Education Research and Perspectives, Vol.38, No.1

Validity and Reliability in Social Science Research

Ellen A. Drost[†]

California State University, Los Angeles

Concepts of reliability and validity in social science research are introduced and major methods to assess reliability and validity reviewed with examples from the literature. The thrust of the paper is to provide novice researchers with an understanding of the general problem of validity in social science research and to acquaint them with approaches to developing strong support for the validity of their research.

Introduction

An important part of social science research is the quantification of human behaviour — that is, using measurement instruments to observe human behaviour. The measurement of human behaviour belongs to the widely accepted positivist view, or empirical-analytic approach, to discern reality (Smallbone & Quinton, 2004). Because most behavioural research takes place within this paradigm, measurement instruments must be valid and reliable. The objective of this paper is to provide insight into these two important concepts, and to introduce the major methods to assess validity and reliability as they relate to behavioural research. The paper has been written for the novice researcher in the social sciences. It presents a broad overview taken from traditional literature, not a critical account of the general problem of validity of research information.

The paper is organised as follows. The first section presents what reliability of measurement means and the techniques most frequently used to estimate reliability. Three important questions researchers frequently ask about reliability are discussed: (1) what

[†] Address for correspondence: Dr. Ellen Drost, Department of Management, College of Business and Economics, California State University, Los Angeles, 5151 State University Drive, Los Angeles, CA 90032. Email: edrost@calstatela.edu.

affects the reliability of a test?, (2) how can a test be made more reliable?, and (3) what is a satisfactory level of reliability? The second section presents what validity means and the methods to develop strong support for validity in behavioural research. Four types of validity are introduced: (1) statistical conclusion validity, (2) internal validity, (3) construct validity and (4) external validity. Approaches to substantiate them are also discussed. The paper concludes with a summary and suggestions.

Reliability

Reliability is a major concern when a psychological test is used to measure some attribute or behaviour (Rosenthal and Rosnow, 1991). For instance, to understand the functioning of a test, it is important that the test which is used consistently discriminates individuals at one time or over a course of time. In other words, reliability is the extent to which measurements are repeatable – when different persons perform the measurements, on different occasions, under different conditions, with supposedly alternative instruments which measure the same thing. In sum, reliability is consistency of measurement (Bollen, 1989), or stability of measurement over a variety of conditions in which basically the same results should be obtained (Nunnally, 1978).

Data obtained from behavioural research studies are influenced by random errors of measurement. Measurement errors come either in the form of systematic error or random error. A good example is a bathroom scale (Rosenthal and Rosnow, 1991). Systematic error would be at play if you repeatedly weighed yourself on a bathroom scale which provided you with a consistent measure of your weight, but was always 10lb. heavier than it should be. Random error would be at work if the scale was accurate, but you misread it while weighing yourself. Consequently, on some occasions, you would read your weight as being slightly higher and on other occasions as slightly lower than it actually was. These random errors would, however, cancel out, on the average, over repeated measurements on a single person. On the other hand, systematic errors do not cancel out; these contribute to the

mean score of all subjects being studied, causing the mean value to be either too big or too small. Thus, if a person repeatedly weighed him/herself on the same bathroom scale, he/she would not get the exact same weight each time, but assuming the small variations are random and cancel out, he/she would estimate his/her weight by averaging the values. However, should the scale always give a weight that is 10lb. too high, taking the average will not cancel this systematic error, but can be compensated for by subtracting 10lb. from the person's average weight. Systematic errors are a main concern of validity.

There are many ways that random errors can influence measurements in tests. For example, if a test only contains a small number of items, how well students perform on the test will depend to some extent on their luck in knowing the right answers. Also, when a test is given on a day that the student does not feel well, he/she might not perform as strongly as he/she would normally. Lastly, when the student guesses answers on a test, such guessing adds an element of randomness or unreliability to the overall test results (Nunnally, 1978).

In sum, numerous sources of error may be introduced by the variations in other forms of the test, by the situational factors that influence the behaviour of the subjects under study, by the approaches used by the different examiners, and by other factors of influence. Hence, the researcher (or science, in general) is limited by the reliability of the measurement instruments and/or by the reliability with which he/she uses them.

Somewhat confusing to the novice researcher is the notion that a reliable measure is not necessarily a valid measure. Bollen (1990) explains that reliability is that part of a measure that is free of purely random error and that nothing in the description of reliability requires that the measure be valid. It is possible to have a very reliable measure that is not valid. The bathroom scale example described earlier clearly illustrates this point. Thus, reliability is a necessary but not a sufficient condition for validity (Nunnally, 1978).

Estimates of reliability

Because reliability is consistency of measurement over time or stability of measurement over a variety of conditions, the most commonly used technique to estimate reliability is with a measure of association, the correlation coefficient, often termed reliability coefficient (Rosnow and Rosenthal, 1991). The reliability coefficient is the correlation between two or more variables (here tests, items, or raters) which measure the same thing.

Typical methods to estimate test reliability in behavioural research are: test-retest reliability, alternative forms, split-halves, inter-rater reliability, and internal consistency. There are three main concerns in reliability testing: equivalence, stability over time, and internal consistency. These concerns and approaches to reliability testing are depicted in Figure 1. Each will be discussed next.

Test-retest reliability. Test-retest reliability refers to the temporal stability of a test from one measurement session to another. The procedure is to administer the test to a group of respondents and then administer the same test to the same respondents at a later date. The correlation between scores on the identical tests given at different times operationally defines its test-retest reliability.

Despite its appeal, the test-retest reliability technique has several limitations (Rosenthal & Rosnow, 1991). For instance, when the interval between the first and second test is too short, respondents might remember what was on the first test and their answers on the second test could be affected by memory. Alternatively, when the interval between the two tests is too long, maturation happens. Maturation refers to changes in the subject factors or respondents (other than those associated with the independent variable) that occur over time and cause a change from the initial measurements to the later measurements (t and t + 1). During the time between the two tests, the respondents could have been exposed to things which changed their opinions, feelings or attitudes about the behaviour under study.

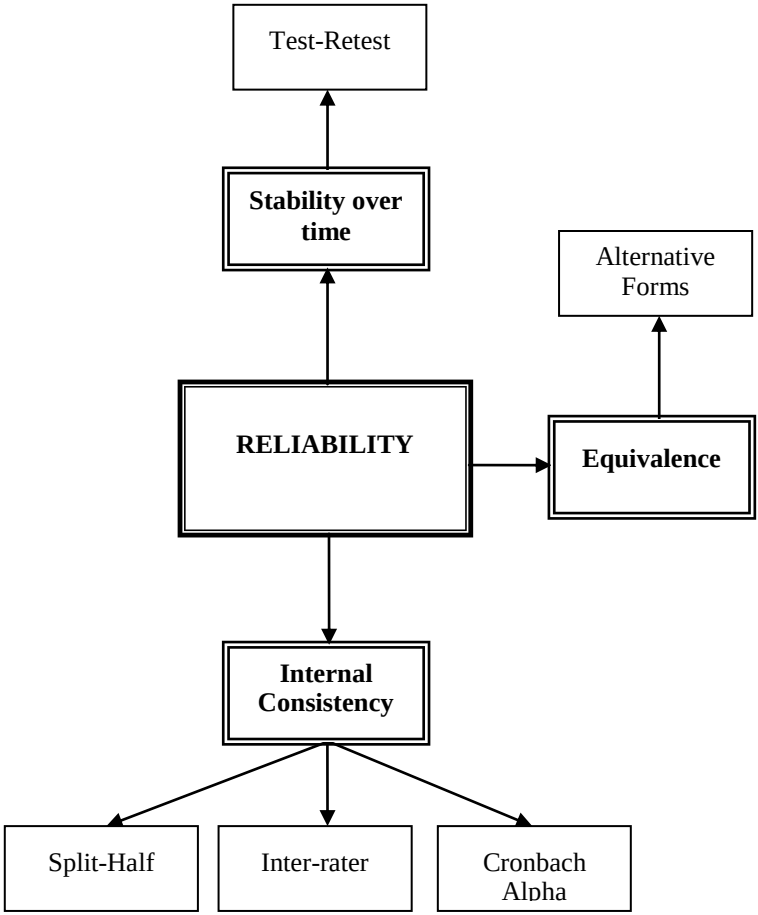

Figure 1. Reliability of Measurement Tests

Alternative forms. The alternative forms technique to estimate reliability is similar to the test retest method, except that different measures of a behaviour (rather than the same measure) are collected at different times (Bollen, 1989). If the correlation between the alternative forms is low, it could indicate that considerable measurement error is present, because two different scales were used. For example, when testing for general spelling, one of the two independently composed tests might not test general spelling but a more subject-specific type of spelling such as business vocabulary. This type of measurement error is then attributed to the sampling of items on the test. Several of the limits of the test-retest method are also true of the alternative forms technique.

Split-half approach. The split-half approach is another method to test reliability which assumes that a number of items are available to measure a behaviour. Half of the items are combined to form one new measure and the other half is combined to form the second new measure. The result is two tests and two new measures testing the same behaviour. In contrast to the test-retest and alternative form methods, the split-half approach is usually measured in the same time period. The correlation between the two halves tests must be corrected to obtain the reliability coefficient for the whole test (Nunnally, 1978; Bollen, 1989).

There are several aspects that make the split-halves approach more desirable than the test-retest and alternative forms methods. First, the effect of memory discussed previously does not operate with this approach. Also, a practical advantage is that the split-halves are usually cheaper and more easily obtained than over time data (Bollen, 1989).

A disadvantage of the split-half method is that the tests must be parallel measures – that is, the correlation between the two halves will vary slightly depending on how the items are divided. Nunnally (1978) suggests using the split-half method when measuring variability of behaviours over short periods of time when alternative forms are not available. For example, the even

items can first be given as a test and, subsequently, on the second occasion, the odd items as the alternative form. The corrected correlation coefficient between the even and odd item test scores will indicate the relative stability of the behaviour over that period of time.

Interrater reliability. When raters or judges are used to measure behaviour, the reliability of their judgments or combined internal consistency of judgments is assessed (Rosenthal & Rosnow, 1991). Below in table format is an example of two judges rating 10 persons on a particular test (i.e., judges rating people's competency in their writing skills).

Judge 1	Rating	Judge 2	Rating
Subject 1	----	Subject 1	----
----	----	----	----
Subject 10	----	Subject 10	----

The correlation between the ratings made by the two judges will tell us the reliability of either judge in the specific situation. The composite reliability of both judges, referred to as effective reliability, is calculated using the Spearman-Brown formula (see Rosenthal & Rosnow, 1991, pp. 51-55).

Internal consistency. Internal consistency concerns the reliability of the test components. Internal consistency measures consistency within the instrument and questions how well a set of items measures a particular behaviour or characteristic within the test. For a test to be internally consistent, estimates of reliability are based on the average intercorrelations among all the single items within a test.

The most popular method of testing for internal consistency in the behavioural sciences is coefficient alpha. Coefficient alpha was popularised by Cronbach (1951), who recognised its general usefulness. As a result, it is often referred to as *Cronbach's alpha*. Coefficients of internal consistency increase as the number of items goes up, to a certain point. For instance, a 5-item test might

correlate .40 with true scores, and a 12-item test might correlate .80 with true scores.

Consequently, the individual item would be expected to have only a small correlation with true scores. Thus, if coefficient alpha proves to be very low, either the test is too short or the items have very little in common. Coefficient alpha is useful for estimating reliability for item-specific variance in a unidimentional test (Cortina, 1993). That is, it is useful once the existence of a single factor or construct has been determined (Cortina, 1993). Next in conclusion of this section, three important questions researchers frequently ask about reliability are considered.

What factors affect the reliability of a test?

There are many factors that prevent measurements from being exactly repeatable or replicable. These factors depend on the nature of the test and how the test is used (Nunnally, 1978). It is important to make a distinction between errors of measurement that cause variation in performance within a test, and errors of instrumentation that are apparent only in variation in performance on different forms of a test.

Sources of error within a test. A major source of error within a test is attributable to the sampling of items. Because each person has the same probability of answering an item correctly, the higher the number of items on the test, the lower the amount of error in the test as a whole. However, error due to item sampling is entirely predictable from the average correlation, thus coefficient alpha would be the correct measure of reliability. Other examples of sources of errors on tests are: guessing on a test, marking answers incorrectly (clerical errors), skipping a question inadvertently, and misinterpreting test instructions.

On subjective tests, such as essay tests, measurement errors are often caused by fluctuations in standards by the individual grader and by the differences in standards of different graders. For example, on an essay examination the instructor might grade all

answers to question 1, then grade all answers to question 2, and so forth. If these scores are independent, then the average correlation among the questions can be used to obtain an accurate estimate of reliability. On the other hand, if half the questions are scored by one person and the other half are independently scored by another person, then the correlation between the two half-tests will provide an estimate of the reliability. Thus, for any test, the sampling of items from a domain includes the sampling of situational factors.

Variation between tests. There are two major sources of error which intervene between administrations of different tests: (1) systematic differences in content of the two tests, and (2) respondents' change with regard to the attribute being measured.

Systematic differences in the content of two tests and in variations in people from one occasion to another cannot be adequately handled by random sampling of items. In this case, the tests should be thought of as random samples of particular occasions, and correlations among tests are allowed to be slightly lower than would be predicted from the correlations among items within tests (Nunnally, 1978). The average correlation among a number of alternative tests completed on different occasions would then be a better estimate of reliability than that given by coefficient alpha for one test administered on one occasion only.

How can I make a test more reliable?

Reliability can be improved by writing items clearly, making test instructions easily understood, and training the raters effectively by making the rules for scoring as explicit as possible (Nunnally, 1978), for instance.

The principal method to make tests more reliable is to make them longer, thus adding more items. For reliability and other reasons in psychometrics, the maxim holds that, other things being equal, a long test is a good test (from Nunnally, p. 243). However, the longer the test, the more likely that boredom and fatigue, among

other factors, can produce attenuation (reduction) in the consistency of accurate responding (Rosenthal & Rosnow, 1991).

What is a satisfactory level of reliability?

A satisfactory level of reliability depends on how a measure is being used. The standard is taken from Nunnally (1978), who suggests that in the early stages of research on predictor tests or hypothesised measures of a construct, reliabilities of .70 or higher will be sufficient. During this stage, Nunnally (1978) maintains that increasing reliabilities much beyond .80 are often wasteful of time and funds, because correlations at that level are attenuated very little by measurement error. To obtain a higher reliability of .90, for instance, requires strenuous efforts at standardisation and probably an addition of items.

On the other hand, in applied settings where important decisions are made with respect to specific test scores, Nunnally (1978) recommends that a reliability of at least .90 is desirable, because a great deal depends on the exact score made by a person on a test. A good example is given for children with low IQs below 70 who are placed in special classes. In this case, it makes a big difference whether the child has an IQ of 65 or 75 on a particular test. Next, the discussion will focus on validity in research.

Validity

Validity is concerned with the meaningfulness of research components. When researchers measure behaviours, they are concerned with whether they are measuring what they intended to measure. Does the IQ test measure intelligence? Does the GRE actually predict successful completion of a graduate study program? These are questions of validity and even though they can never be answered with complete certainty, researchers can develop strong support for the validity of their measures (Bollen, 1989).

There are four types of validity that researchers should consider: statistical conclusion validity, internal validity, construct validity, and external validity. Each type answers an important question and is discussed next.

Statistical conclusion validity

Does a relationship exist between the two variables? Statistical conclusion validity pertains to the relationship being tested. Statistical conclusion validity refers to inferences about whether it is reasonable to presume covariation given a specified alpha level and the obtained variances (Cook & Campbell, 1979). There are some major threats to statistical conclusion validity such as low statistical power, violation of assumptions, reliability of measures, reliability of treatment, random irrelevancies in the experimental setting, and random heterogeneity of respondents.

Internal validity

Given that there is a relationship, is the relationship a causal one? Are there no confounding factors in my study? Internal validity speaks to the validity of the research itself. For example, a manager of a company tests employees on leadership satisfaction. Only 50% of the employees responded to the survey and all of them liked their boss. Does the manager have a representative sample of employees or a bias sample? Another example would be to collect a job satisfaction survey before Christmas just after everybody received a nice bonus. The results showed that all employees were happy. Again, do the results really indicate job satisfaction in the company or do the results show a bias?

There are many threats to internal validity of a research design. Some of these threats are: history, maturation, testing, instrumentation, selection, mortality, diffusion of treatment and compensatory equalisation, rivalry and demoralisation. A discussion of each threat is beyond the scope of this paper.

Construct validity

If a relationship is causal, what are the particular cause and effect behaviours or constructs involved in the relationship? Construct validity refers to how well you translated or transformed a concept, idea, or behaviour – that is a construct – into a functioning and operating reality, the operationalisation (Trochim, 2006). To substantiate construct validity involves accumulating evidence in six validity types: face validity, content validity, concurrent and predictive validity, and convergent and discriminant validity. Trochim (2006) divided these six types into two categories: translation validity and criterion-related validity. These two categories and their respective validity types are depicted in Figure 2 and discussed in turn, next.

Translation Validity. Translation validity centres on whether the operationalisation reflects the true meaning of the construct. Translation validity attempts to assess the degree to which constructs are accurately "translated" into the operationalisation, using subjective judgment – face validity – and examining content domain – content validity.

Face Validity. Face validity is a subjective judgment on the operationalisation of a construct. For instance, one might look at a measure of reading ability, read through the paragraphs, and decide that it seems like a good measure of reading ability. Even though subjective judgment is needed throughout the research process, the aforementioned method of validation is not very convincing to others as a valid judgment. As a result, face validity is often seen as a weak form of construct validity.

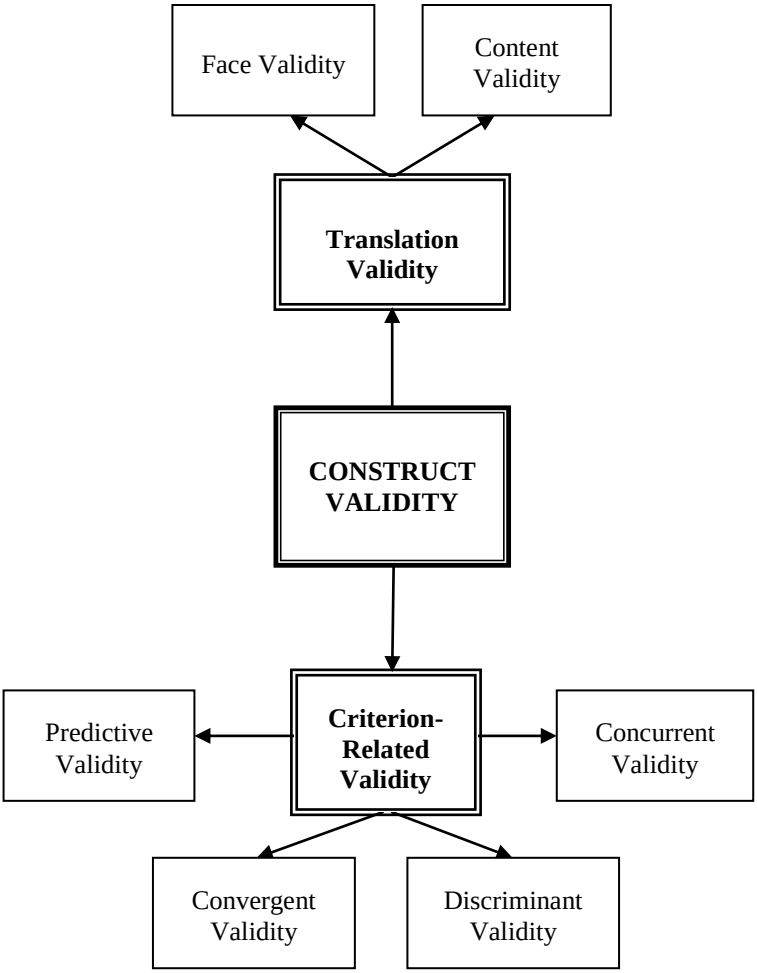

Figure 2. Construct Validity Types

Content validity. Bollen (1989) defined content validity as "a qualitative type of validity where the domain of the concept is made clear and the analyst judges whether the measures fully represent the domain (p.185). According to Bollen, for most concepts in the social sciences, no consensus exists on theoretical definitions, because the domain of content is ambiguous. Consequently, the burden falls on the researcher not only to provide a theoretical definition (of the concept) accepted by his/her peers but also to select indicators that thoroughly cover its domain and dimensions. Thus, content validity is a qualitative means of ensuring that indicators tap the meaning of a concept as defined by the researcher. For example, if a researcher wants to test a person's knowledge on elementary geography with a paper-and-pencil test, the researcher needs to be assured that the test is representative of the domain of elementary geography. Does the survey really test a person's knowledge in elementary geography (i.e. the location of major continents in the world) or does the test require a more advanced knowledge in geography (i.e. continents' topography and their effect on climates, etc.)? There are basically two ways of assessing content validity: (1) ask a number of questions about the instrument or test; and/or (2) ask the opinion of expert judges in the field.

Criterion-related validity. Criterion-related validity is the degree of correspondence between a test measure and one or more external referents (criteria), usually measured by their correlation. For example, suppose we survey employees in a company and ask them to report their salaries. If we had access to their actual salary records, we could assess the validity of the survey (salaries reported by the employees) by correlating the two measures. In this case, the employee records represent an (almost) ideal standard for comparison.

Concurrent Validity and Predictive Validity. When the criterion exists at the same time as the measure, we talk about concurrent validity. Concurrent ability refers to the ability of a test to predict

an event in the present. The previous example of employees' salary is an example of concurrent validity.

When the criterion occurs in the future, we talk about predictive validity. For example, predictive validity refers to the ability of a test to measure some event or outcome in the future. A good example of predictive validity is the use of students' GMAT scores to predict their successful completion of an MBA program. Another example is to use students' GMAT scores to predict their GPA in a graduate program. We would use correlations to assess the strength of the association between the GMAT score with the criterion (i.e., GPA).

Convergent and Discriminant Validity. Campbell and Fiske's (1951) proposed to assess construct validity by examining their convergent and discriminant validity. The authors posited that construct validity can be best understood through two construct-validation processes: first, testing for convergence across different measures or manipulations of the same "thing", and second, testing for divergence between measures and manipulations of related but conceptually distinct "things" (Cook & Campbell, 1979, p. 61). In order to accumulate such evidence, Campbell and Fiske proposed the use of a multitrait-multimethod (MTMM) correlation matrix. This MTMM matrix allows one to zero in on the convergent and discriminant validity of a construct by investigating the intercorrelations of the matrix. The principle behind the MTMM matrix of measuring the same and differing behaviour is that it avoids the difficulty that high or low correlations may be due to their common method of measurement rather than convergent or discriminant validity. The table below shows how the MTMM matrix works.

	Method 1	Method 2	Method 3	Method 4	Method 5
Behaviours	12345	12345	12345	12345	12345

The matrix represents 5 different methods and 5 different behaviours. Convergent validity for Trait 1 is established if Trait 1

(T1) measured by Method 1 (M1) correlates highly with Trait 1 (T1) and Method 2 (M2), resulting in (T1M2) and so on for T1M3, T1M4, and T1M5. Discriminant validity for Trait 1 is established when there are no correlations among TI Ml and the other four traits (T2,T3,T4,T5) measured by all five methods (M2,M3,M4,M5).

A prevalent threat to construct validity is common method variance. Common method variance is defined as the overlap in variance between two variables ascribed to the type of measurement instrument used rather than due to a relationship between the underlying constructs (Avolio, Yammarino & Bass, 1991). Cook and Campbell (1979) used the terms mono-operation bias and mono-method bias, while Fiske (1982) adopted the term methods variance when discussing convergent and discriminant validation in research. Mono-operation bias represents the single operationalisation of a construct (behaviour) rather than gathering additional data from alternative measures of a construct.

External validity

If there is a causal relationship from construct X to construct Y, how generalisable is this relationship across persons, settings, and times? External validity of a study or relationship implies generalising to other persons, settings, and times. Generalising to well-explained target populations should be clearly differentiated from generalising across populations. Each is truly relevant to external validity: the former is critical in determining whether any research objectives which specified populations have been met, and the latter is crucial in determining which different populations have been affected by a treatment to assess how far one can generalise (Cook & Campbell, 1979).

For instance, if there is an interaction between an educational treatment and the social class of children, then we cannot infer that the same result holds across social classes. Thus, Cook and Campbell (1979) prefer generalising across *achieved* (my emphasis) populations, in which case threats to external validity

relate to statistical interaction effects. This implies that interactions of selection and treatment refer to the categories of persons to which a cause-effect relationship can be generalised. Interactions of setting and treatment refer to whether a causal relationship obtained in one setting can be generalised to another. For example, can the causal relationship observed in a manufacturing plant be replicated in a public institution, in a bureaucracy, or on a military base? This question could be addressed by varying settings and then analysing for a causal relationship within each setting.

Conclusion

This paper was written to provide the novice researcher with insight into two important concepts in research methodology: reliability and validity. Based upon recognised and classical works from the literature, the paper has clarified the meaning of reliability of measurement and the general problem of validity in behavioural research. The most frequently used techniques to assess reliability and validity were presented to highlight their conceptual relationships. Three important questions researchers frequently ask about what affects reliability of their measures and how to improve reliability were also discussed with examples from the literature. Four types of validity were introduced: statistical conclusion validity, internal validity, construct validity and external validity or generalisability. The approaches to substantiate the validity of measurements have also been presented with examples from the literature. A final discussion on common method variance has been provided to highlight this prevalent threat to validity in behavioural research. The paper was intended to provide an insight into these important concepts and to encourage students in the social sciences to continue studying to advance their understanding of research methodology.

References

Avolio, B. J. , Yammanno, F. J. and Bass, B. M. (1991). Identifying Common Methods Variance With Data Collected From A Single Source: An Unresolved Sticky Issue. *Journal of Management*, 17 (3), 571-587.

Bollen, K. A. (1989). *Structural Equations with Latent Variables (*pp. 179-225). John Wiley & Sons,

Brinberg, D. and McGrath, J. E. (1982). A Network of Validity Concepts Within the Research Process. In Brinberg, D. and Kidder, L. H., (Eds), *Forms of Validity in Research*, pp. 5-23.

Campbell, D.T. and Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Chapman, L.J. and Chapman, J.P. (1969). Illusory correlations as an obstacle to use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, 74, 271-280.

Cook, T. D. and Campbell, D. T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Boston: Houghton Muffin Company, pp. 37- 94.

Cortina, J. M. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, 78 (1), 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16(3)*, 297-334.

Dansereau, F., Alutto, J.A. and Yammarino, F.J. (1984). *Theory testing in organizational behavior: The variant approach.* Englewood cliffs, NJ: Prentice Hall.

Fiske, Donald W. (1982). Convergent--Discriminant Validation in Measurements and Research Strategies. In Brinberg, D. and Kidder, L. H., (Eds), *Forms of Validity in Research*, pp. 77-93.

Nunnally, J. C. (1978). *Psychometric Theory*. McGraw-Hill Book Company, pp. 86-113, 190-255.

Nunnaly, J. D. and Bernstein, I. H. (1994). *Psychometric Theory*. New York, NY: McGraw Hill.

Podsakoff, P M and Organ, D W (1986). Self-reports in organizational ‘ research: Problems and prospects. *Journal of Management*, 12: 531-544.

Miller, M. B. (1995). Coefficient Alpha: A Basic Introduction from the Perspective of Classical Test Theory. *Structural Equation Modeling*, 2 (3), 255-273.

Rosenthal, R. and Rosnow, R. L. (1991). *Essentials of Behavioral Research: Methods and Data Analysis.* Second Edition. McGraw-Hill Publishing Company, pp. 46-65.

Shadish, W. R., Cook, T.D., and Campbell, D. T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Smallbone, T. and Quinton, S. (2004). Increasing Business Students' Confidence in Questioning the Validity and Reliability of their Research. *Electronic Journal of Business Research Methods*, 2 (2): 153-162. www.ejbrm.com

Trochim, W. M. K. (2006). Introduction to Validity. *Social Research Methods*, retrieved from www.socialresearchmethods.net/kb/introval.php, September 9, 2010.

Williams, L.J., Cote, J.A. and Buckley, M.R. (1989). Lack of method variance in self-reported affect and perceptions at work: Reality or artifact? *Journal of Applied Psychology*, 74: 462-468.