Reliability and validity in research

ASSIGN BUSTER

The two most important and fundamental characteristics of any measurement procedure are reliability and validity. Reliability and validity tells us whether a research being carried out studies what it is meant to study, and whether the measures used are consistent. These two principles are discussed below.

Reliability

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. It is the extent to which a questionnaire, test, observation or any measurement procedure produces the same results on repeated trials. It is important to note the idea of replicability or repeatability of results or observations. Reliability relates to numbers or scores and not humans. Therefore, it would be wrong in the field of research to say that someone is reliable. Take for instance, in a mathematics quiz competition for schools. The degree to which the scores of the panel of judges for each contestant agree is an indication of reliability. Similarly, the degree to which an individual’s responses (i. e., their scores) on a survey would stay the same over time is also a measure of reliability.

A measure can be reliable without being valid. A measure cannot be valid without being reliable. Consider a tape rule that always measures my height as 175cm, taller than my true height. This tape-rule (though invalid as it incorrectly assesses height) is perfectly reliable as it consistently measures my height as 175cm, taller than I truly are. A research example of this phenomenon would be a questionnaire designed to assess the impact of corporate social responsibility on employee commitment that asked questions such as, “ Do you like to swim?”, “ What do you like to eat more, pizza or hamburgers?” and “ What is your favorite colour?”. As you can readily imagine, the responses to these questions would probably remain stable over time, thus, demonstrating highly reliable scores. However, are the questions valid when one is attempting to measure impact of corporate social responsibility on employee commitment? Of course not, as they have nothing to do with the research objective.

Assessing the Four Types of Reliability

There are four aspects of reliability, namely: equivalence, stability and internal consistency (homogeneity). It is important to understand the distinction between these three as it will guide one in the proper assessment of reliability given the research protocol.

Test-Retest reliability (also called Stability): This answers the question, “ Will the scores or results be stable over the time a measure is administered”. Stability is said to occur when the same or similar scores are obtained with repeated testing with the same group of respondents. Stability is assessed through a test-retest procedure that involves administering the same measurement instrument to the same individuals under the same conditions after some period of time. The reliability coefficient is expected to be highly correlated. For example, if a classroom achievement test is administered today and the test is given two weeks later, it is expected to have a reliability coefficient of r = 0. 75 or more.

Parallel forms reliability (also called Equivalence): This answers the question, “ Are the two forms of the test or measure equivalent?” If different forms of the same test or measure are administered to the same group; one would expect that the reliability coefficient will be high. Equivalence measures the level of agreement between two or more instruments that are administered at nearly the same point in time. It is measured through a parallel forms procedure in which one administers alternative forms of the same measure to either the same group or different group of respondents. This administration of the various forms occurs at the same time or following some time delay. The higher the degree of correlation between the two forms, the more equivalent they are.

Internal consistency reliability or homogeneityanswers the question, “ How well does each item measure the content or construct under consideration?” It is an indicator of reliability for a test or measure which is administered once. One expects the correlation between responses to each test item to be highly correlated with the total test score. For example, an employee job satisfaction (attitude scale) or a classroom achievement test which is administered once. Internal consistency is estimated via the split-half reliability index, coefficient alpha (Cronbach, 1951) index or the Kuder-Richardson formula 20 (KR-20) (Kuder & Richardson, 1937) index.

Inter-rater reliability: Different raters, using a common rating form, measure the object of interest consistently. Inter-rater agreement answers the question, “ Are the raters consistent in their ratings?” One expects that the reliability coefficient will be high, if the observers rated similarly. For example, three senior sales trainers rating the closing skills of a novice sales representative or master teachers rating the teaching effectiveness of a first or second year teacher.

At this point, it is important to understand the two main questions reliability helps to answer;

- What is considered a ‘ good’ or ‘ adequate’ value? and
- How does one improve the reliability of a survey instrument?

The general convention in research has been prescribed by Nunnally and Bernstein (1994) who state that one should strive for reliability values of . 70 or higher. Reliability values increase as test length increases (Gulliksen, 1950) That is, the more items you have in your scale to measure the construct of interest the more reliable your scale will become.

Various indices for assessing the reliability of measures have been proposed. We shall look at them in relations to the type of reliability they measure.

Various Reliability Indices

Parallel forms procedure: This procedure measures Equivalence. Here, one administers alternative forms of the same measure to either the same group or different group of respondents. This administration of the various forms occurs at the same time or following some time delay. The higher the degree of correlation between the two forms, the more equivalent they are. In practice the parallel forms procedure is seldom implemented, as it is difficult, if not impossible, to verify that two tests are indeed parallel (i. e., have equal means, variances, and correlations with other measures). Indeed, it is difficult enough to have one well-developed instrument to measure the construct of interest let alone two. Another situation in which equivalence will be important is when the measurement process entails subjective judgments or ratings being made by more than one person.

Test-retest procedure: This is used to measure stability. It involves administering the same measurement instrument to the same individuals under the same conditions after some period of time. Test-rest reliability is estimated with correlations between the scores at Time 1 and those at Time 2 (to Time x). Two assumptions underlie the use of the test-retest procedure. The first required assumption is that the characteristic that is measured does not change over the time period. The second assumption is that the time period is long enough that the respondents’ memories of taking the test at Time 1 does not influence their scores at the second and subsequent test administrations.

Split-half reliability index: This index is used in estimating internal consistency. The split-half estimate entails dividing up the test into two parts (e. g., odd/even items or first half of the items/second half of the items), administering the two forms to the same group of individuals and correlating the responses.

Other indices for measuring internal consistency are the coefficient alpha (Cronbach) index, the Kuder-Richardson formula 20 (KR-20) and the (Kuder & Richardson, 1937) index. This indices represent the average of all possible split-half estimates. The difference between the two is when they would be used to assess reliability. Specifically, coefficient alpha is typically used during scale development with items that have several response options (i. e., 1 = strongly disagree to 5 = strongly agree) whereas KR-20 is used to estimate reliability for dichotomous (i. e., Yes/No; True/False) response scales.

Validity

Validity examines how truthful the research results are. It is the extent to which the instrument measures what it purports to measure. In other words, does the research instrument allow you to hit “ the bull’s eye” of your research object? For example, a test that is used to screen applicants for M. sc admissions into a UK university is valid if its scores are directly related to future academic performance of students either in research thesis or course work. Researchers generally determine validity by asking a series of questions, and will often look for the answers in the research of others.

There are various types of validity. They include: content validity, face validity, criterion-related validity (or predictive validity), construct validity, factorial validity, concurrent validity, convergent validity and divergent (or discriminant validity). The four most discussed types will be explained here.

Construct validity:-

Wainer and Braun (1998) describe the validity in quantitative research as “ construct validity”. The construct is the initial concept, notion, question or hypothesis that determines which data is to be gathered and how it is to be gathered.

Construct validity is the degree to which an instrument measures the trait or theoretical construct that it is intended to measure. For example, if one were to develop an instrument to measure intelligence that does indeed measure IQ, then this test is construct valid.

Content validity:-

This pertains to the degree to which the instrument fully assesses or measures the construct of interest. For example, say we are interested in evaluating employees' attitudes toward a training program within an organization. We would want to ensure that our questions fully represent the domain of attitudes toward the training program. The development of a content valid instrument is typically achieved by a rational analysis of the instrument by raters (ideally 3 to 5) familiar with the construct of interest (Michael J. Miller 2004)

Face validity:-

This a component of content validity and is established when an individual reviewing the instrument concludes that it measures the characteristic or trait of interest. For instance, if a quiz in this class comprised items that asked questions pertaining to research methods you would most likely conclude that it was face valid. In short, it looks as if it is indeed measuring what it is designed to measure.

Criterion-related:-

Criterion related validity is assessed when one is interested in determining the relationship of scores on a test to a specific criterion. An example is that scores on a job test for fresh graduate should be related to relevant criteria such as youth service corps completion(for students in Nigeria), class of degree, etc. Conversely, an instrument that measures colour inclinations would most assuredly demonstrate very poor criterion-related validity with respect to graduate job placement.

Conclusion

In conclusion, it is important to note that one’s ability to answer a research question is only as good as the instruments developed or ones method of data collection. A well-developed survey instrument will better provide a researcher with quality data with which to answer a question or solve a problem. Finally, recall that for something to be valid it must be reliable but it must also measure what it is intended to measure.