Understanding the Psychometric Testing of a Measuring Instrument in Business Research

Dr. Sanjay Kumar
Dean-Academics, New Era College of Science and Technology,
Ghaziabad (U.P.) India.

ABSTRACT

The purpose of this paper was to understand the issues involved in the psychometric testing of a measuring instrument in business research. The two most important testing tools for measuring instrument are validity and reliability. Validity refers to whether a measuring instrument is measuring what it purports to. Three types of validity were discussed in this paper: (1) transaction validity, which assesses how well the measuring instrument samples the content domain being measured; (2) criterion validity, which assesses how well the measuring instrument correlates with other measures of the construct of interest; and (3) construct validity, which assesses how well the measuring instrument represents the construct of interest. How consistently does the measuring instrument measures whatever it does measure? This is the issue of reliability. Reliability is the degree to which the measuring instrument is dependable, consistent and replicable over time, over the instruments and other groups of respondents. Three types of reliability were discussed in this paper:(1) reliability as stability is a measure of consistency over time and over similar samples.(2) reliability as equivalence: If equivalent forms of a measuring instrument yield similar results, then the measuring instrument can be said to demonstrate this form of reliability and (3) reliability as internal consistency which examines the inter-item correlations within a measurement device and indicates how well the items fit together conceptually.

Key Words: Validity, Reliability, Criterion Validity, Construct Validity and Internal Consistency.

I INTRODUCTION

Quantitative research always depends on a measuring instrument. Two very important concepts that researcher must understand when they use measuring instrument are validity and reliability. Validity and reliability are jointly called the "psychometric properties" of a measuring instrument. Both are the yardsticks against which the adequacy and accuracy of our measurement procedures are evaluated in business research. A measure can be reliable but not valid, if it is measuring something very consistently but is consistently measuring the wrong construct. Likewise, a measure can be valid, but not reliable if it is measuring the right construct, but not doing so in a consistent manner. Hence, reliability and validity are both needed to assure adequate measurement of the constructs of interest. The purpose of this paper was to understand the issues involved in the psychometric testing of a measuring instrument in business research. The qualitative approach is used to describe and discuss the psychometric testing of a measuring instrument.

II VALIDITY

Validity is the most important consideration in developing and psychometric testing of a measuring instrument. Validity refers to whether a questionnaire is measuring what it purports to (Bryman & Cramer 1997). There are several different types of validity (Polgar & Thomas 1995, Bowling 1997). A validity can be tested using either theoretical or empirical approach. Both approaches are necessary for the validation of a measuring instrument. Theoretical testing of a validity focuses on how well the measuring construct is represented in an operational manner. This type of validity is termed as translational validity. This is also called representational validity. There are two popular methods (face validity and content validity) to evaluate the translational validity.

Face validity refers to whether measuring instrument seems to be a reasonable measure of its underlying construct "on its face". It is the easiest validation process to undertake, but it is the weakest form of validity. It evaluates the appearance of the questionnaire in terms of feasibility, readability, consistency of style and formatting, and the clarity of the language used (Haladyna 1999; Trochim 2001; Devon 2007).To determine the face validity of a measuring instrument, a face validity form is developed for respondents to assess each item in terms of the clarity of the wording; the likelihood the target audience would be able to answer the questions, the layout and style on a Likert scale of 1-4, strongly disagree= 1, disagree= 2, agree= 3, and strongly agree= 4. All respondents rate each item and items rated at three or four on a Likert scale of 1-4 are accepted as face validity. The feedback is taken on the items rated below three and modified as per need of the face validity.

Content validity refers to expert opinion concerning whether the scale items represent the proposed construct, the questionnaire is intended to measure. Content validity indicates the content reflects a complete range of the attributes under study and is usually undertaken by seven or more experts (Pilot & Hunger 1999). To estimate the content validity of a measuring instrument, the researchers clearly define the conceptual framework of the measuring construct by undertaking a thorough literature review and seeking expert opinion. Once the conceptual framework was established, a panel of seven or more purposely chosen experts in the relevant areas is employed to review the draft of the measuring instrument to ensure it is consistent with the conceptual framework. Each expert independently rated the relevance of each item on the measuring instrument to the conceptual framework using a Likert scale of 1-4 (1=not relevant, 2=somewhat relevant, 3=relevant, 4=very relevant). The Content Validity Index (CVI) is used to estimate the validity of the items (Lynn 1996). According to the CVI index, a rating of three or four indicates the content is valid and consistent with the conceptual framework (Lynn 1996). For instance, if five of eight content experts rate an item at three or four, the CVI would be 5/8=0.62, which does not meet the 0.87 (7/8) level required, and indicates the item should be dropped (Devon 2007). Theoretical approach of validity is an initial step in establishing validity, but is not sufficient by itself. Therefore, empirical approach of validity must also be demonstrated to develop a complete valid tool. The empirical approach of validity testing focuses on how a given measuring instrument is related to external criteria. This approach of validity testing is based on empirical data collected by a researcher. There are two types of validity **(criterion validity and construct validity)** under the empirical approach of validity. Criterion validity measures how well a measuring instrument predicts an outcome for another measuring instrument. It is useful for predicting performance in another situation. There are two popular methods to evaluate the criterion validity. These are concurrent validity and Predictive validity.

Concurrent validity is the relationship between scores on a newly developed test and previously developed test obtained at the same time. For instance, a researcher has developed an English language aptitude test and needs evidence that the test really measures English language aptitude. The researcher could select a well-known and previously validated English language aptitude test (criterion), administer it and the new English language aptitude test to a group of students, and determine the correlation between the two sets of scores. A substantial correlation between the new aptitude test and the widely accepted test is evidence that the new aptitude test is also measuring English language aptitude. The high correlation reflects high concurrent validity and low correlation reflects low concurrent validity.

Predictive validity is the relationship between scores on a newly developed test and scores on a criterion test available at a future time. For instance, a researcher has developed an English language aptitude test and needs evidence that the test really predict performance in English language courses. At the gathering predictive validity evidence of an English language aptitude test, one would look at the relationship between scores on the test and the scores students eventually earned in a future English language course (criterion). If a relationship is demonstrated, the scores on an aptitude test could be used later to predict performance in English language courses. In the case of a new scholastic aptitude test, predictive validity evidence would involve administering the test to a sample of high school students and then putting the scores away until the students complete their first semester of college. When the students' college scores become available, one would correlate the test scores and college scores. If the correlation is high, one has evidence for the usefulness of the aptitude test for predicting college achievement.

Criterion validity is a second step in establishing validity of a measuring instrument, but is also not sufficient by itself. Therefore, construct validity must also demonstrated to develop a complete valid tool. Construct validity relates to how well the items in the questionnaire represent the underlying conceptual structure. **Construct validity** refers to the degree to which the items of a measuring instrument relate to the relevant theoretical construct (Kane 2001; Devon 2007). Construct validity refers to the degree to which the items on meaning instrument relates to its theoretical construct. It is the degree to which a meaning instrument measures what it claims for the measurement purpose.

Campbell and Fiske (1959), Brock-Utne (1996) and Cooper and Schindler (2001) suggest that construct validity is addressed by convergent and discriminant techniques. Convergent and discriminant validity must also demonstrate by correlating the measure with related and/or dissimilar measures (Bowling 1997). Convergent techniques imply that different methods for researching the same construct should give a relatively high inter-correlation, while discriminant techniques suggest that using similar methods for researching different constructs should yield relatively low inter-correlations. Factor analysis is one of the best statistical technique for measuring the discriminant validity.

Factor analysis is a statistical technique which is very much used for the development of a measuring instrument in business research. This statistical technique clusters the items of a measuring instrument into common factors, interpret each factor of the measuring instrument to the items having a high loading on it and summaries the items into a small number of factors (Bryman & Cramer 1999). Loadings refers to the correlation between an item and a factor (Bryman & Cramer 2005). A factor is a list of items which belongs to the same group.Related items are grouped together under a factor, because they represent the construct and unrelated items that do not belong together, do not represent the construct and should be defected. (Munro 2005). In brief, factor analysis is that statistical method which clusters similar issues together and separates them from others.

III RELIABILITY

Reliability is the degree to which the measuring instrument is consistent or dependable. If we use this measuring instrument to measure the same construct multiple times, we do get pretty much the same result every time. Reliability refers to the degree to which a measuring instrument is consistent and dependable in measuring what it is intended to measure. This meaning of reliability is supported by Haladyna (1999) and Devon (2007). They define reliability, as consistency in the measurement of a questionnaire and how well the items fit together, conceptually. Validity is the primary necessity to test the reliability of a measuring instrument. If a test is not valid, then reliability is useless. Therefore, a measuring instrument may be reliable but not valid (Beanland et al. 1999; Pilot & Hunger 1999, Devon et al. 2007).

Reliability is the degree to which the measuring instrument is dependable, consistent and applicable over time, over the instruments and other groups of respondents. There are three principle types of reliability: stability, equivalence and internal consistency.

Reliability as stability is a measure of consistency over time and over similar samples. A reliable measuring instrument will yield similar data from similar respondents over a period of time. In the experimental research design this would mean that if a test and then retest are undertaken within an appropriate time span, then similar results would be obtained. This is a measure of temporal stability of the measuring instrument. This type of reliability is also called **test-retest reliability**. Test retest reliability can be measured by applying the same measurement instrument on the same sample at two different points of time on the assumption that there will be no change in the construct under study. (Trochim 2001; Devon, 2007). A high correlation between the scores at the two time points indicates the instrument is stable over time (Haladyna 1999; Devon et al. 2007).

The duration of time between the two tests is always debatable. The shorter the time interval, the higher the correlation between the two tests, the longer the time interval, the lower the correlation (Trochim, 2001). Generally, it is considered that a longer time gap may change the observation due to random error and it will provide lower test-retest reliability. In addition to stability over time, reliability can also be stabled over a similar sample. In the experimental research design this would mean that if we administer a test simultaneously to groups of students who are similar on significant characteristics, then similar results would be obtained.

Reliability as equivalence is measured in two ways. It may be achieved first through using equivalent forms or alternative forms of a measuring instrument. If equivalent forms of a measuring instrument yield similar results, then the measuring instrument can be said to demonstrate this form of reliability. This type of reliability might also be demonstrated if the equivalent forms of a measuring instrument yield similar results if applied simultaneously to similar samples. Here reliability can be measured through a t-test, through the demonstration of a high correlation coefficient and through the demonstration of similar means and standard deviations between two groups. Second, reliability as equivalence may be achieved through inter-rater reliability. Inter-rater reliability is a measure of how reliable the score is when different people rate the same performance on a measurement instrument. It gives a score of how much homogeneity there is in the ratings given by different people for the same performance on the same measuring instrument. Low inter-rater reliability is a sign of poor measuring instrument and high inter-rater reliability is a sign of good measuring instrument.

Reliability as Internal consistency is a measure of consistency between different items of the same construct. Internal consistency examines the inter-item correlations within a measuring instrument and indicates how well the items fit together conceptually (Nunnally & Bernstein 1994; Devon et al. 2007). Internal consistency is measured in two ways: Split-Half reliability and Cronbach's alpha correlation coefficient (Trochim 2001).

Split-half reliability is a measurement of consistency between two equal parts of a measuring instrument.The items of the measuring instrument can be divided into two equal parts on any logical basis.It is a type of reliability in which a measuring instrument is divided into two parts and the score of the same sample is computed on both the parts.Coefficient of correlation between the two scores is the measure of split-half reliability.It is one of the easiest way of establishing reliability of a measuring instrument.This reliability is directly proportional to the length of the measuring instrument i.e reliability increases with the length of the measuring instrument and vice-versa.

Cronbach's alpha, a reliability measure designed by Lee Cronbach in 1951, is the most common statistic to estimate reliability for internal consistency. This statistic uses inter-item correlations to determine whether constituent items are measuring the same domain (Bowling1997, Bryman & Cramer 1997, Jack & Clarke 1998). If the items show good internal consistency, Cronbach's alpha should exceed 0.70 for a developing questionnaire or 0.80 for a more established questionnaire (Bowling 1997, Bryman & Cramer 1997). The alpha is recommended $\geq$0.90 for measuring instruments used in clinical settings (Nunnally & Bernstein 1994) and alpha>0.70 is acceptable for a new measuring instrument (DeVellis 1991; Devon et al. 2007).

Cronbach's alpha is equivalent to the average of the all possible split-half estimates and is the most frequently used reliability statistic to establish internal consistency reliability (Trochim 2001; Devon et al. 2007). If an instrument contains two or more subscales, Cronbach's alpha should be computed for each subscale as well as the entire scale (Nunnally & Bernstein 1994; Devon et al. 2007). It is usual to report the Cronbach's alpha statistic for each subscale within a measuring instrument rather for the entire measuring instrument.

IV CONCLUSION

The paper discussed the procedures by which a reliable and valid measuring instrument can be developed. Validity and reliability are the two major psychometric characteristics of a measuring instrument. The researchers must understand the importance of validity and reliability when they use measuring instrument in any business research. If a piece of research is invalid and unreliable, then it is worthless. Validity and reliability is thus a necessary requirement of a measuring instrument. The paper is very useful for researchers who are interested in developing a valid and reliable measuring instrument in business research.

REFERENCES

[1] Bryman, A. & Cramer, D (1997). Quantitative Data Analysis with SPSS for Windows. Routledge, London.

[2] Polgar, S. & Thomas, S. (1995). Introduction to Research in the Health Sciences. Churchill Livingstone, Melbourne.

[3] Bowling, A. (1997). Research Methods in Health. Open University Press, Buckingham.

[4] Haladyna, T. (1999). Developing and Validating multiple-choice test items. New Jersey: Lawrence Erlbaum.

[5] Trochim, W.M.K. (2001). The Research Methods Knowledge Base. Cincinnati: Atomic Dog.

[6] Devon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J. & Lazzara, D. J.(2007). A psychometric Toolbox for testing Validity and Reliability. Journal of Nursing scholarship, 39 (2), 155-164.

[7] Pilot, D. & Hunger, B. (1999). Nursing research: principals and methods. Philadelphia: Lippincott Williams & Wilkins.

[8] Lynn, M.R. (1996). Determination and quantification of content validity. Nursing Research, 35, 382-385.

[9] Kane, M. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38, 319-342.

[10] Hunter, J.E. & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting errors and bias in research findings. Newsbury Park: Sage Publications.

[11] Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. Psychological Bulletin, 56, 81–105.

[12] Brock-Utne, B. (1996) Reliability and validity in qualitative research within education in Africa. International Review of Education, 42 (6), 605–21.

[13] Cooper, D. C. & Schindler, P. S. (2001) Business Research Methods (seventh edition). New York: McGraw-Hill.

[14] Bryman, A. & Cramer, D. (2005). Quantitative Data Analysis with SPSS12 and 13. A Guide for Social Scientists. East Sussex Routledge.

[15] Jack, B. & Clarke, A. (1998) The purpose and use of questionnaires in research. Professional Nurse 14, 176–179.

[16] Nunnally, J.C. & Bernstein, I.H. (1994). Psychometric theory. New York: McGraw-Hill.

[17] Robson, C. (2002) Real World Research (second edition). Oxford: Blackwell.