Chapter 8

Reliability

Reliability of a test is a criterion of test quality relating to accuracy of psychological measurements. The higher the reliability of the test the relatively more free it would be of measurement errors. Some regard it as stability of results in repeated testing, i.e. the same individual or object is tested in the same way so that it yields the same value from moment to moment provided that the thing measured has itself not changed in the meantime. The concept of reliability underlines the computation of error of measurement of a single score; whereby we can predict the range of fluctuations likely to occur in a single individual score as a result of irrelevant chance factor.

The test reliability, in its broadest sense indicates the extent to which individual differences as in a test scores are attributable to true differences in the chance scores under consideration and the extent to which they are attributable to chance scores. In technical terms, the measures of test-reliability make it possible to estimate as to what proportion of total test score is error variance. The more the error the lesser the reliability. Practically speaking, this means that if we can estimate the error variance in any measure, we can also estimate the measure of reliability. This brings us two equivalent definitions of reliability.

1. Reliability is the proportion of the 'true' variance to the total obtained variance of the data yielded by measuring instrument.

2. It is the proportion of error variance to the total obtained variance of the data yielded by measuring instrument subtracted from 1.00. The index of 1.00 indicates perfect reliability.

Types of Reliability: There are five methods to measure reliability of a test. These are, (i) Test-Retest Method, (ii) Method of Parallel Form, (iii) Split-half Reliability (iv) The method of Rational Equivalence and (v) Cronbach Alpha. These five methods are discussed in detail below.

1. Test-Retest Method

The most frequently used method to find the reliability of a test is by repeating the same test on a second occasion. The reliability coefficient (*r*) in this case would

metric data — 93

interval / rated

be the correlation between the score obtained by the same person on two administrations of the test. An error variance corresponds to the random fluctuations of performance from one test session to another test session. The problem related to this test is the controversy about the interval between two administrations. If the interval between tests is long (say six months) and subjects are young children, growth changes will effect the retest scores. In general, it increases the initial scores by various amounts and tends to lower the reliability coefficient. Owing to the difficulty in controlling the factor which influences scores on retest, the retest method is generally less useful than are the other methods.

The formula used to find the test-retest reliability is the Pearson Product moment formula.

$$\text{reliability} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2]\,[N\Sigma Y^2 - (\Sigma Y)^2]}}$$

Take, for example, a test to measure numerical ability having 60 items was administered to a group of 20 students twice with a gap of 15 days. The two sets of scores obtained were Test score (X) and Retest Score (Y) as given below:

= 0.31 0·25

2. Method of Parallel Form (External)

To overcome the difficulty of practice and time interval in case of test-retest method, the method of parallel or alternate form is used. Using the equivalent or parallel forms, has some advantage like lessening the possible effect of practice and recall. But this method presents an additional problem of construction and standardization of the second form. According to Freeman, both forms should meet all of the test specifications as follows:

(a) The number of items should be same.

(b) The kinds of items in both should be uniform in respect to content, operations or traits involved, levels and range of difficulty, and adequacy of sampling.

(c) The items should be uniformly distributed as to difficulty.

(d) Both test forms should have the same degree of item homogeneity in the operations or traits being measured. The degree of homogeneity may be shown by intercorrelations of items with subtest scores, or with total-test scores.

(e) The means and the standard deviations of both forms should correspond closely.

(f) The mechanics of administering and scoring should be uniform.

Freeman states that the above are the ideal criteria of equivalent forms, but complete uniformity in all respects cannot be expected. However, it is necessary that uniformity be closely approximated.

Those parallel forms are administered on the group of individuals and the correlation coefficient is calculated between one form and the other. For instance, 1937 Stanford-Binet Scale has form *L* and form *M*. The content of the forms was derived from one and the same process of standardization. The correlation of 0.91 is obtained between these two forms for chronological age of seven years. The formula and method used to find reliability with the help of parallel or alternative techniques is the same as used in test-retest method.

Split-half Reliability (Internal)

This method has advantage over the retest method is that only one testing is needed. This technique is also better than the parallel-form method to find reliability because only one test is required. In this method test is scored for the single-testing to get two halves so that variation brought about by difference between the two testing situations is eliminated. A marked advantage of the split-half technique lies in the fact that chance errors may affect scores on the two halves of the test in the same way thus tending to make the reliability coefficient too high. This follows because the test is administered only once. The larger the test, lesser the probability that the effects of temporary and variable disturbances will be cumulative in one direction and the more accurate the estimate of score reliability.

The two halves can be made by counting the number of odd-numbered items answered correctly as one half and the number of even numbered items answered correctly as other half. In other words, odd-items and even-items are scored separately and those are considered as two separate halves. There are other methods also to split the test items into two halves, like items 1 and 2 will go to first score, 3 and 4 will go to second score, 5 and 6 will go to first score, and 7 and 8 will go to second score and so on. The other method to divide into two halves is to consider first fifty percent items as one half and the second fifty percent items as another half. Whenever the difficulty level of the test items are not same, we apply odd-even method and if the difficulty level is same; we apply the first half and second half method to divide the test into two halves. Once the two halves have been obtained for each individual score, these halves will be correlated with the help of Pearson Product moment formula, namely

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

667 = 0.174

650

575 ## 4 Method of Rational Equivalence (Internal)

672

598 The coefficient of internal consistency could also be obtained with the help of Kuder Richardson formula number 20. One of the techniques for items analysis is
675 item difficulty index. Item difficulty is the proportion or percentage of those
667 answering correctly to an item, say Symbol P is used to represent the difficulty
700 index. Suppose an item X has $p = 0.74$. This means items X was answered
648 correctly by 74 percent of those who answered the item.

783

624 To compute reliability with the help of Kuder-Richardson formula nunber 20, the following procedure is used. Firstly write the first column in a worksheet
550 showing the number of items. The second column should give the difficulty value
483 (p) of each item obtained during item analysis. The third column is given as q
672 where $q = 1 - p$. The fourth column is taken as $(p)(q)$. This column is the product
702 of column 2 and column 3.

13165

The Kuder-Richardson formula no. 20 is

$$\text{Reliability} = \frac{N}{N-1}\left[1 - \frac{\Sigma p\,q}{\sigma_t^2}\right]$$

where N is the number of items on the test σ_t^2 is the variance of the test.

For the above formula, there is one basic assumption that there is some difficulty level for all the test items. In other words, the same proportion of individuals (but not the same individuals) answer each item correctly. It has been observed that the above formula holds true even when assumption of equal item difficulty is not satisfied. The formula of Rational Equivalence cannot yield strictly comparable results as obtained by other methods of finding reliability. The actual differences obtained by the method of rational equivalence and split-half method is never large and is often negligible within the acceptable range. Two forms of a test are equivalent when the corresponding items like a_1, a_2, b_1, b_2 etc. are interchangeable, and when the item-item correlations are the same for both the forms.

Cronbach Alpha

The Kuder-Richardson formula is applicable to find internal consistency of tests whose items are scored as right or wrong, or according to some other all or none system. Some tests, however, may have multiple-choice items. On a personality inventory, however, there are more than two response categories. For such tests, a generalized formula has been derived known as coefficient alpha (Cronbach, 1951). In this formula, the value of Σpq is replaced by $\Sigma\sigma_i^2$, the sum of variance of item scores. The procedure is to find the variance of all individuals scores for each item and then to add these variances across all items.

The formula is

$$r_{tt} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_t^2}\right)$$

Factors Affecting Reliability

(i) *Variability of Ages*: The variability of the group affects the reliability coefficient. The reliability will be higher for a group having wider range and the reliability will be small for a group having small variation of the trait or ability assessed. This is illustrated in Fig. 8.1 below:

For instance if we have a group completely homogeneous with respect to chronological age (CA). For this the range of test scores will be from extremely low to extremely high. As there is no deviation in age, hence the correlation of CA with test scores would be zero. Further if there is small variation in a group with respect to CA, the correlation of CA with test scores would be lower and if there is a large variation in CA then correlation coefficient will be larger. Therefore while interpreting a reliability coefficient of a test, it is necessary to know the range for which the test is standardized.

(ii) *Variability of scores*: As discussed above with regard to variability of age, the reliability is affected in the same way with the variability in the measured

Fig. 8.1. Increase in test reliability with increase in variability of a group.

scores. When a variation among the testees is small, the correlation between two sets of scores may also be lowered by chance and by minor psychological factors. Because the testees in such a group are closely clustered the chances in scores and relative position produced by extraneous factors are more significant than they would be in a widely divergent group.

(iii) *Time Interval between Testings*: When there is a time interval between test and retest, the retest results will be affected due to differences in individual performances and also due to the change in the environmental conditions. If the time interval has been quite long, namely in case of young children — an individuals retest results may be influenced due to their growth tempo or due to enduring conditions like emotional experiences.

(iv) *Effects of Practice and Learning*: Practice on the test will help in learning and this in turn can affect the reliability of a test. For example, therapy or counselling may modify an individuals attitudes, values, and behaviour sufficiently to produce significant differences in test-retest results in case of personality test.

(v) *Consistency in scores*: Lack of agreement among scorers will drastically affect the reliability coefficient. This is generally true in case of tests in which entirely objective scoring is not available. For such tests, it is advisable to know the extent of agreement in scoring among the competent psychologists who have scored the same set of responses.

(vi) *Effect of test length*: The reliability of a test is directly dependent on the test length, i.e., the number of items in a test. Suppose a test of 40 items has a reliability of 0.60. Now we increase the items to 120 by adding 80 more homogeneous items to the previous 40 items. The reliability of the new test of 120 items will be increased. Similarly by shortening the length of the test, the reliability of the test will also be decreased. How much will be the reliability of the test change after increasing/decreasing the items is given by a formula called Spearman-Brown formula.

The term reliability in psychological research refers to the consistency of a research study or measuring test.

For example, if a person weighs themselves during the course of a day they would expect to see a similar reading. Scales which measured weight differently each time would be of little use.

The same analogy could be applied to a tape measure which measures inches differently each time it was used. It would not be considered reliable.

If findings from research are replicated consistently they are reliable. A correlation coefficient can be used to assess the degree of reliability. If a test is reliable it should show a high positive correlation.

Of course, it is unlikely the exact same results will be obtained each time as participants and situations vary, but a strong positive correlation between the results of the same test indicates reliability.

There are two types of reliability – internal and external reliability.

- Internal reliability assesses the consistency of results across items within a test.
- External reliability refers to the extent to which a measure varies from one use to another.

Assessing Reliability

Types of Reliability

INTERNAL (extent to which a measure is consistent within itself.)	EXTERNAL (the extent to which a measure varies from one use to another.)
split-half method: measures the extent to which all parts of the test contribute equally to what is being measured.	**test re-test**: measures the stability of a test over time.
	Inter-rater: to the degree to which different raters give consistent estimates of the same behavior

Split-half method

The split-half method assesses the internal consistency of a test, such as psychometric tests and questionnaires. There, it measures the extent to which all parts of the test contribute equally to what is being measured.

This is done by comparing the results of one half of a test with the results from the other half. A test can be split in half in several ways, e.g. first half and second half, or by odd and even numbers. If the two halves of the test provide similar results this would suggest that the test has internal reliability.

The reliability of a test could be improved through using this method. For example any items on separate halves of a test which have a low correlation (e.g. r = .25) should either be removed or re-written.

The split-half method is a quick and easy way to establish reliability. However it can only be effective with large questionnaires in which all questions measure the same construct. This means it would not be appropriate for tests which measure different constructs.

For example, the Minnesota Multiphasic Personality Inventory has sub scales measuring differently behaviors such depression, schizophrenia, social introversion. Therefore the split-half method was not be an appropriate method to assess reliability for this personality test.

Test-retest

up

The test-retest method assesses the external consistency of a test. Examples of appropriate tests include questionnaires and psychometric tests. It measures the stability of a test over time.

A typical assessment would involve giving participants the same test on two separate occasions. If the same or similar results are obtained then external reliability is established. The disadvantages of the test-retest method are that it takes a long time for results to be obtained.

Beck et al. (1996) studied the responses of 26 outpatients on two separate therapy sessions one week apart, they found a correlation of .93 therefore demonstrating high test-restest reliability of the depression inventory.

This is an example of why reliability in psychological research is necessary, if it wasn't for the reliability of such tests some individuals may not be successfully diagnosed with disorders such as depression and consequently will not be given appropriate therapy.

The timing of the test is important; if the duration is to brief then participants may recall information from the first test which could bias the results. Alternatively, if the duration is too long it is feasible that the participants could have changed in some important way which could also bias the results.

Inter-rater reliability

The test-retest method assesses the external consistency of a test. This refers to the degree to which different raters give consistent estimates of the same behavior. Inter-rater reliability can be used for interviews.

Note, it can also be called inter-observer reliability when referring to observational research. Here researcher when observe the same behavior independently (to avoided bias) and compare their data. If the data is similar then it is reliable.

Where observer scores do not significantly correlate then reliability can be improved by:

- Training observers in the observation techniques being used and making sure everyone agrees with them.
- Ensuring behavior categories have been operationalized. This means that they have been objectively defined.

For example, if two researchers are observing ‘aggressive behavior’ of children at nursery they would both have their own subjective opinion regarding what aggression comprises. In this scenario it would be unlikely they would record aggressive behavior the same and the data would be unreliable.

However, if they were to operationalize the behavior category of aggression this would be more objective and make it easier to identify when a specific behavior occurs.

For example, while “aggressive behavior" is subjective and not operationalised, “pushing" is objective and operationalized. Thus researchers could simply count how many times children push each other over a certain duration of time.