CHAPTER - X

THE RELIABILITY AND VALIDITY OF THE TEST

1. Reliability of a test
2. Definitions of the term 'reliability'
3. Methods of estimating reliability
   (1) The test-retest method
   (2) The parallel-form method
   (3) The split-half method
   (4) The method of rational equivalence
4. Measuring the reliability of the present test
5. The reliability studies
6. The concept of validity
7. Types of validity
8. Approaches selected for finding validity.

****
Measuring devices are used in almost all sciences. These devices must fulfil certain conditions, if they are to be of service. They must indicate such results as are authentic, dependable and serviceable within the limits specified. Whenever something is measured, it is liable to cause error in results. This is true with the measurement of physical, biological and behavioural sciences. Mursell (1) has indicated the following four classes of possible errors which can affect any instrument of measurement whatsoever.

(1) All measurement is subject to constant error.

(2) All measurement is subject to variable errors.

(3) All measurement is subject to personal errors.

(4) All measurement is subject to errors of interpretation.

This is true of psychological tests as well, to a certain extent. If we administered
the tests more than once, variations of results will be obtained with the same psychological tests. They are sometimes considered, due to chance factors but sometimes they are due to actual differences among the persons taking the tests and they affect the results. Generally, two essential characteristics, namely, reliability and validity, are considered while selecting a psychological test.

1. **RELIABILITY OF A TEST**

   The reliability of a test refers to the consistency of scores obtained by the same subjects on different occasions. If the same subjects are measured twice with a perfectly reliable instrument, the correlation between the two sets of scores is nearly equal to 1.00. Here reliability refers to the extent to which a measuring device yields consistent results upon testing and retesting.

2. **DEFINITIONS OF THE TERM 'RELIABILITY':**

   Anastasi (2) defines reliability of a test as,

   "the consistency of the scores obtained by the same individuals on different occasions or
with different sets of equivalent forms".

As Freeman points out;

"the reliability of a test is its ability to yield consistent results from one set of measures to another; it is the extent to which the obtained test scores are free from such internal defects as will produce errors of measurement inherent in the items and their standardization". (3)

Wrightstone and others gave a more adequate and precise picture of the term, when they say,

"reliability is that quality which is judged by the accuracy, consistency, or constancy of the measure obtained". (4)

The various definitions show that reliability stands for precision of measurement and consistency of results, but to what extent it can be expected with the psychological instrument is an important point for consideration. There are certain conditions which affect the reliability of the test. Mursell quotes Symonds and Walker, "The cause of unreliability may be classified as
those which are in the test itself, those which are in the person who takes it, and those which are in the person who gives it" (5). If the same test is used and the same person gives it to a pupil at two different occasions, then the results might be affected by the way a person takes the test on two different occasions. Personal characteristics of the tests, such as fluctuation of attention, motivation, health and emotional status also count. When tests are given to a whole class on two different occasions, naturally all the subjects may not take them with the same seriousness. Consequently, the correlation with two results might come to less than 1.00. Freeman, (F.S.) makes the following observations: (6).

"No test presently available is in itself perfectly reliable; scores for the same individuals, obtained on repeated testing, are not completely stable. Not only are there likely to be some different chance determinants in operation at different times, but it is quite normal for human beings to vary in performance, generally within fairly narrow limits, from one occasion to another. Such variation is expected quite aside from changes that occur
as part of the process of growth and development".

3. METHODS OF ESTIMATING RELIABILITY:

Measures of reliability are classified as relative reliability and absolute reliability. Relative reliability is expressed in terms of the reliability coefficient and absolute reliability in terms of the standard error of measurement. For measuring reliability of tests several methods are used. The following four popular methods are widely used:

(1) The test-retest method
(2) The parallel-form method
(3) The split-half method
(4) The method of rational equivalence.

(1) The Test-retest Method:

In this method the same subjects are tested with the same test on two occasions and the reliability coefficient is calculated from the two sets of scores. Such a reliability is known as the coefficient of stability. There are advantages and disadvantages in getting the reliability coefficient by this method of re-testing with the
same form. Administering the identical test form twice has an advantage of providing completely the same content and procedure on both the occasions. But, it requires double the time. Further the experience gained by the subject while taking the first test helps the subject when he takes the test the second time. If the interval between retests is not very long, the testee can recall his experiences. If these factors affect appreciably, the correlation will be high.

Many investigators have demonstrated that, in general, longer intervals between repeated tests will result in lowering the reliability coefficient.

If pupils are retested a number of times, they might get advantage in different ways, such as improvement in the skill of taking the test and their attitudes towards the test. Pupils who take the test a second time are no longer the same in all respects as when they took the test for the first time. All these factors influence the reliability coefficient. The main argument against the use of this method is based on the practice effect. Some psychologists do not reco-
mend the use of this method. Gulliksen writes,

"This method of repetition of the same test at different times should, in general, not be used, since it will give a spuriously high coefficient and the amount of error is not easy to determine" (7).

A counter argument on this aspect is given by Freeman, (F.S.) (8)

"In the first place the number of test items in both individual and group tests is so large that it is extremely difficult to recall a significant number of them, especially when the persons taking the test are working under pressure and must rapidly shift attention from one problem to another".

Gulliksen considers practice effect for giving a spuriously high coefficient, while Freeman, (F.S.) does not consider this point so serious.

2. **The parallel form Method**:

This technique is also referred to as the method of equivalent forms. With this technique,
the subjects are tested with form A on the first occasion and with another equivalent form B of the same test on the second occasion. These two forms of the same test are said to be parallel or equivalent, because they are made up of the same types of items covering the same materials. They also have the same means and variances. If one form correlates to a certain extent with some other measure, then the other form correlates to the same degree.

A Pearson product moment correlation coefficient is computed between the two sets of the scores obtained on the two forms. This method also has some advantages and disadvantages. The main criticism against the use of the same form is the effect of practice but in using equivalent form the effect of specific practice and recall are considerably lessened. The mental process is identical but the items in the two tests are not the same. Thus, the memory factor which was important with the test-retest method is ruled out here. Some of the factors such as learning, growth and change are still present. The main problem is also with the construction and standardization of the second form. In the second form
of the tests, the number of items should be the same, method and processes must be identical and the range of difficulty and distribution of items should be practically the same. The statistical results such as mean, median and standard deviation should correspond closely with the first test and other administration procedures should also be identical. Looking to these factors, it is difficult to have uniformity in all respects in the two tests.

(3) The Split-half Method:

In this method, the test is first divided into two equivalent 'halves'. The main point to consider in this approach is how to split the test to obtain the most nearly comparable halves. The first half and second half of the test cannot always be compared. In such comparisons there will be difference in nature and difficulty of the items. The cumulative effects of warming up, practice, fatigue, boredom and other factors influence progressively from the beginning to the end of the test:

The common practice is to divide the test by taking the odd numbered items as one part and
even numbered as the other. Other ways of making up two half-tests which will be comparable in content, difficulty and susceptibility to practice are employed, but the odds-even split is extensively used. The score pairs for the two halves used to calculate the reliability coefficient, often, corrected by the Spear-man Brown formula for shortening defects (9).

(4) The Method Of Rational Equivalence:

The method of rational equivalence stresses the intercorrelations of the items in the test and the correlations of the items with the test as a whole. The following data are required for determining the reliability of the test.

(a) The number of items in the test
(b) The standard deviation of the test scores
(c) The mean score of the test

The above data was, then, substituted in the well known Kuder - Richardson formula. (10)
MEASURING THE RELIABILITY OF THE PRESENT TEST:

The various methods discussed above are available for measuring the reliability of tests, but most of them are inapplicable in estimating the reliability of a handwriting scale. The important problem which faces the use of the present scale is how stable are the measures provided by the scale. When stability of measures was considered as an important factor in this regard, the test-retest method is the most appropriate method for estimating reliability. Therefore the three separate studies were conducted by the test-retest method. The details are discussed below.

THE RELIABILITY STUDIES:

Each of the three studies in the reliability of the present scale was restricted to a particular standard of a school. This was done to achieve a certain amount of economy in time and labour. The first study was conducted with 115 students belonging to four divisions of standard VII of the Municipal Jahangirpura Shala No.1 of Asarwa, Ahmedabad. The handwriting scale was administered on two occasions, one week apart, under the same
test conditions. The first test was taken by all the 115 students. While, during the second testing only 108 students of the original group were present. The specimens were scored and each student got a pair of scores, one for the first testing and the other for the second testing. Considering the score in the first testing as the \( X \) variables, and the score in the second testing as the \( Y \) variables, Pearson's product-moment coefficient of correlation was found out by the usual methods.

The study was repeated with two other groups of students, the first consisting of 84 girls of Std. IX of the New Era High School of Godhra and the second consisting of 76 boys of Std. X of Boys High School, Dholka. The interval between the two tests was 15 days and one month respectively. These reliability coefficients were also worked out by the usual methods. The details of the above three studies are presented in the table No. XXIX.

The time intervals between the test and the retest was made shorter in lower classes while it was progressively increased for the higher classes, because the rate of growth would be greater at the lower age groups.
TABLE NO. XXIX:
TEST RETEST RELIABILITY COEFFICIENTS.

<table>
<thead>
<tr>
<th>Study No.</th>
<th>Name of the School from which samples were drawn.</th>
<th>Standards</th>
<th>Number of pupils</th>
<th>Time interval between test and retest</th>
<th>Reliability coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Municipal Jhangirpura Shala No.1, Ahmedabad.</td>
<td>VII</td>
<td>108</td>
<td>1 week</td>
<td>.762</td>
</tr>
<tr>
<td>2</td>
<td>New Era High School, Godhra.</td>
<td>IX</td>
<td>-</td>
<td>15 days</td>
<td>.713</td>
</tr>
<tr>
<td>3</td>
<td>Boys High School, Dholka.</td>
<td>X</td>
<td>76</td>
<td>1 month</td>
<td>.685</td>
</tr>
</tbody>
</table>
A NOTE ON THE RELIABILITY OF THE PRESENT SCALE:

The above studies reveal that the present scale possesses a sufficiently high degree of reliability. This statement is well corroborated by the results obtained by Dr. Pillai and others in their research project on handwriting wherein they say,

"The reliability coefficient of the present scale by the test-retest method were found to be .74, .68 and .66 for intervals of one week, fifteen days and one month between two testings respectively". (10-b).

Thus, this scale provides consistent measures when administered at different periods of time.

6. THE CONCEPT OF VALIDITY:

The concept of validity has been interpreted in varying terms by different authors. Bingham defines the validity of a test:

"the closeness of agreement between the scores and some other objective measure of that
According to Travers, "A test is valid, if it actually measures what it claims to measure" (12).

Garrett observes, "The validity of a test or of any measuring instrument depends upon the fidelity with which it measures what it purports to measure". (13).

Wrightstone and others gave an almost similar explanation for the term when they say that, "validity (of a measuring instrument) is that characteristic which indicates the degree to which the instrument measures or provides a diagnosis of the psychological characteristics that it purports to measure". (14).

Thus, one very important characteristic of a test is validity. A test is valid to the degree to which it actually measures what it purports to measure or predict. An index of validity shows how closely the criterion performance could have been predicted from the test scores. The validity helps us in keeping a direct check on how far the test fulfills its function. If a psychological test does not have a sufficiently high degree of validity, it cannot fulfill the purpose for which it is meant. The new
test is validated against some internal criteria. Cronbach discusses two basic approaches to validity (1) logical analysis and (2) empirical analysis. He explains that in logical analysis, one attempts to judge precisely what the test measures and in empirical analysis, one attempts to show that the test is correlated with some other variable and therefore measures the same thing.

7. **TYPES OF VALIDITY**

The concept of validity, by its very definition, is a relative in characteristics. Validity of a measuring device is highly specific. A test may be valid for one purpose, and no other. If we administer a spelling test for the purpose of determining a student's achievement in grammar, it is to use a test for an invalid purpose. As knowledge about the specific nature of validity has grown, measurement experts have realized that various kinds of validity are necessary for the various purposes of different tests. Caggott has rightly pointed out,

"Validity is a relative term. A test is valid for a particular situation if it is not generally
valid" (15). Thus, it is obvious that validity should not be looked upon as an absolute characteristic of a measuring instrument, but is relative in that it is intimately associated with the purpose of the user. In this sense, today construct validity, face validity, content validity, factorial validity and concurrent validity are recognized.

(a) Construct Validity:

The construct validity of a test is the extent to which the test may be said to measure a "theoretical construct" or trait. Nunnally has explained this term with illustration in the following words:

"Construct validation consists of defining a measure in terms of numerous research findings. This is essentially the way in which intelligence tests have gained meaning. Enough research has been done with these instruments to know how the underlying function grows with the child and how intelligence test scores relate to numerous other variables. This gradual defining of the instrument in terms of what it does is the major approach to validating the same psychological measurement" (16).

As regards the approach of finding it, Freeman
(F.S.) suggests that construct validity of a given test might also be demonstrated by finding substantial correlation with other tests that have been shown to measure satisfactorily the mental processes or traits in question.

(b) **Face Validity**

This term refers to the acceptability of the test and test situation by the examinee or user, in terms of apparent uses to which the test is to be put. A test has face validity when it appears to measure the variable to be tested. The contents of the test are selected in such a way that they seem to be relevant for the objectives fixed and hence there is no further effort to confirm an assumption objectively. Face validity is not considered validity in the technical sense. It does not refer to what the test actually measures, but to what it appears superficially to measure.

(c) **Content Validity**

The content validity is essentially the systematic examination of the test content with a view to ascertaining whether it covers a representative
sample of the field to be tested. This method is commonly used in evaluating achievement tests. Each item selected is a sample of knowledge which the test intends to measure and a collection of all items should include a sample of the variable to be tested. Items should be first judged by more than one specialist and the final selection must be done on item analysis basis. Statistical procedures are employed to refine the original selection of items.

(d) **Factorial Validity**:

"A Glossary of Measurement terms" describes factor analysis as a method of analysing the inter-correlations among a set of variables, such as test scores. Factor analysis attempts to account for the inter-relationships in terms of some underlying 'factors' preferably fewer in number than the original variables. It reveals how much of the variation in each of the original measures arises from or is associated with each of the hypothetical factors. The factorial process aims at identifying a list of functional units within a test and how much loading of each is on the total performance of the test.

Freeman, (F.S.) writes the following about the
Freeman, (F.S.), writes, thus about the utility of factorial analysis. "Factorial analysis is intended to reduce the number of variables, or test categories needed to represent an individual's abilities or traits for specified purposes. It is a technique that yields clusters of correlations from which one can infer the existence of underlying variables. The psychologist must make the inference of their existence primarily on the basis of his psychological insights into the intellectual or behavioural processes involved in the tests included within each cluster of correlations" (17).

(e) **Concurrent Validity**

Originally psychologists used to speak of validation of the test with other tests. Now psychologists prefer the term "concurrent validity". This means that the new test is validated by correlating it, or comparing it, with some source of recent information for agreement.

If a student does well on some standardized reading test that measures all aspects of reading, he should also do well in completing and understanding history and geography reading assignments. The most common criteria employed for concurrent
validation are rating and correlating with other tests. Sometimes new tests are correlated with rating by school teachers and instructors, or job supervisors or others under whom the testees work.

8. **APPROACHES SELECTED FOR FINDING VALIDITY**:

The types of validity which applies in the case of a handwriting scale are construct validity, face validity and concurrent validity. Out of these construct validity is the most appropriate when we consider a handwriting scale. For the present scale, the concept of construct validity means the extent to which handwriting quality is measured, where the term quality is to be taken as the sum total of all the ten components discussed earlier.

The validity of the present scale was studied by four different methods. They are as follows:

1. Validity by assumption and definition
2. Validity by single teacher rating of the test specimens as external criteria.
3. Validity by Multiple Teacher Ratings of independent written specimens as the external criteria.
4. Assessing the Percentile ranks of excep-
tionally fine specimens of handwriting.

The details regarding the four validity studies are described below:

(1) **Validity by Assumption and Definition**:

A close examination of the present scale would reveal that it tends to measure handwriting quality. Hence it may be said that the scale possesses validity by assumption. Again handwriting has been carefully defined in terms of its components, and an attempt is made to measure all the defined characteristics. Thus, this scale also possesses the validity by definition.

(2) **Validity Study With Single Teacher Ratings Of The Test Specimens As External Criteria**:

The standardization sample consists of 2069 pupils of Gujarat. From this sample, a 10% sample was drawn by the stratified random sampling technique in the following manner:

First of all, all the specimens were grouped standardwise and then the specimens in each standard were arranged in the alphabetical order of the
names of the pupils. Starting with the first specimen of Std. V, the specimens were again allotted serial numbers, till the last specimen in standard XI was covered. Now starting with the first specimen (serial number 1) every tenth specimen was drawn (viz. serial number 1, 11, 21, 31, etc.) till 206 specimens were obtained. These 206 specimens contained proportionate samples of the general group at all educational levels.

The obtained sample of 206 specimens was again used to draw a smaller sample of 51 (one fourth of the sample), by following exactly the same procedure as discussed above. For this the specimens were again serially numbered and then the specimens were drawn at intervals of four, starting with serial number 1 till 51 specimens were drawn.

The 51 specimens drawn by stratiﬁed sampling were given to an expert Gujarati Teacher having 30 years' experience of teaching at primary, secondary and primary training college. He was instructed to give ranks according to merits. The general data at the top of the specimens had to be removed in order to introduce greater objectivity of judgement and hence this part was cut out and serial numbers
above were given at the top. The teacher was specifically told that ranking of the specimens was to be made for quality. The rank assigned was noted down on a piece of paper against serial number. Then these specimens were scored by the investigator using the scale. The scores were then converted into ranks. Thus, every specimen received two ranks, one given by the expert teacher and the other arising out of the scores. The rank-pairs obtained for the 51 specimens were used to work out the rank coefficient of correlation (rho), using the formula given by Garrett (18). The value of rho for the two types of ranking found out to be 0.672. The coefficient indicates that the scores provided by the scale are broadly in agreement with the ranking assigned by the teacher. The coefficient also indicates a high degree of validity for the present scale.

The following two limitations have come in the study:

(1) The same specimen has been used for the teacher rating and therefore teacher ratings obtained cannot be considered as a perfect external criterion.
A single teacher gave ranking. Therefore it is very probable that holistic judgement may be given by a single teacher.

Validity by Multiple-teacher Ratings of Independent Written Specimens as External Criteria:

The two limitations of the above study were overcome by this study. To make the external criteria really external, independent written specimens of the pupils were provided. Instead of a single teacher ranking the specimens, three teachers were asked to rank the specimens. The average of the ranks assigned by these three teachers was taken as the external criteria. The specimens were taken from std. VIII of the Oriental High School, Ahmedabad. The study made use of 40 pupils. Their handwriting specimens were scored with the help of the scale. These scores were utilized to rank their handwriting quality. These ranks served as the \( x \) variable for calculating the coefficient of correlation.

The Gujarati composition note books of the same set of 40 pupils were given to three teachers who have long experience of teaching Gujarati. They were asked to rank the pupils according to the quality
of their handwriting. They were also instructed to make use of the page with the best quality of writing as the specimen for rating. Thus, each pupil obtained three ranks from the three teachers. The average of the three ranks is taken as the final rank which served as the Y variable for calculating the rank coefficient of correlation (rho). The resulting coefficient was .542.

The above value is a better index of validity of the test as compared with the value obtained in the previous study. In this study, the advantage of utilizing handwriting products attained from classroom situations is the important factor to be considered. Therefore, the validity coefficient indicates that scores obtained with the scale substantially agree with the general assessment of handwriting quality made by the teachers.

It may be noted that even higher coefficient would have been obtained if a more heterogeneous group had been subjected to study by the same procedure.

(4) Assessing The Percentile Ranks Of Exceptionally Fine Specimens Of Handwriting:

A new method of validation is attempted here.
10 adults, who were known to possess very good handwriting in Gujarati language, were given the present test. The specimens were then scored using the scale. These scores were converted into percentile ranks by using the table of norms for the highest group (Std. X-XI) of boys and girls as the case may be. Since separate norms for adults were not available they were compared with the writing of high school pupils. It was seen that 9 out of 10 adults obtained percentile ranks in the range of 91 to 98.

Thus, it may be interpreted that the test helps to discriminate fine handwriting specimens from specimens having poor quality of handwriting.

A NOTE ON THE VALIDITY OF THE PRESENT SCALE:

The above different studies describe that the present scale possesses a tolerably high degree of validity in estimating handwriting quality in Gujarati language. The scope of validity studies, using an external criteria, suffers from certain inherent limitations and internal validity is not possible for the present handwriting scale. Inspite
of all these limitations it may be claimed that the present scale provides the best instrument for measuring handwriting quality in Gujarati language.
BIBLIOGRAPHY


