C H A P T E R - 6

RELIABILITY AND VALIDITY ESTIMAES OF THE TESTS

6.1 Introduction :

The two qualities of a single personnel selection test in which the research worker is particularly interested are its reliability and its validity. By a perfectly reliable measurement we mean that it is completely accurate or free from error. The same "yard stick" applied to the same individual or object in the same way should yield the same value from moment to moment, provided that the thing measured has itself not changed in the meantime.[2]

Thorndike explains its meaning as " A measurement procedure is reliable to the extent that repeated measurement gives consistent results for the individual consistant in that his score remains substantially the same when the measurement is repeated, or in what his standing in the group shows little change. A measurement procedure is valid in so far as it correlates with some measure of success in the job for which it is being used as a predictor."[2]

Hence, standard forms or schedules for evaluating a test are in use in a country like U.S.A., where the number of available tests is prolific. Such schedules include,

the details about the reliability and validity of tests. It is better not to use any test, rather than use an imperfect one. In view of this, it is inevitable that a test maker checks the accuracy and precision of the measurement procedure, as well as the extent to which the test measures what it was intended to, before releasing the newly developed test for general use.[3]

6.2 Reliability :

When Shakespear said, "Consistency thou art a jewel", he said the truth. But a person can be consistently wrong and a test also can measure consistently something which is wrong.[4] Whenever anything is measured, whether in the physical, biological or behavioural sciences, there is some possibility of chance error. This is true of psychological tests as well. The variations of results obtained with the same test administered more than once, using the same persons as subjects, or within the parts of test given only once, are due not only to chance factors, which should be eliminated so far as possible, but also to actual differences among the individuals taking the test and to whatever defects may be inherent in the instrument itself. Thus the reliability of a test depends upon the degree of 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.

Reliability concerns the precision of measurement of what is measured. Some random error is involved in scientific measurements. For example, a metal ruler expands and contracts with changing temperatures. This introduces error into measurements and consequently lowers precision of measurements made with this ruler. Psychological measures also contain a portion of error analogous to that of the metal ruler.

No educational achievement test, no other type of mental test, and indeed no physical measurement has ever achieved the reliability coefficient of 1.00. Error is unavoidably involved in any measurement, but the goal of measurement specialists in all fields is to reduce these inevitable error of measurement to a reasonable minimum.

A careful distinction should be made between test validity and test reliability. The way of clarifying the concept of reliability is to contrast it with validity. These two definitions point out the contrast.

1 The term "reliability" means the consistency with which a set of test scores measure whatever they do measure.

2 The term "validity" means the accracy with which a set of test scores measure what they ought to measure.

If the perforations on a target made by successive shots from a rifle are all clustered closely, the rifle is performing reliably. If those perforations are all clustered in the bull's-eye, the rifle is also performing validly.[5] A test can have high reliability and not be valid for any particular purpose. For example, we might use the weight of individuals to predict college grades. Whereas weight may be measured very preciously and thus be highly reliable, it would be invalid as a predictor of college grades. In order for a test to be highly valid, it must be highly reliable also. High reliability is a necessary but not sufficient condition for high validity.

Reliability of measures can also be considered an issue relating to generalizability. Essentially, relaibility concerns the extent to which measurements of particular traits are repeatable under the same conditions. Repeatability of measurement is a fundamental necessity in all areas of science. If a chemist found that two thermometers measured the temperature of

a particular liquid differently, he could not trust either instrument.[6]

Guilford gives the basic definition of reliability as "The reliability of any set of measurements is logically defined as the proportion of their variance." He further states that, one should speak of the reliability of a certain instrument applied to a certain population under certain conditions.[7] Ebel gives the operational definition of test reliability as follows :

> The reliability coefficient for a set of scores from a group of examinees is the coefficient of correlation between that set of scores and another set of scores on an equivalent test obtained independently from the members of the same group.[8]

From the above definition it is implied that

1 It is the property of a test when applied to a particular group of examinees and not a property of test itself.

2 Coefficient of correlation provides a relative rather than an absolute measure of agreement between the pairs of scores for the same persons. As coefficient of correlation is used as a measure of reliability if the differences between scores

for the same person are small relative to the differences between scores for different persons, then the test will tend to show a high reliability. But if the differences between scores for the same person are large relative to the differences between persons, then the scores will show low reliability.

3 The operational definition calls for two or more independent measures obtained from equivalent tests of the same trait for each member of the group.

Psychologists try to give the meaning of reliability in their own ways. But all of them put stress on the fact that reliability of an instrument, may it be psychological or biological, is the consistency of the results obtained by that particular instrument.

F. S. Freeman puts it as :

> The term reliability has two closely related but somewhat different connotations in psychological testing. First, it refers to the extent to which a test is internally consistent, that is, consistency of results obtained throughout the test when administered once. In other words how accurately is the test measuring at a particular time ?
>
> Second, reliability refers to the extent to which a measuring device yields consistent results upon testing and retesting.9

Thorndike's views on this are similar but they are expressed differently. He says, "In evaluating the consistency of a set of measurements there are two somewhat different aspects to be considered. These may be spoken of as absolute consistency and relative consistency. The degree of absolute consistency is seen in the actual amount of variation which results when a particular measuring instrument is applied more than once to the same individual..... For the evaluation of relative consistency, some statement is required of the degree to which individuals keep the same relative standing inthe group when two equivalent forms of a test are applied to all the members of the group."[10]

Rose and Stanley put it as :

> By reliability is meant the degree to which the test agrees with itself. To what extent can two or more forms of the test be relied upon to give the same results, or the same test to give the same results when repeated ? If the scores on the test are stable under these conditions, the test is said to be reliable. In a word, reliability means consistency.[11]

From the above discussion it follows that the test results should be consistent. Thus there are several meanings sttached to reliability. It includes dependability, consistency and

stability. Each signifies something different as applied to measurements. Even the same term has slightly different meanings as applied to different measurement operations.

6.2.1 Factors Affecting Reliability Estimates :

As discussed above it is impossible to devise an instrument even in physical sciences, which gives fully reliable results. One or the other factor, affecting reliability estimates, lessens it to some extent. Hence reliability is always talked in terms of degrees. Some possible sources of variation in psychological measurements are listed by Thorndike as follows :

1. Lasting and general characteristics of the individual
2. Lasting but specific characteristics of the individual
3. Temporary but general characteristics of the individual
4. Temporary and specific characteristics of the individual
5. Systematic or chance factors affecting the administration of the test or the appraisal

of test performance

6. Variance not otherwise accounted for chance[12]

Thorndike furchar says,

> There will also be variance which is associated only with the one particular set of measurements, that is, which will not be reproduced another time. This may be designated as, error variance'. The existence of this 'error variance' corresponds to the fact of unreliability, and its amount relative to the total of all variance is a measure of the degree of unreliability.[13]

Reliability is also affected by the following factors.

1. Range of ages : The reliability is low if the variation of the trait or ability being measured is small. If the group is of wider range, the coefficient is higher.

2. Range of scores : The coefficient of correlation will be lowered by restricting the group's range of variation in the trait being measured.

3. The time interval between testings : When reliability estimates are based upon odd-even correlations or upon scores of two equivalent forms administrered

at a single sitting, the results are relatively uniformly affected by the examinees's physical condition and attitudes, and prevailing environmental conditions. When there is a time interval, the retest results will be affected by the normally expected fluctuations in individual performances and by changes in environmental conditions.

4. The effect of practice and learning : Such effect depends upon the content of the test, the length of the interval and upon the examinee's experiences during the interval.
5. Reliability of subtests : Other factors being equal, the reliability of test increases with increase in lengths although not in direct proportion.
6. Consistency of scores : Some tests are not entirely objective in scoring, because the examiner at times finds it necessary to judge the correctness or quality of responses.[14]

6.3 Methods of Estimating Reliability :

Various methods of estimating test reliability are known. The following four methods will be considered, in view of their