REVIEW ARTICLE

GOODNESS OF MEASUREMENT: RELIABILITY AND VALIDITY

Shweta Bajpai[1], Ram Bajpai[2]
[1] Department of Psychology, National University of Study and Research in Law, Ranchi, Jharkhand, India
[2] Department of Community Medicine, Army College of Medical Sciences, New Delhi, India

Correspondence to: Ram Bajpai (rambajpai@hotmail.com)

DOI: 10.5455/ijmsph.2013.191120133 **Received Date: 07.10.2013** **Accepted Date: 19.01.2014**

ABSTRACT
The two most important and fundamental characteristics of any measurement procedure are reliability and validity and lie at the heart of competent and effective study. However, these phenomena have often been somewhat misunderstood or under emphasized. How productive can be any research, if the instrument used does not actually measure what it purports to? How justifiable the research that is based on an inconsistent instrument? What constitutes a valid instrument? What are the implications of proper and improper testing? This paper attempts to explore these measurement related concepts as well as some of the issues pertaining thereto.

Key-Words: Item Analysis; Error; Validity; Reliability; Alpha

Introduction

Across disciplines, researchers often not only fail to report the reliability of their measures, but also fall short of grasping the inextricable link between scale validity and effective research.[1,2] It is important to make sure that the instrument we developed to measure particular concept is indeed accurately measuring the variable i.e., we are actually measuring the concept that we supposed to measure. The scales developed can often be imperfect, and errors are prone to occur in the measurement of scale variables. The use of better instrument will ensure more accuracy in results, which will enhance the scientific quality of research. Hence, we need to assess the "goodness" of the measures developed and reasonably sure that the instrument measures the variables they are supposed to, and measures them accurately.[3] First, an item analysis of the responses to the questions tapping the variable is carried out and then the reliability and the validity of the measures are established.

True Score Model

Multi-item scales should be evaluated for accuracy, reliability and applicability. Measurement accuracy refers to capturing the responses as the respondent intended them to be understood. Errors can result from either systematic error, which affects the observed score in the same way on every measurement, or random error, which varies with every measurement.[4] This model provides a framework for understanding the accuracy of measurement. According to this model;

$$\boldsymbol{X_O = X_T + X_S + X_R} \qquad \textbf{(1)}$$

Where, X_O = the observed score or measurement; X_T = the true score of the characteristic; X_S = systematic error; X_R = random error

If the random error in equation (1) is zero then instrument is termed as reliable and if both systematic error as well as random error are zero then instrument considered as valid. The total measurement error is the sum of the systematic error, which affects the model in a constant fashion, and the random error, which affects the model randomly. Systematic errors occur due to stable factors which influence the observed score in the same way on every occasion that a measurement is made. However, random error occurs due to transient factors which influence the observed score differently each time.[4]

Item Analysis

Item analysis is carried out to see if the items in the instrument belong there or not. Each item is examined for its ability to discriminate between those subjects whose total scores are high and those with low scores. In item analysis, the means between the high-score group and the low-score group are tested to detect significant differences through the t-values. The item with a high t-value are then included in the instrument.[3] Thereafter, tests for the reliability of the instrument are carried out and the validity of the measure is established. The various forms of reliability and validity are depicted in Figure 1.

Validity

Validity is a test of how well an instrument that is developed measures the particular concept it is intended to measure as shown in Figure 2. In other words, validity concerned with weather we measure the right concept or not.[3] For example, when we ask a set of questions with the hope that we are tapping the concept, how can we be reasonably certain that we are indeed measuring the concept we set out to measure and something else? This can be determined by applying certain validity tests. Several types of validity test are used to test the goodness of measures and writers use different terms to denote them.

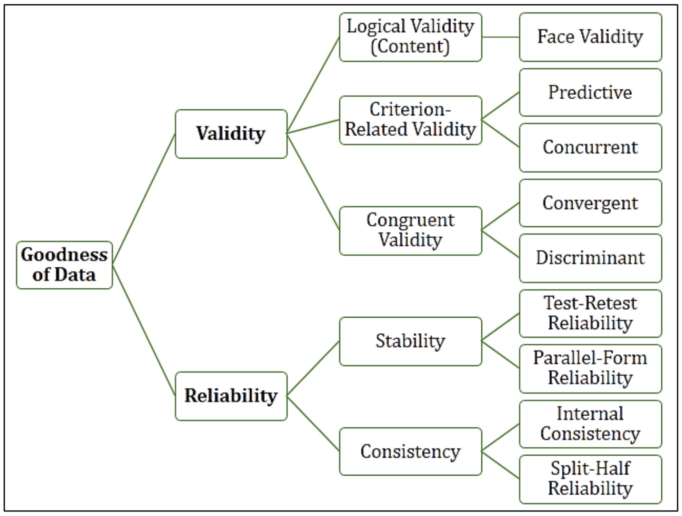

Figure-1: Forms of reliability and validity

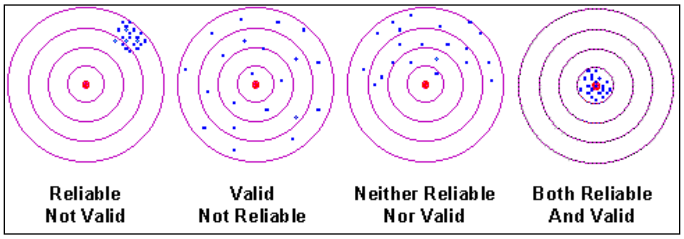

Figure-2: Validity and reliability of instrument

Content Validity

It ensures that the measure includes an adequate and representative set of items that tap the concept. The more the scale items represent the domain or universe of the concept being measured, the greater the content validity. It is a function of how well the dimensions and elements of a concept have been delineated. The development of content valid of an instrument is typically achieved by a rational analysis of the instrument by raters (ideally 3 to 5) familiar with the construct of interest. Specifically, raters will review all of the items for readability, clarity and comprehensiveness and come to some level of agreement as to which items should be included in the final instrument. In short, a panel of judges can attest to the content validity of the instrument.

Crocker and Algina suggest employing the following four steps to effectively evaluate content validity: (i) identify and outline the domain of interest; (ii) gather resident domain experts; (iii) develop consistent matching methodology; and (iv) analyze results from the matching task.[5]

Face Validity: it is considered as a basic and minimum index of content validity. Face validity indicates the items that are intended to measure a concept, do, on the face of it, look like they measure the concept. In brief, it looks as if it is indeed measuring what it is designed to measure.

Criterion-Related Validity

It is established when the measure differentiates individuals on a criterion it is expected to predict. This can be done by establishing concurrent validity or predictive validity.

Concurrent Validity: It is established when the scale discriminates individuals who are known to be different i.e., they should score differently on the instrument. In other words, the degree to which an instrument can distinguish individuals who differ on some criterion measured or observed at the same time. For example, based on current observed behaviours, who should be released from an institution?

Predictive Validity: It indicates the ability of the measuring instrument to differentiate among individuals with reference to a future criterion. In other words, how adequately will an instrument be in differentiating between the performance and behaviour of individuals on some future criterion? For example, how well do GRE scores predict future grades?

Construct Validity

Construct validity testifies to how well the results obtained from the use of the measure fit the theories around which the test is designed. This is assessed through convergent and discriminate validity. For example, if one were to develop an instrument to measure intelligence that does indeed measure IQ, than this test is valid. Construct validity is very much an ongoing process as one refines a theory, if necessary, in order to make predictions about test scores in various settings and situations.

Crocker and Algina (1986) provide a series of steps to follow when pursuing a construct validation study[5]: (i) Generate hypotheses of how the construct should relate to both other constructs of interest and relevant group differences; (ii) Choose a measure that adequately represents the construct of interest; (iii) Pursue empirical study to examine the relationships hypothesized; and (iv) Analyze gathered data to check hypothesized relationships and to assess whether or not alternative hypotheses could explain the relationships found between the variables.

Convergent Validity: It is established when the scores obtained with two different instruments measuring the same concept are highly correlated.

Discriminant Validity: It is established when, based on theory, two variables are predicted to be uncorrelated, and the scores obtained by measuring them are indeed empirically found to be so.

There are following methods could be used to check construct validity of the instrument: (1) Correlational analysis; (2) Factor analysis; and (3) The multitrait, multimethod matrix of correlations

Finally, it is important to note that validity is a necessary but not sufficient condition of the test of goodness of a measure. A measure should not only be valid but also reliable.

Reliability

If a measurement device or procedure consistently assigns the same score to individuals or objects with equal values, the instrument is considered reliable. In other words, the reliability of a measure indicates the extent to which it is without bias and hence insures consistent measurement cross time and across the various items in the instruments as given in Figure 2.

It is an indication of the stability (or repeatability) and consistency (or homogeneity) with which the instrument measures the concept and helps to assess the "goodness" of a measure.[3,6]

This property is not a stagnant function of the test. Rather, reliability estimates change with different populations (i.e. population samples) and as a function of the error involved. More important to understand is that reliability estimates are a function of the test scores yielded from an instrument, not the test itself.[2] Low internal consistency estimates are often the result of poorly written items or an excessively broad content area of measure.[5] However, other factors can equally reduce the reliability coefficient, namely, the homogeneity of the testing sample, imposed time limits in the testing situation, item difficulty and the length of the testing instrument.[5-9]

Stability

It is defined as the ability of a measure to remain the same over time despite uncontrolled testing conditions or respondent themselves. Two methods to test stability are test-retest reliability and parallel-form reliability.

Test-Retest Reliability: The reliability coefficient obtained by repetition of the same measure on a second time is called the test-retest reliability. When a questionnaire containing some items that are supposed to measure a concept is administered to a set of respondents now, and after some time ranging from few days to weeks or months, again administered on the same respondents. The correlation coefficient calculated between two set of data and if it found to be high, better the test-retest reliability.

Parallel-Form Reliability: when responses on two comparable sets of measures tapping the same construct are highly correlated, known as parallel-form reliability. It similar to test-retest method except order or sequence of questions or sometimes wording of questions has changed at second time. If two such comparable forms are highly correlated (say more than 0.7), we may be fairly certain that the measures are reasonably reliable.

Internal Consistency

The internal consistency of measures is indicative of the homogeneity of the items in the measure that tap the construct. In other words, the items should "hang together as a set", and be capable of independently measuring the same concept. Consistency can be examined through the inter-item consistency and split-half reliability.

Inter-Item Consistency: It is a test of consistency of respondents' answers to all concepts, they will be correlated with one another.

a. *Cronbach's alpha:* In statistics, Cronbach's alpha (α) is a coefficient of internal consistency and widely used in social sciences, business, nursing, and other disciplines. Cronbach's alpha is actually an average of all the possible split-half reliability estimates of an instrument and it is commonly used as an estimate of the reliability of a psychometric test for a sample of examinees.[10] It was first named alpha by Lee Cronbach in 1951, as he had intended to continue with further coefficients.[10] The measure can be viewed as an extension of the Kuder–Richardson Formula 20 (KR-20), which is an equivalent measure for dichotomous items. Cronbach's Alpha is not robust against missing data. Suppose that we measure a quantity which is a sum of $\boldsymbol{K}$ components (*K-items* or *testlets*): $\boldsymbol{X = Y_1 + Y_2 + \ldots\ldots + Y_k}$. Cronbach's α is defined as

$$\boldsymbol{\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_{Y_i}^2}{\sigma_X^2}\right)} \qquad \textbf{(2)}$$

Where, $\boldsymbol{\sigma_X^2}$ is variance of observed total test scores and $\boldsymbol{\sigma_{Y_i}^2}$ is variance of component $\boldsymbol{i}$ for the current sample of persons. The theoretical value of alpha varies from zero to 1, since it is the ratio of two variances. However, depending on the estimation procedure used, estimates of alpha can take on any value less than or equal to 1, including negative values, although only positive values make sense. Higher values of alpha are more desirable. Some professionals, as a rule of thumb, require a reliability of 0.70 or higher (obtained on a substantial sample) before they will use an instrument.**[4]** As a result, alpha is most appropriately used when the items measure different substantive areas within a single construct.**[11-13]**

b. *Kuder–Richardson Formula 20 (KR-20):* KR-20, first published in 1937 is a measure of internal consistency reliability for measures with dichotomous choices and it is analogous to Cronbach's α.**[14,15]** A high KR-20 coefficient (e.g.,>0.90) indicates a homogeneous test. Values can range from 0.00 to 1.00 (sometimes expressed as 0 to 100), with high values indicating that the examination is likely to correlate with alternate forms (a desirable characteristic). The KR-20 may be affected by difficulty of the test, the spread in scores and the length of the examination. Since Cronbach's α was published in 1951, there has been no known advantage to KR-20 over Cronbach alpha. 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.

Split-Half Reliability: It reflects the correlations between two halves of an instrument. The estimates will vary depending on how the items in measure are split into two halves. Split-half reliabilities may be higher than Cronbach's alpha only in the circumstances of there being more than one underlying responses dimension tapped by measure and when certain other conditions are met as well.

Conclusion

In health care and social science research, many of the variables of interest and outcomes that are important derived from abstract concepts are known as theoretical constructs. Therefore, using tests or instruments that are valid and reliable to measure such constructs is a crucial component of research quality. Reliability and validity of instrumentation should be important considerations for researchers in their investigations. Well-trained and motivated observers or a well-developed survey instrument will better provide quality data with which to answer a question or solve a problem. In conclusion, remember that your ability to answer your research question is only as good as the instruments you develop or your data collection procedure.

References

1. Henson RK. Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha. Meas Eval Coun Dev. 2001;34:177-188.
2. Thompson B. Understanding reliability and coefficient alpha, really. In: Thompson B, editor. Score Reliability: Contemporary Thinking on Reliability Issues. Thousand Oaks, CA: Sage; 2003. p. 5.
3. Shekharan U, Bougie R. Research Methods for Business: A Skill Building Approach. 5th ed. New Delhi: John Wiley. 2010
4. Malhotra NK. Marketing Research: An Applied Orientation. 4th ed. New Jersey: Pearson Education, Inc. 2004.
5. Crocker L, Algina J. Introduction to Classical and Modern Test Theory. Philadelphia: Harcourt Brace Jovanovich College Publishers. 1986.
6. Zikmund WG. Business Research Methods. 7th ed. Ohio: Thompson South-Western. 2003.
7. Mehrens WA, Lehman IJ. Measurement and Evaluation in Education and Psychology. 4th ed. Orlando, Florida: Holt, Rinehart and Winston Inc. 1991.
8. Devellis RF. Scale Development: Theory and Applications, Applied Social Research Methods Series. Newbury Park: Sage. 1991.
9. Gregory RJ. Psychological Testing: History, Principles and Applications. Boston: Allyn and Bacon. 1992.
10. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297–334.
11. Tavakol M and Dennick R. Making sense of Cronbach's alpha. Int Jour Med Edu. 2011;2:53-55.
12. Zinbarg R, Revelle W, Yovel I, Li W. Cronbach's, Revelle's, and McDonald's: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika. 2005;70:123–133.
13. McDonald RP. Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum Associates Inc; 1999.
14. Kuder GF, Richardson MW. The theory of the estimation of test reliability. Psychometrika. 1937;2:151–160.
15. Cortina JM. What Is Coefficient Alpha - an Examination of Theory and Applications. Journal of Applied Psychology. 1993;78:98–104.

Cite this article as: Bajpai SR, Bajpai RC. Goodness of Measurement: Reliability and Validity. Int J Med Sci Public Health 2014;3:112-115.
Source of Support: Nil
Conflict of interest: None declared