



Reliability and Validity Issues in Research

Ganesh Thanasegaran
Department Of Management & Marketing
Faculty Of Economics & Management
Universiti Putra Malaysia
E-mail: ganesh@econ.upm.edu.my

ABSTRACT

Instrument validity and reliability lie at the heart of competent and effective study. However, these phenomena have often been somewhat misunderstood or under emphasized. How productive can any research be if the instrument used does not actually measure what it purports to? How legitimate or justifiable is 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. In so doing, it is hoped that the questions raised may, to some extent at least, be answered.

Keywords: Reliability, Validity, Estimates

Introduction

Across disciplines, competent researchers often not only fail to report the reliability of their measures (Henson, 2001; Thompson, 1999), but also fall short of grasping the inextricable link between scale validity and effective research. At best, measurement error affects the ability to find significant results in one's data. At worst, measurement error can significantly damage the interpretability of scores or the function of a testing instrument.

What is Reliability?

Reliability is the degree to which measures are free from error and therefore yield consistent results (i.e. the consistency of a measurement procedure). If a measurement device or procedure consistently assigns the same score to individuals or objects with equal values, the instrument is considered reliable. Reliability involves the consistency, or reproducibility, of test scores i.e., the degree to which one can expect relatively constant deviation scores of individuals across testing situations on the same, or parallel, testing instruments.

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. These facts underscore the importance of consistently reporting reliability estimates for each administration of an instrument, as test samples, or subject populations, are rarely the same across situations and in different research settings. More important to understand is that reliability estimates are a function of the test scores yielded from an instrument, not the test itself (Thompson, 1999). Accordingly, reliability estimates should be considered based upon the various sources of measurement error that may be involved in test administration (Crocker & Algina, 1986). Two dimensions underlie the concept of reliability: repeatability (or stability over time) and internal consistency (or homogeneity of the measure) (Zikmund, 2003 p 300).

Any remaining errors or omissions rest solely with the author(s) of this paper.

RELIABILITY ISSUES IN RESEARCH

Reliability Estimation

Repeatability, or stability-over-time reliability, may be measured with the test-retest method, whereby the same scale or measure is administered to the same respondents at two separate points in time (Zikmund, 2003 p 300), i.e. comparing the scores from repeated testing of the same participants with the same test. Reliable measures should produce very similar scores, e.g. IQ tests typically show high test-retest reliability. However, test-retest procedures may not be useful when participants may be able to recall their previous responses and simply repeat them upon retesting.

Internal consistency, or homogeneity, may be measured by using either the split-half method, alternate-form method, or Cronbach's alpha method. The split-half method is one that measures the degree of internal consistency by checking one half of the results of a set of scaled items against the other half, i.e. comparing scores from different parts of the test. The method demands equal item representation across the two halves of the instrument. Clearly the comparison of dissimilar sample items will not yield an accurate reliability estimate. One can ensure equal item representation through the use of random item selection, matching items from one half to the next, or assigning items to halves based on an even/odd distribution (Crocker & Algina, 1986).

The alternate-form method is one that measures the correlation between alternative instruments, designed to be as equivalent as possible, administered to the same group of subjects (Zikmund, 2003 pp 300-301), i.e. by comparing scores from alternate forms of the test. In cases where administering the exact same test will not necessarily be a good test of reliability, we may use equivalent/alternate forms reliability. As the name implies, two or more versions of the test are constructed that are equivalent in content and level of difficulty, e.g. professors use this technique to create makeup or replacement exams because students may already know the questions from the earlier exam.

The most common method of assessing internal consistency reliability estimates is through the use of coefficient alpha. Though there are three different measures of coefficient alpha, the most widely used measure is Cronbach's coefficient alpha. Cronbach's alpha is actually an average of all the possible split-half reliability estimates of an instrument (Crocker & Algina, 1986; DeVellis, 1991; Gregory, 1992; Henson, 2001). Cronbach's alpha is a reliability coefficient that measures inter-item reliability or the degree of internal consistency/homogeneity between variables measuring one construct/concept i.e. the degree to which different items measuring the same variable attain consistent results. This coefficient varies from 0 to 1 and a value of 0.6 or less generally indicates unsatisfactory internal consistency reliability (Malhotra, 2004). In the social sciences, acceptable reliability estimates range from .70 to .80 (Nunnally & Bernstein, 1994).

When it is impractical or inadvisable to administer two tests to the same participants, it is possible to assess the reliability of some measurement procedures by examining their internal consistency. This type of reliability assessment is useful with tests that contain a series of items intended to measure the same attribute. Scores on different items designed to measure the same construct should be highly correlated, e.g. math tests often require solving several examples of the same type of problem - scores on these questions will normally represent an ability to solve this type of problem.

Factors Affecting Reliability

Low internal consistency estimates are often the result of poorly written items or an excessively broad content area of measure (Crocker & Algina, 1986). 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 (Crocker & Algina, 1986; Mehrens & Lehman, 1991; DeVellis, 1991; Gregory, 1992).

Group homogeneity is particularly influential when one is trying to apply a norm-referenced test to a homogenous test sample. In such circumstances, the restriction of range of the test group (i.e. low variability) translates into a smaller proportion of variance explained by the test instrument, ultimately deflating the reliability coefficient. It is essential to bear in mind the intended use of the instrument when considering these circumstances and deciding how to use an instrument (Crocker & Algina, 1986; Mehrens & Lehman, 1991; Gregory, 1992).

Imposed time constraints in a test situation pose a different type of problem, i.e. time limits ultimately affect a test taker's ability to fully answer questions or to complete an instrument. As a result, variance in test taker ability to work at a specific rate becomes enmeshed in that person's score variance. Ultimately, test takers who work at similar rates have higher degrees of variance correspondence, artificially inflating the reliability of the testing instrument. Clearly, this situation becomes problematic when the construct that an instrument intends to measure has nothing to do with speed competency (Crocker & Algina, 1986; Mehrens & Lehman, 1991; Gregory, 1992).

The relationship between reliability and item difficulty addresses a variability issue once again. Ultimately, if the testing instrument has little to no variability in the questions (i.e. either items are all too difficult or too easy), the reliability of scores will be affected. Aside from this, reliability estimates can also be artificially deflated if the test has too many difficult items, as they promote uneducated guesses (Crocker & Algina, 1986; Mehrens & Lehman, 1991).

Lastly, test length also factors into the reliability estimate. Simply, longer tests yield higher estimates of reliability. However, one must consider the reliability gains earned in such situations, as infinitely long tests are not necessarily desirable.

Reliability & Power

Assessing scale reliability is crucial to maximizing power in one's study. Simply put, unreliable scales decrease the statistical power of an instrument. This is important in many ways. Most notably, as power decreases, larger sample sizes are necessary to find significant results. An increase in statistical effect size is also observed with an increase in instrument reliability and subsequent power gained. Additionally, reliable instruments introduce less error into the statistical measurement and resulting analysis. Still, the significant results may well be meaningless if the instrument is faulty (DeVellis, 1991).

What is Validity?

Validity has been defined by "the extent to which [a test] measures what it claims to measure" (Gregory, 1992, p.117). A measure is valid if it measures what it is supposed to measure, and does so cleanly – without accidentally including other factors. The focus here is not necessarily on scores or items, but rather inferences made from the instrument i.e. the behavioral inferences that one can extrapolate from test scores is of immediate focus. In order to be valid, the inferences made from scores need to be "appropriate, meaningful, and useful" (Gregory, 1992, p. 117). These distinctions illuminate the inextricable link between validity and reliability. For example, a testing instrument can *reliably* measure something other than the supposed construct, but an unreliable measure cannot be valid (Crocker & Algina, 1986; Gregory, 1992). Reliability is a necessary but insufficient (on its own) condition for validity. In other words, a valid instrument must (by definition) be reliable, but a reliable instrument may not necessarily be valid.

Violations of instrument validity severely impact the function and functioning of a testing instrument. In some ways, validity inadequacies impart even more serious consequences on an instrument than its reliability counterpart. This can be substantiated in the sense that validity is a comprehensive construct that cannot be definitively measured in any one given statistic, and that this instrumental testing property is often even less understood than reliability (Crocker & Algina, 1986; Gregory, 1992).

Validity Issues In Research

Effective validity studies not only demand the integration of multiple sources of evidence, but also must continually take place over time, i.e. a measure cannot be deemed valid in a simple instance of study. Rather, multiple studies must be implemented over different samples, and the collection of validity evidence must cover specified areas (Crocker & Algina, 1986; Gregory, 1992; Messick, 1995). Moreover, in recent years researchers have expanded the understanding of validity to comprise more dimensionality than previously recognized.

Specifically, Messick (1995) has criticized traditional approaches to validity, asserting that researchers have narrowly focused attention on compartmentalized distinctions at the expense of fostering the development of a unified concept. Alternatively, he purports an integrative approach to instrument validity

that not only focuses on conventional test score issues, but also emphasizes the significance of score implication and their social use. Still, this unified concept of validity is best understood and examined within the context of its four discrete facets: content validity, construct validity, criterion validity and consequential validity (Messick, 1995).

Content Validity

Content validity considers whether or not the items on a given test accurately reflect the theoretical domain of the latent construct it claims to measure. Items need to effectively act as a representative sample of all the possible questions that could have been derived from the construct (Crocker & Algina, 1986; DeVellis, 1991; Gregory, 1992).

In the social sciences where theories and constructs involved are innately intangible (e.g. anxiety, intelligence, depression, etc.), their measurement depends on the operationalisation of variables deemed to be representative of the domain. In this respect, there is no clean set of exhaustive measures that represents any given construct. Rather, there exists an almost infinite sampling domain from which questions can be drawn. In this instance, content validity becomes more of a qualitative judgment than an absolute definitive measure (Crocker & Algina, 1986; DeVellis, 1991; Gregory, 1992).

Crocker and Algina (1986) suggest employing the following four steps to effectively evaluate content validity: 1) identify and outline the domain of interest, 2) gather resident domain experts, 3) develop consistent matching methodology, and 4) analyze results from the matching task.

Construct Validity

The construct validity of a measure “is directly concerned with the theoretical relationship of a variable (e.g. a score on some scale) to other variables. It is the extent to which a measure ‘behaves’ the way that the construct it purports to measure should behave with regard to established measures of other constructs” (DeVellis, 1991). Previously referred to as “congruent validity”, the term “construct validity” was first formulated by a subcommittee (Meehl, P.E. and Challman, M.C.) of the American Psychologists Association’s Committee on Psychological Tests (Cronbach & Meehl, 1955).

In practice, as constructs are not readily observable, items or variables, that act as representations of the construct and serve to measure examinee scores with respect to the paradigm, must be developed. This is facilitated through the delineation of the construct itself, i.e. a construct needs to be both operationalised and syntactically defined in order to measure it effectively (Benson, 1998; Crocker & Algina, 1986; Gregory, 1992).

The operationalisation of the construct involves developing a series of measurable behaviors or attributes that are hypothesized to correspond to the latent construct. Defining the construct syntactically involves establishing hypothesized relationships between the construct of interest and other related constructs or behaviors (Benson, 1998; Crocker & Algina, 1986; Gregory, 1992).

Crocker and Algina (1986) provide a series of steps to follow when pursuing a construct validation study: 1) generate hypotheses of how the construct should relate to both other constructs of interest and relevant group differences, 2) choose a measure that adequately represents the construct of interest, 3) pursue empirical study to examine the relationships hypothesized, and 4) analyze gathered data to check hypothesized relationships and to assess whether or not alternative hypotheses could explain the relationships found between the variables.

Different validity coefficients are yielded as a result of investigations into construct validity. When comparing the measured construct to other constructs based on hypothesized relationships, one expects to see either convergent or discriminant validity, i.e. convergent validity coefficients should arise when considering two constructs hypothesized to be related. Further, the correlation between the two should be moderately high in order to contend test validity, as we want the measure to be correlated to other scales purporting to measure the same, or related, thing. In contrast, when one is looking at the relationship between the scale of interest and a construct that is not hypothesized to be related, we aim to find discriminant validity coefficients. In other words, we want little to no correlation between our measure and unrelated constructs.

Criterion Validity

Criterion validity refers to the ability to draw accurate inferences from test scores to a related behavioral criterion of interest. This validity measure can be pursued in one of two contexts: predictive validity or concurrent validity. In criterion-oriented validity, the investigator is primarily interested in some criterion which he wants to predict. If the criterion is obtained some time after the test is given, predictive validity is being studied. However, if the test scores and criterion are “determined at essentially the same time”, then concurrent validity is being examined (Cronbach & Meehl, 1955).

In predictive validity, researchers are interested in assessing the predictive utility of an instrument. For example, the purpose of standardized testing such as the SAT and GMAT are to predict performance outcomes in college and graduate business school respectively. The test scores ultimately become the basis from which decisions are made (Crocker & Algina, 1986; Gregory, 1992).

Researchers look for a high degree of correlation between the criterion variable and scores on the testing instrument, in order to assert good criterion validity. Validity coefficients are ultimately derived from the correlation between these components. From this, one can calculate a coefficient of determination for the measures by squaring the validity coefficient. This tells the researcher what percentage of variance in the criterion variable is accounted for by the testing measure, or predictor variable (Crocker & Algina, 1986; Gregory, 1992).

Though concurrent validity also looks at the correlation between criterion and test scores, the two measures are taken one right after the other in this instance, i.e. there is no extended time period between the testing measure and the criterion measure, and the examiner is more interested in the measure’s ability to reflect current ability than any predictive aptitude. Measures with established concurrent validity ultimately have the ability to act as “shortcut[s] for obtaining information that might otherwise require the extended investment of professional time”, i.e. validated measures may be used as precursory diagnostic tools in professional settings (Gregory, 1992).

Consequential Validity

In recent years, more emphasis has been placed on the social utility and bias of interpretation in test scores. Messick (1995) has been at the forefront of this push for the consideration of consequential validity within the context of a measure’s construct validity.

Consequential validity refers to the notion that the social consequences of test scores and their subsequent interpretation should be considered not only with the original intention of the test, but also cultural norms (Messick, 1995). This idea points to both the intended and unintended consequences of a measure, which may be either positive or negative.

Consider the deleterious effects of an invalid instrument that creates unmerited distinctions between racial or ethnic groups on an academic measure. It is critical that distinctions of this kind are supported with valid measurement instruments, particularly if they go against the grain of social norms.

This concern is wholly based in the scale’s construct validity. Moreover, any inferences made from test scores are contingent upon the construct validity of the scale. Therefore adequate evidence, as collected through all four facets of validity study, is crucial in developing an argument for the value of test scores (Messick, 1995).

General Factors Affecting Validity

An integral issue at hand in establishing validity coefficients is the actual relationship between the two variables, or constructs, that one is interested in. Beyond this, comparable measurement issues that affected the nature of reliability coefficients also affect validity coefficients, i.e. the more heterogeneous the groups are, the higher the correlations between two measures will ultimately be.

This phenomenon is most readily observable in samples with a restriction of range problem. When the data range is limited, the scores become more homogenous and the resulting correlation coefficients derived are artificially inflated. An important point to note is that the more effective an instrument is at screening individuals for a particular purpose, the less heterogeneous the resulting sample will be, which in turn results in a smaller validity coefficient.

Conclusion

Reliability and validity of instrumentation should be important considerations for researchers in their investigations. The goal of achieving measurement validity and reliability can be accomplished partly by a push for quality item writing, an insistence on reporting reliability data across studies, sound theoretical bases for construct measurement and the accurate operationalisation of constructs.

This objective imparts a direct responsibility on behalf of all examiners in a given field, i.e. it is essential for researchers to actively measure the reliability and validity of instrument scores over populations and time. The continual nature of both these processes should not be undermined or overlooked. Moreover, it is critical for this type of information to be easily accessible in order to facilitate the understanding and sharing of this knowledge. Without credible instrumentation that is monitored and measured over time, research results can become meaningless.

References

- Benson, J. (1998) Developing a Strong Program of Construct Validation: A Test Anxiety Example, *Educational Measurement: Issues and Practice*, **17**: 10-17.
- Crocker, L., and Algina, J. (1986) *Introduction to Classical and Modern Test Theory*, Harcourt Brace Jovanovich College Publishers: Philadelphia
- Cronbach, L.J., & Meehl, P.E. (1955) *Construct Validity in Psychological Tests*, *Psychological Bulletin*, **52**: 281-302.
- Devellis, R.F. (1991) *Scale Development: Theory and Applications*, *Applied Social Research Methods Series 26*, Sage: Newbury Park.
- Gregory, R.J. (1992) *Psychological Testing: History, Principles and Applications*, Allyn and Bacon: Boston
- Henson, R.K. (2001) Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha, *Measurement and Evaluation in Counseling and Development*, **34**:177-188.
- Malhotra, N.K. (2004) *Marketing Research: An Applied Orientation* (4th edn) Pearson Education, Inc: New Jersey.
- Mehrens, W.A., and Lehman, I.J. (1991) *Measurement and Evaluation in Education and Psychology*, (4th edn) Holt, Rinehart and Winston Inc: Orlando, FL.
- Messick, S. (1995) Validity of Psychological Assessment: Validation of Inferences from Persons Responses and Performances as Scientific Inquiry into Score Meaning, *American Psychologist*, **50** (9):741-749.
- Nunnally, J.C., and Bernstein, I.H. (1994) *Psychometric Theory*, (3rd edn), Mcgraw-Hill: New York
- Thompson, B. (1999) Understanding Coefficient Alpha, Really, *Paper Presented at the Annual Meeting of the Education Research Exchange*, College Station, Texas, February 5, 1999.
- Zikmund, W.G. (2003) *Business Research Methods*, (7th edn), Thompson South-Western: Ohio