Validity and Reliability Issues in Educational Research

Oluwatayo, James Ayodele [PhD]

Institute of Education,
Ekiti State University, Ado-Ekiti, Nigeria
Email. ayotayor@yahoo.com

Doi:10.5901/jesr.2012.v2n2.391

Abstract: This paper discussed validity and reliability issues in educational research. It was suggested that validity and reliability can be applied to both quantitative and qualitative educational research and that threats to educational research can be attenuated by paying attention to validity and reliability throughout the research.

Keywords: Validity, reliability, types of validity and reliability, threats to validity and reliability

1. Introduction

Educational research has been defined by Ary, Jacobs and Razavieh (2002) as the application of the scientific approach to the study of educational problems. By scientific approach, educational research involves standards and procedures for demonstrating the empirical warrant of its findings, showing the match between its statements and what is happening or has happened in the world (Cuff & Payne, 1979).

Broadly, educational research can be divided into two categories namely, quantitative and qualitative research. The quantitative research uses objective measurement and statistical analysis of numeric data to understand and explain phenomena while qualitative research focuses on understanding and phenomena from the perspective of the human participants in the study (Cohan, Manion & Morrison, 2008). However, due to the multiplicity of measuring instruments available to researchers conducting either quantitative or qualitative research, the need to set criteria for the evaluation of such instruments is inevitable.

Technically, there are two most important criteria for measuring devices in research. They are validity and reliability. These two terms have application in educational research. Hence this paper discusses the concepts of validity and reliability, types of validity and reliability and threats to validity and reliability in research.

2. Concept of Validity

Historically, validity has been defined as the degree to which a test or measuring instrument actually measures what it purports to measure or how well a test or a meaning instrument fulfils its function (Anastasi & Urbina, 1997). However, recent views of validity seem not to be on the instrument itself but on the interpretation and measuring of the scores derived from the instruments. For example, Ary, Jacobs & Razavieh (2002) conceptualise validity as the extent to which theory and evidence support the proposed interpretation of test scores for an intended purpose. Relatedly, Whiston (2005) views validity as the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests. Similarly, Kaplan & Saccuzzo (2005) view validity as the evidence for inferences made about a test score. Further, McBurney & White (2007) view validity as an indication of accuracy in terms of the extent to which a research conclusion corresponds with reality. The foregoing suggests that validity hinges on the extent to which meaningful and appropriate inferences or decisions are made on the basis of scores derived from the instrument used in a research.

3. Types of Validity

The literature on psychometric validation is saturated with the different types of validity that are used in research (e.g. Campbel & Fiske, 1959; Andrews, 1984; Cronbach, 1989; Sireci, 1998; Cohen, Manion & Morrison, 2008). Indeed, Cohen, Manion & Morrison (2008) listed several kinds of validity including content validity, criterion-related validity, construct validity, internal validity, external validity, concurrent validity, face validity, jury validity, predictive validity, consequential validity, systemic validity, ecological validity, cultural validity, descriptive validity, interpretive validity, theoretical validity and evaluative validity. However, in educational research, there are four types of validity that are of much importance. They are face, content, construct and criterion-related validity.

4. Face Validity

Face validity refers to researchers' subjective assessments of the presentation and relevance of the measuring instrument as to whether the items in the instrument appear to be relevant, reasonable, unambiguous and clear. Several authors have commented on the status of face validity in research. Most of these authors believe that face validity is not truly an indicator of validity and hence should not be considered as one (e.g. Kaplan & Saccuzzo, 2005; Whiston, 2005). The argument is that face validity does not offer evidence to support conclusions drawn from the test scores and that an instrument may look valid without measuring what it is intended to measure. However, Anastasi & Urbina (2007) argue that face validity is a desirable feature of tests, noting that if tests originally designed for children and developed with a classroom setting are extended for adult's use, such tests are likely to meet with resistance and criticism because of their lack of face validity. In other word, if the content of a measuring instrument appears irrelevant, inappropriate, silly or childish, there is every likelihood that the result obtained might provide false information and misleading decisions.

Practically, the quantitative assessment of face validity can be achieved by having experts in the field of study or psychometrically unsophisticated interested persons to rate the suitability of the measuring instrument for its intended use. The criteria for assessment may include:

i. the structure of the instrument in terms of construction and well-thought out format
ii. the clarity and unambiguity of items
iii. appropriateness of difficulty level for the respondents
iv. correct spelling of difficult words
v. spacing of items between lines
vi. adequacy of instruction on the instrument
vii. reasonableness of items in relation to the perceived purpose of the instrument
viii. legibility of printout
ix. attractiveness of paper used
x. other criteria that satisfy face validity

It may sound illogical to conclude that face validity is irrelevant in research. Indeed, it is an aspect of validity that researchers in both quantitative and qualitative should pay attention to when reporting their validity procedure. It is not enough to just report that the researcher's supervisor and experts in Tests and Measurement judge the instrument as satisfying face validity, their comments on the criteria earlier mentioned should be presented to make the research meaningful and robust.

5. Content Validity

Content validity is a theoretical concept which focuses on the extent to which the instrument of measurement

shows evidence of fairly and comprehensive coverage of the domain of items that it purports to cover. According to Babbie (2007), content validity shows the degree to which a measure covers the range of meanings included within a concept. In another dimension, Cohen, Manion & Morrison (2008) define content validity as a form of validity that ensures that the elements of the main issue to be covered in a research are both a fair representation of the wider issue under investigation and that the elements chosen for the research sample are addressed in depth and breadth. Essentially, careful sampling of items is the basic requirement to ensure their representatives.

In looking for content validity evidence in educational research, the focus is on determination of whether a measuring instrument has been constructed adequately or its items have a fair sample of the total potential content. For example, if a test is chosen as a measuring instrument, the establishment of content validity evidence requires good logic, intuitive skills and perseverance. In other words, the content of the items must be carefully evaluated. Meanwhile, Messick (1986) suggests that test developers must consider the wording of the items and the appropriateness of the reading level as part of content validity.

In literature, the determination of content validity evidence is often made by experts' judgment. Indeed, there are several methods for aggregating judgment into an index of content representation. These include multiple judges rating, statistical methods and test specification method. In multiple judges rating, the researcher consults the experts in the field of the research to rate each item in the instrument in terms of its match or relevance to the content (Rubio, BergWeger, Tebb, Lee & Rauch, 2003). The implication is that every correction or modification made by the experts must be effected and reported in the final presentation of the research.

In the statistical method, the most frequently used method is factor analysis. Factor analysis is used to determine whether items in the instrument fit into conceptual domain (Sireci, 1998).

Test specification method provides the organisational framework for the development of the instrument (Whiston, 2005). A common first step in test specification is to identify the goals or the content areas to be covered. This is followed by defining the behavioural change, affective or cognitive changes that the researcher intends to measure. The match between the content areas and behavioural domain is helpful in enabling the researcher to clearly articulate the intended purpose(s) of the instrument. In the final procedure of test specifications, the researcher must take into consideration the level of the respondents whether the instrument is designed for children or adults, the level of complexity of the items, and the appropriate reading level (Kane, 2001).

6. Construct Validity

Construct validity is a device commonly used in educational research. It is based on the logical relationships among variables. According to Walden (2012), construct validity refers to whether the operational definition of a variable actually reflects the theoretical meanings of a concept. In other words, construct validity shows the degree to which inferences are legitimately made from the operationalisations in one's study to the theoretical constructs on which those operationalisations are based.

Practically, construct validity comprises two elements namely, convergent validity and discriminant validity. The convergent validity requires that the scores derived from the measuring instrument correlate with the scores derived from similar variables (Campbell & Fiske, 1959; Brock-Utme, 1996; Cooper & Schindler, 2001). For example, if a researcher is interested in assessing a child's intelligence, he or she may construe intelligence to be ability to demonstrate a high result in an intelligence test. In order to establish construct validity, the researcher needs to be sure that his or her construction of a particular issue agrees with other constructions of the same underlying issue such as creativity, anxiety, motivation and so on. A high correlation coefficient implies construct validity of the new instrument or research.

In another dimension, convergent validity can be achieved by correlating scores obtained between the scale and sub-scales together. The inter-correlations from the multitrait-multimethod matrix are used to

support convergence validity (Campbell & Fiske, 1959). The principle behind this technique of validation is that different methods of measuring the same construct should yield similar results.

The discriminant validity suggests that using similar methods for researching different constructs should yield relatively low inter-correlations. That is, the construct in question is different from other potentially similar constructs. Such discriminant validity can also be yielded by factor analysis, which clusters together similar issues and separates them from others (Cohen, Manion & Morrison, 2008).

7. Criterion-Related Validity

A criterion is a standard of judgment or an established standard against which other measure is compared (Kaplan & Saccuzzo, 2005). Therefore, criterion-related validity covers correlations of the measure with another criterion measure, which is accepted as valid (Bowling, 2009). In other words, criterion-related validity is where a high correlation coefficient exists between the scores on a measuring instrument and the scores on other existing instrument which is accepted as valid.

In literature (e.g. Wolf, 1994; Whiston, 2005; Cohen, Manion & Morrison, 2008; Bowling, 2009), there exists two types of criterion-related validity: concurrent validity and predictive validity. There is a time lag between when the instrument is administered and the time when the criterion information is gathered.

In concurrent validity, the measures and criterion measures are taken at the same time because they are usually designed to provide diagnostic information that can help guide educational development of the learners (Kaplan & Saccuzzo, 2007). On the other hand, predictive validity is achieved if the data acquired at the first round of research correlates highly with data acquired at a future date. For example, if the results of examination taken by a student at his entry into the university (say, Unified Tertiary Matriculation Examination scores) correlate highly with the examination results at his or her graduation level, then it might be concluded that the first examination demonstrated strong predictive validity (Oluwatayo, 2004).

The criterion-related validity evidence can be achieved by three broad methods: correlational method, regression method and decision theory or group separation method. The correlational method uses the statistics technique of correlations to determine the magnitude and direction of relationship between the measure (independent variable) and criterion measure. The first step is selection of an appropriate group to use in the validation study, followed by the administration of the instrument. If the focus is on concurrent validity, the next step is to collect the criterion data and correlate the scores on the instrument with the criterion information (scores); the result of the calculation being a validity coefficient. By squaring the validity coefficient, a percentage of variance in the criterion that is accounted for by the instrument is obtained, called a coefficient of determination, which provides an indication of the amount of shared variance between the two variables. However, if the focus of the research is on predictive validity, this involves a time lapse where the researcher would wait until it is appropriate to gather the criterion information and then correlate. In most discussions of validity coefficients, a question arises about the magnitude of the coefficient and how large a validity coefficient should be. Kaplan & Saccuzzo (2005) indicate that validity coefficients are rarely larger than 0.60 and that they often fall in the 0.30 and 0.40 range while Anastasi & Urbina (1997) suggest that validity coefficients should at least be statistically significant, which means that the coefficient has a low probability of occurring by chance.

The regression method is based on the premise that a straight line, called the regression line, can describe the relationship between the instrument's scores and the criterion scores. The scores on an instrument are plotted in relation to the scores on the criterion. The line that best fits the points is the regression line. Once a regression line is established, it becomes possible to predict performance on the criterion based on scores on the instrument.

The decision theory or group separation or expectancy table concerns whether the scores of the instrument correctly differentiate, that is, do testees who score high on the test get the high grades, as predicted? Decision theory attempts to assist in the selection and placement by taking available information

and putting it into a mathematical form (Cronbach & Glesser, 1965). In this way, it is possible to determine whether the instrument accurately predicts those testees who would succeed in college or university versus accurately predict those who will not succeed.

8. Concept of Reliability

Reliability is one of the most desirable technical merits in any educational research though its meaning differs in quantitative and qualitative research. Quantitative research assures the possibility of replication. That is, within a certain limit of experimental error or random error, if the same methods are used with the same sample, then the results should be the same (Cohen, Manion & Morrison, 2008). In a more explicit way, Bowling (2009) views reliability in quantitative research as synonymous to dependability, consistency, reproducibility or replicability over time, over instruments and over groups of respondents. Indeed, for a research to be reliable, it must demonstrate that if it were to be carried out on a similar group of respondents in a similar context, similar results would be obtained.

On the other hand, qualitative research strives to record the multiple interpretations of intention in and meanings given to situations and events (Brock-Utme, 1996). Consequently, reliability in qualitative research is regarded as a fit between what researchers record as data and what actually occurs in the natural setting that is being researched. Meanwhile, Bogdan & Bilken (1992) have earlier argued that qualitative research is not to strive for uniformity but accuracy and comprehensiveness of courage, noting that two researchers who are studying a single setting may come up with very different findings but both sets of findings being reliable. Interestingly, Winter (2000), Stenbacka (2001) and Golafshani (2003) suggest that reliability in qualitative research should be replaced with terms such as credibility, neutrality, confirmability, dependability, consistency, applicability, trustworthiness and transferability. In addition, LeCompete & Preissle (1993) suggest that the canons of reliability for quantitative research may be simply unworkable for qualitative research because typical quantitative research methods require a degree of control and manipulation of phenomena whereas in qualitative research, the control and manipulation of variables may distort the natural occurrence of phenomena. Nevertheless, the focus in the application of reliability in educational research is to determine whether a particular technique applied repeatedly to the same object would yield the same result each time.

9. Types of Reliability in Educational Research

There are three principal types of reliability in educational research: stability, equivalence and internal consistency.

Reliability as stability/reproducibility/repeatability/replicability: This is a measure of consistency over time and over similar samples (Cohen, Manion & Morrison, 2008). Expectedly, a reliable instrument for a piece of research should produce similar data from similar respondents over time. For example, in the experimental and survey models of research, this would mean that if a test and then a retest were carried out within an appropriate time span, then similar results would be obtained. The correlation coefficient using test-retest method is usually estimated using Pearson statistics or t-test. Whiston (2005) suggests that if the correlation coefficient using test-retest were 0.80 (80% correct observation and 20% error) or higher, then the reliability can be guaranteed. Similarly, Cohen, Manion & Morrison (2008) suggest that the statistical significance of the correlation coefficient must be 0.05 or higher if reliability is to be guaranteed. However, Bland & Altman (1986) believe that correlations are a weak measure of test-retest reliability and recommend the use of confidence intervals to assess the size of the difference between the scores. Rationally, both statistical techniques can be used in educational research to justify the relevance of test-retest method in research.

Nevertheless, it is significant to note that stability over a sample is particularly useful in piloting tests and questionnaires. Consequently, Cooper & Schindler (2001) suggest that in using test-retest method, the following must be ensured:

- the time period between the test and retest must not be so long that situational factors change;
- the time period between the test and retest must not be so short that the participants would remember the first test;
- the participants may have become interested in the field and may have followed it up themselves between the test and retest times.

Reliability as equivalence

Reliability as equivalence is of two sorts: alternate or parallel form and inter-rater form. Estimating reliability using alternate or parallel form requires developing two forms of an instrument using the same content-domain, the same test specifications, the same number of items, the same items format and similar difficulty and discriminating indices. In this method, the respondents are given one form of the instrument initially and then assessed with a second alternate or parallel form of the instrument. The scores derived from the two instruments are then correlated to estimate the reliability coefficient. In another way, the two sets of instruments can simultaneously be administered on two homogenous groups and their scores compared using either Pearson statistics or t-test statistics. A correlation coefficient of at least 0.80 is accepted as producing comparable responses (Bowling, 2009).

In inter-rater reliability, the focus is on the extent to which the results obtained by two or more raters agree for similar or the same ratees. This method is pertinent to a team of researchers gathering structural observational or semi-structural interview data where each member of the team would have to agree on which data would be entered in which categories. For observational data, reliability is addressed in the training sessions for research-assistants where they are intimated with the material to ensure parity in how they enter the data. The simplest level of calculating inter-rater agreement is using percentage (Bowling, 2009). This is expressed by

Reliability as internal consistency

Reliability as internal consistency tests for the homogeneity of items in a measuring instrument. Bowling (2009) defines it as the extent to which the items relating to a particular dimension in an instrument tap only this dimension and no other. The internal consistency demands that the instrument or test be administered once on the intended group of respondents and their scores collated for analysis using the appropriate statistical tools.

In educational research, much attention is placed on internal consistency using split-half, item-total correlations, Kuder-Richardson-20 & 21 and Cronbach alpha (Cronbach, 1951).

Split-half

Split-half reliability assumes that the items in an instrument can be split into two matched halves in terms of contents and cumulative degree of difficulty. This is often achieved by assigning all the odd numbered items to one group and all even numbered items into another. Essentially, a testees's marks on one half is expected to match his or her marks on the other half. The calculation follows by correlating the marks in the odd items with the marks in the even items using Pearson's statistics and corrected for the whole items using Spearman-Brown formula:

$$\text{Reliability} = \frac{2r}{1+r} \quad \text{where } r = \text{the actual correlation between the halves of the instrumen}$$

Bryman & Crammer (1990) suggest the reliability level is acceptable at 0.80 and above.

Item-Total Correlations

The items-total correlations refer to the extent to which the score of each item in an instrument correlates with the total score of all the items in the instrument. For example, if a researcher has 50 items on his or her instrument, he or she is expected to present 50 correlation coefficients in the final analysis. The usual rule of thumb for item-total correlations is that items should correlate with the total scale score by more than 0.20 to satisfy reliability and scaling assumptions (Streiner & Norman, 2003). Statistically, if the items in the instrument have dichotomous responses (e.g. Yes/No, Agree/ Disagree, True/False and so on), the Point-biserial correlation is usually recommended. However, if items have more than two responses (in continuum) such as Strongly Agree, Agree, Disagree, and Strongly Disagree, the Product Moment Correlation is usually used (Kline, 1986). Practically, item-total correlations are inflated in scales with additional items so that the redundant items are deleted.

Kuder-Richardson-20 & 21 (KR-20&21)

Kuder & Richardson (1937) develop procedures for determining homogeneity of items. Probably, the best known index of homogeneity is KR-20; which is based on the proportion of correct and incorrect responses to each of the items on a test. The formula for KR-20 is:

$$r_{20} = \frac{k}{k-1}\left[\frac{S_x^2 - \sum pq}{S_x^2}\right]$$

where: r_{20} = reliability of the whole test
k = number of items in the test
S^2_x = variance of scores on the total test
p = proportion of correct responses on a single item
q = proportion of incorrect responses on the same

The product *pq* is computed for each item, and the products are summed over all items to give $\sum pq$. KR-20 is applicable to tests whose items are dichotomous (True/False, Right/Wrong). However, KR-21 assumes that all items in the test are of equal difficulty and computationally simpler. The formula for KR-21 is:

$$r_{21} = \frac{k}{k-1}\left[1 - \frac{M(K-M)}{KS^2}\right]$$

where: r_{21} = reliability coefficient of the whole test
k = number of items in the test
S^2 = variance of scores
M = Mean of the scores

Ary, Jacobs & Razavieh (2002) remark that KR-21 method is by far the least time consuming of all the reliability estimation procedures because it involves only one administration of the test and employs only easily available information. Consequently, the authors recommend it to teachers for classroom use.

Cronbach-Alpha or Coefficient Alpha

Cronbach (1951) developed a formula for estimating the internal consistency of an instrument in which the items are not scored dichotomously (e.g. Yes/No, Right/Wrong, True/False, Agree/Disagree, etc). The formula, widely known as Cronbach-alpha or Coefficient alpha is given by:

$$\alpha = \left[\frac{N}{N-1}\right]\left[\frac{S_x^2 - \sum S_i^2}{S_x^2}\right]$$

where:

N = number of items on the instrument

S_i^2 = variance of individual item score

S_x^2 = sum of variances of scores of individual items

S_x^2 = variance of the total test scores

Cronbach-alpha is widely used in educational research when instrument for gathering data have items that are scored on a range of values, such as essay tests in which different items have different scoring points or attitude scales in which the item responses are in continuum (e.g. Strongly Agree=4, Agree=3, Disagree=2, Strongly Disagree=1) because it takes into consideration the variance of each item. Indeed, Whiston (2005) points out that when the scoring of items is not dichotomous, then the appropriate method of estimating reliability is Cronbach-alpha. Moreover, Ary, Jacobs & Razavich (2002) emphasise that if the test items on an instrument are heterogenous, that is, measuring more than one trait or attribute, the reliability index is best computed using Cronbach-alpha. However, there arises a question as to what should be the acceptable magnitude of correlation when using Cronbach-alpha. For example, Nunnally (1994) suggests 0.70 (70% of variance reliable) and above while some authors suggest that it is acceptable if it is 0.67 (Cohen et al, 2008). However, Whiston (2005) notes that Cronbach-alphas are usually low and conservative estimates of reliability.

Threats to Validity and Reliability in Educational Research

There are many threats to validity and reliability in educational research including biases and errors in the conceptualisation of the research, the research design, sampling and process of the study.

Conceptual bias: This arises from the faulty logic of the researcher, leading to faulty conceptualisation of the research problem, faulty interpretations and conclusions.

Design bias: This emanates from the studies which have faulty design, methods, sampling procedures and the use of inappropriate techniques of analysis. The resultant effect is that there exists a difference between the observed value and the true value.

Sampling bias: This occurs when the researcher's sample does not represent the population of interest. It may also occur when there exists differences in the selection of samples for the comparison groups in a research or when intact classes are employed as experimental or control groups.

Process bias: This is the sum of all errors from the sampling method to data collection and analysis. It is a known fact that when invalid or unreliable instruments are employed to generate data, such data would lead to statistical regression and distort the results of the research.

However, threats to validity and reliability in educational research can be attenuated if the researcher clearly defines his or her research problem, uses the appropriate research design, selects representative and

unbiased sample, uses valid and reliable instrument for data collection, employs the appropriate statistical tools for analysis and avoids *Type I* and *Type II* errors in interpreting the results.

Summary and Conclusion

This paper discussed validity and reliability issues in educational research, touching the concepts of validity and reliability, types of validity and reliability and threats to validity and reliability in educational research. Comparatively, validity and reliability are related because concepts such as concurrent validity, predictive validity, convergent validity and discriminant validity are based on evidence from the reliability coefficients. However, it needs to be pointed out that a measure may have high reliability without supporting evidence for its validity. Notwithstanding, it could be concluded in this paper that validity and reliability are touchstones of all types of educational research and hence suggests that threats to validity and reliability in educational research can be attenuated if attention is paid to the validity and reliability throughout the research.

References

Anastasi, A. & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River. NJ: Prentice-Hall.

Anastasi, A. & Urbina, S. (2007). Psychological testing (2nd impression). Pearson. NJ: Prentice-Hall.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modelling approach. *Public Opinion Quarterly*. 48: 409—422.

Ary, D.; Jacobs, L. C. & Razavich, A. (2002). Introduction to research in education (6th ed.). Wadsworth Thomson Leaving. Chapter 9: 241—274.

Babbie, E. (2007). The practice of social research (11th ed.). Belmont, USA: Thomson high Education.

Bland, J. M. & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1: 307—310.

Bogdan, R. G. & Bilken, S. K. (1992). Qualitative research for education (2nd ed.) Boston, MA: Allyn & Bacon. 48.

Bowling, A. (2009). Research methods in health: Investigative Health and Health Services (3rd ed.). New York: McGraw-Hill. 162—176.

Brock-Utme, B. (1996). Reliability and validity in qualitative research within education in Africa. *International Review of Education*. 42(6): 605—621.

Bryman, A. & Cramer, D. (1990). Quantitative data analysis for social scientists. London: Routledge. 71.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-traits multi-method matrix. *Psychological Bulletin*. 56: 81—105.

Cohen, L.; Manion, L. & Morrison, K. (2008). Research methods in education (6th ed.). London & New York: Routledge Taylor & Francis Group. 133—164.

Cooper, D. C. & Schindler, P. S. (2001). Business research methods (7th ed.). New York: McGraw-Hill.

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

Cronbach, L. J. & Glesser, G. C. (1965). Psychological tests and personnel decisions. Urbana: University of Illinois Press.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*. 52: 281—302.

Cronbach, L. J. (1989). Construct validity after thirty years. In R. Linn (ed.) *Intelligence measurement, theory and public policy*. Urbana: University of Illinois Press.

Cuff, E. G. & Payne, G. C. F. (1979). Perspectives in sociology. London: Allen & Unwin.

Golafshani, N. (2003). Understanding reliability and validity in qualitative assessment. *The Quantitative report*. 8(4): 597—607.

Kane, M. T. (2001). Current concept in validity theory. *Journal of Educational Measurement*. 38(4): 319—342.

Kaplan, R. M. & Saccuzzo, D. P. (2005). Psychological testing: Principles, applications and issues (6th ed.) Thomson Wadsworth. 132—154.

Kline, P. (1986). A handbook of psychological testing. London: Methuen.

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of reliability. *Psychometrika*. 2: 151—160.

LeCompete, M. & Preissle, J. (1993). Ethnography and qualitative design in educational research (2nd ed.). London: Academic Press.

McBurney, D. H. & White, T. L. (2007). Research methods (7th ed.). Thomson Wadsworth. 169.

Messick, S. J. (1998[b]). Test validity: A matter of consequence. *Social Indicators Research*. 45(1-3): 35—44.

Nunnally, J. (1994). Psychometric theory. 3rd ed. New York: MCGraw-Hill.

Oluwatayo, J. A. (2004). Mode of entry and performance of university undergraduates in science courses. An unpublished Ph.D. thesis, University of Ado-Ekiti, Nigeria.

Rubio, D. M.; Berg-Weger, M.; Tebri, S. S.; Lee, E. S. & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*. 27(2): 94—104.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*. 45(1—3): 83—117.

Stenbacka, C. (2001). Qualitative research requires quality concepts of its own. *Management Decision*. 39(7): 551—555.

Streiner, G. I. & Norman, D. R. (2003). Health measurement scales; A guide to their development and use. 3rd ed. Oxford: Oxford University Press.

Walden, U. (2012). Educational social psychology. www.experiment-reserch. com, www.alleydog.com 1998—2012.

Whiston, S. C. (2005). Principles and applications of assessment in counselling. 2nd ed. Thomson Brooks/Cole. 43—74.

Winter, G. (2000). A comparative discussion of the notion of validity in qualitative and quantitative research. *The Quantitative Report*. 4(3-4). www.nova.edu/SSS/QR/QR4-3/winter.internal.

Wolf, R. M. (1994). The validity and reliability of outcome measure. In A. C. Tuïjuman & T. N. Postlethwaile (eds.). *Monitoring the standards of education*. Oxford: Pergamon. 121—132.