under identical conditions are required. Since the conditions under which evaluation data are obtained are never identical, it is hardly possible to get to such measures. Hence, some substitute for such ideal procedure has to be adopted. Several methods of estimating reliability are introduced in that context.

4.1.2 Methods of Estimating Reliability

(a) Test-Retest Method

To estimate the reliability of a test by test-retest method, the same test is administered twice over the same sample with reasonable time interval between the two administrations. The scores so obtained are correlated and the coefficient of correlation is considered as the measure of reliability, as it indicates the stability of test scores over a given period of time. If the individuals scoring high on one administration of the test tend to score high on the other administration of the test, and the remaining individuals tend to stay in their same relative positions on both the administrations of the test, then
the results can be said to be highly stable. Such stability would result into a large correlation coefficient. The important factor in interpreting measures of reliability by this method is the time interval between the two administrations. If the time is short, say one or two days, there is a possibility that the constancy of results may be inflated because of the pupils' memory. On the other hand, if the time interval is too long, say about a year or more, the results of retesting would be affected not only by the instability of the testing procedure but also by the developmental changes of the individuals from time to time. The most preferable time interval between the testings largely depends upon the use to be made of the results. Generally, a period of two to six weeks is preferred.

(b) Retest Using Equivalent Forms

Two forms of the test having same set of specifications (test content, difficulty etc), are constructed independently. Such forms are known as equivalent forms, parallel forms or alternate forms of the test. The two forms of the test are administered to the same group of individuals in close succession. Thus two sets of measures are obtained for the same characteristics to be measured. The test scores are then correlated. The correlation
coefficient indicates the degree to which, both the forms of the test measure, the same aspects of behaviour. This reliability coefficient reflects the extent to which the test represents an adequate sample of the characteristics being measured, but it does not refer to the stability of scores for the individuals.

In estimating reliability by equivalent forms, one need not bother about the interval of time between two testings. Some how the labourious procedure of constructing two equivalent forms of the same test restricts its use. Sometimes equivalent forms are used keeping a time interval between the administrations of the two forms of a test. In such a case the reliability coefficient provides measure of stability as well as equivalence. It takes into consideration all the three sources of variation viz., the stability of testing procedure, the constancy of the pupil characteristic being measured and the representativeness of the sample of behaviour included in the test.

(c) Split-half Method

The reliability of test scores can also be estimated by a single administration of a single test form. The test is administered to the group in the usual manner and then for scoring purpose, it is divided into two halves.
Normally the odd-numbered items and even numbered items are grouped to get two sets of measures for the same individuals. These two scores are then correlated. The correlation coefficient indicates the degree to which the two halves of the test are equivalent. It provides a measure of internal consistency. A reliability coefficient is calculated for two half-tests. To estimate the reliability of the scores based on the full-length test, a formula known as Spearman-Brown formula is applied.

\[
\text{Reliability on full test} = \frac{2 \times \text{Reliability on half test}}{1 + \text{Reliability on half test}}
\]

The reliability coefficient obtained by split-half method indicates the extent to which the sample of the test items is an adequate sample of the content being measured. However, it tells nothing about changes in the individual from one time to another.

(d) **Kuder-Richardson Method**

Kuder-Richardson developed a formula which is useful to estimate the reliability of test scores from a single administration of a single form of a test. The formula provides a measure of internal consistency. The formula is
\[
\rho = \left(\frac{n}{n-1}\right) \left(\frac{S^2 - \sum pq}{S^2}\right)
\]

Where \(\rho\) = reliability coefficient

\(n\) = no. of items in the test

\(s\) = SD of the test scores

\(p\) = the percent passing a particular item

\(q\) = the percent failing the same item

As the above formula is based on the proportion of persons passing each item and the standard deviation of the total scores, the computation is rather cumbersome.

In order to make the computation simpler Kuder-Richardson gave a modified version of the formula.

\[
\text{Reliability coefficient} = \frac{n}{n-1} \left(1 - \frac{M(n-M)}{nS^2}\right)
\]

Where \(n\) = no. of items in the test

\(M\) = Mean of the test scores

\(S\) = SD of test scores

While estimating reliability by Kuder-Richardson formula it is assumed that the test items are homogeneous.

(e) The Standard Error of Measurement

If an individual is tested repeatedly on the same
test, some variation in his test scores is likely to occur. The amount of variation in test scores would be directly related to the reliability of testing procedure. Large variation indicate low reliability while small variation indicate high reliability. It is not practical to administer the same test, on the same sample again and again. Yet it is possible to estimate the expected amount of variation in test scores. Such estimate is called the standard error of measurement. If an individual is tested repeatedly under identical conditions 68% of scores are likely to fall within 1 standard error, 95% scores would fall within 2 standard errors and 99% test scores would fall within 3 standard errors. Standard error of measurement on its conversion into reliability establishes the fact that high reliability coefficients are associated with small standard errors in specific test scores and low reliability coefficients are associated with high errors. Two special advantages of standard error of measurement are that the estimates are in the same units as the test scores and, it is likely to remain constant for different groups. Standard error of measurement cannot be used to compare two tests having different types of scores.

4.1.3 Reliability of the Present Test

It was necessary to estimate the reliability of the
present test, as it is an integral part of standardization procedure. From different methods of estimating reliability of a test, test-retest method was found to be the most appropriate one. Looking to the nature of the present test, split-half method or use of parallel forms was not practical. Application of Kuder-Richardson formula was also not possible.

(a) Estimation of Reliability by Test-retest Method

From among the students who were already administered the test once, about 100 students were selected for retest. In selecting these students, any specific method of sampling was not thought to be necessary. One of the schools was selected by the researcher and the test was administered in the same classes, where it was administered once. The test was readministered after about 3 weeks, under identical conditions as far as possible. Time interval of three weeks was considered long enough to avoid the effect of memory and not too long such that the development of the students may cause variation in the test scores. The researcher went to the school and administered the test in the same way as before. All possible care was taken to maintain identical testing procedure. Another set of test scores for the same students was obtained. The product-moment
correlation was calculated. The scatter diagram is presented in Appendix 3. The correlation coefficient is found to be 0.62.

A good estimate of the effect of lengthening or repeating a test can be obtained using the Spearman-Brown prophecy formula:

\[ r_{nn} = \frac{n}{1+(n-1)} r_{ii} \]

Where \( r_{nn} \) = the correlation between \( n \) forms of a test and \( n \) alternate forms.

\[ r_{ii} \] = the reliability coefficient of test 1.

Taking \( n = 2 \) for the test

\[ r_{22} = \frac{2 \times 0.62}{1 + 0.62} \]

\[ = 0.76 \]

(b) The Standard Error of Measurement

Standard error of measurement scores for the present test was calculated, since it is considered to be a better way of expressing the reliability. The standard error of measurement takes into account the variability within the group as well as the self correlation of the test. The following formula was applied for the calculation:
\[ \sigma_{sc} = \sigma_1 \times \sqrt{1 - r_{ii}} \]

Where, \( \sigma_{sc} \) = Standard error of an obtained score i.e. standard error of measurement.

\( \sigma_1 \) = Standard Deviation of test scores

\( r_{ii} \) = reliability coefficient of the test.

SE of measurement was found to be 2.81.

4.1.4 The Index of Reliability

The mean of the scores obtained by administering repeatedly the same test on the same individual under identical conditions is known as the 'true score' of that individual. The correlation between a set of obtained scores and their corresponding true scores is expressed as

\[ r_{lc} = \sqrt{r_{ii}} \]

Where \( r_{lc} \) = Correlation of obtained and true scores

\( r_{ii} \) = reliability coefficient of the test.

The coefficient \( r_{lc} \) is known as the index of reliability. The index of reliability gives the maximum correlation, which the given test is capable of yielding in its present form.
For the present test, index of reliability is found to be 0.78 which is satisfactorily high.

4.2 VALIDITY

4.2.1 The Concept

In selection or construction of an evaluation instrument, an important consideration would be to inquire, whether the test results serve the specific purpose for which they are intended. The evaluation instrument should measure what it intends to measure. It should measure all of what is expected to measure and nothing but what is expected 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 acceptable criteria". Any measurement device need to be evaluated against criteria, regarded by the experts as the best evidence of the traits to be measured by it. Validity is always related to the results of the test or evaluation instrument. It is considered in terms of specific degree such as high validity, moderate validity or low validity. Evaluation results are not simply valid, the degree of validity differs for different interpretations to be made. Validity is always concerned with the specific use to be made of evaluation results and with the soundness of proposed
interpretations. Selection of satisfactory validation criteria and demonstration of an appropriate degree of validity are fundamental in test construction.

4.2.2 Types of Validity

In educational and psychological measurement three basic types of validity are generally used. The three types are (a) Content Validity (b) Criterion related Validity and (c) Construct validity.

(a) Content Validity

This type of validity is concerned with the content of the measurement tool. The term content is assumed to include subject matter content and instructional objectives. Content validity of a test indicates the extent to which the test measures a representative sample of the subject-matter content and the behavioural changes under consideration.

If a test appears, on superficial examination, to be relevant as regarding the subject matter content and the instructional objectives, it is considered to have face validity. Face validity confirms only that the test looks like an appropriate measurement instrument. For most of
the psychological tests the test constructor satisfies himself with the face validity but, it is not desirable to accept it as the substitute for content validity for achievement tests. For establishing content validity of the achievement tests the major topics of subject-matter and major behavioural changes are listed. They are assigned weightage based on their relative importance. A table of specifications is prepared, showing the relative emphasis of each subject matter content and each type of behavioural change. The test is constructed or selected in accordance with the table of specifications. As this procedure includes analysis of the test items it is also known as Rational Validity or Logical Validity.

(b) **Criterion Related Validity**

Sometimes the test scores are to be used to predict future performance or to get an estimate of present performance on some other valued measurement. When it is related to future performance, it is called predictive validity. In case of estimating present status, it is known as concurrent validity. The test administered first is called predictor and it predicts the performance on some of the test i.e. criterion. The second measure of performance may be obtained simultaneously or in future, depending upon the use to be made. A rank-difference correlation is computed between the two sets of measures. The
correlation coefficient (rho) indicates how far the scores on one test (predictor) are predictive of the scores on the second test (criterion). The accuracy of prediction increases as the correlation coefficient becomes larger. A major problem in estimating criterion-related validity is that of having a satisfactory criterion of success.

(c) Construct Validity

Construct validity is useful when test performance is to be interpreted in terms of some psychological trait or quality called construct. In order to explain any aspect of behaviour, some construct is assumed to be existing. In determining the construct validity certain steps have to be followed. The steps include -

1. identifying the constructs influencing test performance;
2. framing hypotheses regarding test performance;
3. verifying the hypotheses by logical and empirical means.

The evidence of construct validity may be obtained by any of the following procedures:
1. Analysis of the mental process required by the test items.

2. Comparison of the scores of known groups.

3. Comparison of scores before and after some particular treatment.

4. Correlation with other tests.

Construct validity can be determined for all types of evaluation procedures. The aim of establishing construct validity is to identify the nature and strength of all factors influencing performance on the test.

4.2.3 Validity of the Present Test

The Thematic Apperception Test, being a projective technique, deals with the inner, concealed aspects of personality dynamics which are revealed only indirectly by the interpretation of subtle signs in the individuals' interpretation of what he perceives. Hence content validity cannot be determined for a TAT. The projective techniques are concerned with what the responses are presumed to signify. While inquiring into the validity of the projective techniques, the primary focus should be upon, their validity as a measurement of psychological constructs.
The validity of the present test was established by two methods, at two different stages of standardization procedure.

(a) **Face Validity of the Present Test**

The picture cues, prepared for the present TAT were given to experts to determine the face validity. Twenty experts working in the field of education measurement were given the set of pictures and were requested to evaluate, whether the pictures have any cue depicting concern for excellence. The experts opined that the pictures were appropriate for the purpose. The experts examined the pictures and gave their opinion in favour of the instrument being prepared. Thus, face-validity of the TAT has been considered satisfactory.

(b) **Construc Validity of the TAT**

As there is no adequate single method of establishing construct validity, several methods are used to obtain evidence for construct validity. One of the methods is to correlate the test scores with the scores on some other test measuring the same psychological trait or quality. The scores of any particular test can be expected to correlate fairly well with the scores of other test which is supposed to measure the same ability or trait. For establishing
validity of the present test this method was adopted. The researcher searched for some other instrument to measure concern for excellence, but no such tool is available for Gujarati students. The concept of concern for excellence is having similarity with the concept of achievement motivation. A value judgement inventory for Gujarati students (Vide Appendix K) prepared by C.C. Pathak was available. The inventory measures achievement value which is required for the concern for excellence. The researcher visited one of the schools where TAT was administered. A sample of 100 students of standards VIII, IX and X from among those who were administered the TAT, was selected. These students were administered the Value Judgement Inventory. Achievement value score for each individual was obtained. Two sets of measures for the same group of students were prepared. The scores on the TAT and the scores for achievement value were correlated using product-moment method. The scatter diagram is presented in Appendix F. The correlation coefficient was calculated. It is found to be 0.69, which suggests that the present test is satisfactorily valid.