

5.1. Reliability

What does the term reliability mean? Reliability means Trustworthy. A test score is called reliable when we have reasons for believing the test score to be stable and objective. For example if the same test is given to two classes and is marked by different teachers even then it produced the similar results, it may be considered as reliable. Stability and trustworthiness depends upon the degree to which score is free of chance error. We must first build a conceptual bridge between the question asked by the individual (i.e. are my scores reliable?) and how reliability is measured scientifically. This bridge is not as simple as it may first appear. When a person thinks of reliability, many things may come into his mind – my friend is very reliable, my car is very reliable, my internet bill-paying process is very reliable, my client's performance is very reliable, and so on. The characteristics being addressed are the concepts such as consistency, dependability, predictability, variability etc. Note that implicit, reliability statements, is the behaviour, machine performance, data processes, and work performance may sometimes not reliable. The question is “how much the scores of tests vary over different observations?”

5.1.1 Some Definitions of Reliability:

According to Merriam Webster Dictionary:

“Reliability is the extent to which an experiment, test, or measuring procedure yields the same results on repeated trials.”

According to Hopkins & Antes (2000):

“Reliability is the consistency of observations yielded over repeated recordings either for one subject or a set of subjects.”

Joppe (2000) defines reliability as:

“...The extent to which results are consistent over time and an accurate representation of the total population under study is referred to as reliability and if the results of a study can be reproduced under a similar methodology, then the research instrument is considered to be reliable.” (p. 1)

The more general definition of the reliability is: The degree to which a score is stable and consistent when measured at different times (test-retest reliability), in different ways (parallel-forms and alternate-forms), or with different items within the same scale (internal consistency).

5.2 Types of Reliability

Reliability is one of the most important elements of test quality. It has to do with the consistency, or reproducibility, of an examinee's performance in the test. It's not possible

to calculate reliability exactly. Instead, we have to estimate reliability, and this is always an imperfect attempt. Here, we introduce the major reliability estimators and talk about their strengths and weaknesses.

There are six *general classes of reliability estimates*, each of which estimates reliability in a different way. They are:

i) Inter-Rater or Inter-Observer Reliability

To assess the degree to which different raters/observers give consistent estimates of the same phenomenon. That is if two teachers mark same test and the results are similar, so it indicates the inter-rater or inter-observer reliability.

ii) Test-Retest Reliability:

To assess the consistency of a measure from one time to another, when a same test is administered twice and the results of both administrations are similar, this constitutes the test-retest reliability. Students may remember and may be mature after the first administration creates a problem for test-retest reliability.

iii) Parallel-Form Reliability:

To assess the consistency of the results of two tests constructed in the same way from the same content domain. Here the test designer tries to develop two tests of the similar kinds and after administration the results are similar then it will indicate the parallel form reliability.

iv) Internal Consistency Reliability:

To assess the consistency of results across items within a test, it is correlation of the individual items score with the entire test.

v) Split half Reliability:

To assess the consistency of results comparing two halves of single test, these halves may be even odd items on the single test.

vi) Kuder-Richardson Reliability:

To assess the consistency of the results using all the possible split halves of a test.

Let's discuss each of these in turn.

5.2.1. Inter-Rater or Inter-Observer Reliability

Whenever we observe or activities of humans, we have to think about the procedure for reliable and consistent results. For this two or more than two observers are assigned to observe the students or teachers. So how do we determine whether two observers are being consistent in their observations? We probably should establish inter-rater reliability by considering the similarity of the scores awarded by the two observers. After all, if we use data to establish reliability, and we find that reliability is low. We should have to focus upon the criteria established for the observation. And if it is tried first in the actual situation then it may help to develop the reasonable criteria for the observation, and may be more objective.

There are two major ways to actually estimate inter-rater reliability. If your measurement consists of categories -- the raters are checking off which category each observation falls in -- you can calculate the percent of agreement between the raters. For instance, let's say you had 100 observations that were being rated by two raters. For each observation, the rater could check one of three categories. Imagine that on 86 of the 100 observations, the raters checked the same category. In this case, the percent of agreement would be 86%. OK, it's a crude measure, but it does give an idea of how much agreement exists, and it works no matter how many categories are used for each observation.

The other major way to estimate inter-rater reliability is appropriate when the measure is a continuous one. There, all you need to do is calculate the correlation between the ratings of the two observers. For instance, they might be rating the overall level of activity in a classroom on a 1-to-7 scale. You could have them give their rating at regular time intervals (e.g., every 30 seconds). The correlation between these ratings would give you an estimate of the reliability or consistency between the raters.

One might think of this type of reliability as "calibrating" the observers. There are other things one could do to encourage reliability between observers, even without estimating it. For instance, in a psychiatric unit where every morning a nurse had to do a ten-item rating of each patient on the unit. Of course, it's difficult to count on the same nurse being present every day, so there is a need to find a way to assure that any of the nurses would give comparable ratings. The way we did, it was to hold weekly "calibration" meetings where we would have all of the nurses ratings for several patients and discuss why they chose the specific values they did. If there were disagreements, the nurses would discuss them and attempt to come up with rules for deciding when they would give a "3" or a "4" for a rating on a specific item. Although this was not an estimate of reliability, it probably went a long way towards improving the reliability between raters.

Activity 5.1: Develop an essay type test for any class, administer it, get it marked from two raters and then compare the marks given by the two raters for each question.

5.2.2. Test-Retest Reliability

Test-retest is a statistical method used to determine a test's reliability. The test is performed twice; in the case of a questionnaire, this would mean giving a group of participants the same questionnaire on two different occasions.

This form of reliability is used to judge the consistency of results across items on the same test. Essentially, you are comparing test items that measure the same construct to determine the tests internal consistency. When you see a question that seems very similar to another test question, it may indicate that the two questions are being used to gauge reliability. Because the two questions are similar and designed to measure the same thing, the test taker should answer both questions the same, which would indicate that the test has internal consistency.

We estimate test-retest reliability when we administer the same test to the same sample on two different occasions. This approach assumes that there is no substantial change in the construct being measured between the two occasions. The amount of time allowed between measures is critical. We know that if we measure the same thing twice that the correlation between the two observations will depend in part by how much time elapses between the two measurement occasions. The shorter the time gap, the higher the correlation; the longer the time gap, the lower the correlation. This is because the two observations are related over time -- the closer in time we get the more similar the factors that contribute to error. Since this correlation is the test-retest estimate of reliability, you can obtain considerably different estimates depending on the interval.

Activity 5.2: Develop a test of English for sixth grade students, administer it twice with a gap of six weeks, find the relationship between the scores of students between 1st and 2nd administration.

5.2.3. Split-Half Reliability

Suppose you have to develop a test of 30 items and you want to know that how reliable the test is? What you have to do is to administer the test, mark it and divide it in to two parts, in such a way that place all the even numbered items (2,4,6.....) in one half and the odd numbered items (1,3,5.....) in the second. Calculate the reliability by using the Spearman-Brown prophecy formula given below.

Actually in split-half reliability we randomly divide all items that claim to measure the same contents into two sets. We administer the entire instrument to a sample of students and calculate the total score for each randomly divided half. The split-half reliability estimate is simply the correlation between these two total scores.

Normally a single test is used to make two shorter alternate forms. This method has the advantage that only one test administration is required, and therefore memory and the practice and maturation effects are not involved. Furthermore, it does not require two tests. So it has many advantages over parallel form and test-retest methods, therefore it is the most frequently used method of finding internal consistency of the classroom tests. The formula used for the reliability of the full test is Spearman-Brown prophecy formula as given below.

$$\text{Reliability of the Full Test} = \frac{2(\text{reliability of the half test})}{1 + (\text{reliability of the half test})}$$

5.2.4 Parallel-Form Reliability

In parallel form reliability we have to create two different tests from the same contents to measure the same learning outcomes. The easiest way to accomplish this is to write a large set of questions that address the same contents and then randomly divide the questions into two sets. Now it's time to administer both instruments to the same students at the same time. The correlation between the two parallel forms is the estimate of reliability. One major problem with this approach is that you have to be able to write lots of items that reflect the same contents. This is often no easy to do job. Furthermore, this approach makes the assumption that the randomly divided halves are parallel or equivalent. Even by chance, this will sometimes not be the case. The parallel forms approach is very similar to the split-half reliability described earlier. The major difference is that parallel forms are constructed so that the two forms can be used independent of each other and considered equivalent measures. For instance, we might be concerned about a testing threat to internal validity. If we use Form A for the pretest and Form B for the posttest, we minimize that problem. It would even be better if we randomly assign individuals to receive Form A or B on the pretest and then switch them on the posttest. With split-half reliability we have an instrument that we wish to use as a single measurement instrument and only develop randomly split halves for purposes of estimating reliability.

Activity 5.3: Make two tests of mathematics and compare its reliability through Parallel-Forms Reliability method.

5.2.5. Internal Consistency Reliability

In internal consistency reliability estimation, we use our single test. The test is administered to a group of students on one occasion to estimate reliability. In effect we judge the reliability of the instrument by estimating how well the items that reflect the same content give similar results. We are looking at how consistent the results are for different items for the same construct within the measure. There are a wide variety of internal consistency measures that can be used.

5.2.6. Kuder Richardson Reliability

The estimates of internal consistency of the test are commonly calculated by using Kuder-Richardson methods. These measures to extent to which items within one form of

the test have as much in common with one another as do the items in that one form with corresponding items in an equivalent form. The strength of this estimate of reliability depends upon the context to which the entire test represents a single, fairly consistent measure of a concept.

Normally these estimates are lower than the split halves but estimates higher than the test-retest and parallel form estimates. These techniques are also called item total correlations. There are different techniques to estimate the internal consistency of the test using K-R procedures, but two of them are more frequently used by the measurement experts. The first KR-20 is difficult to calculate as it is based on the information of the percentages of the students passing each item on the test. However, it gives more accurate results (Kubiszyn and Borich, 2003). The KR-20 formula is given below.

KR20 Formula

$$r = \frac{n}{n-1} \left[1 - \frac{\sum P_i^2}{n} \right]$$

Where “pq” provides a test score error variance for an "average" person, we know that the sampled people vary, i.e., the variance of their raw scores is greater than zero. Persons with high or low scores have less score error variance than those with scores near fifty percent correct where the score error variance is maximum. Since the "average" person variance used in the KR20 formula is always larger than the lower score error variance of persons with extreme scores, it must always overestimate their score error variances.

The second formula, which is easier to calculate but slightly less accurate is called KR21. It requires only the information about the number of items, the mean of the test score and the standard deviation. The formula KR21 is as under.

$$r_1 = \frac{n\sigma^2 m(n-m)}{\sigma^2(n-1)}$$

Studies indicated that this formula provide good results even when the item difficulties are not consistent.

5.3 Factors Affecting Reliability

Reliability of the test is an important characteristic as we use the test results for the future decisions about the students' educational advances and for the job selection and many more. The methods to assure the reliability of the tests have been discussed. Many examples have been provided in order to in-depth understanding of the concepts. Here we shall focus upon the different factors that may affect the reliability of the test. The degree of the affect of each factor varies from the situation to situation. Controlling the factor may improve the reliability and otherwise it may lower the consistency of production of scores. Some of the factors that directly or indirectly affect the test reliability are given as under.