Reliability and Validity: A Sine Qua Non for Fair Assessment of Undergraduate Technical and Vocational Education Projects in Nigerian Universities

John A.C., Ph. D
Professor ,Department of Technology Education, Modibbo Adama University of Technology, P.M.B. 2076, Yola

Abstract

The aim of the study was to examine the importance of reliability and validity as necessary foundation for fair assessment. The concepts of reliability, validity, fair assessment and their relationships were analysed. Qualities of fair assessment were discussed. a number of recommendations were made to make assessors be more cautious in award of marks or grades.

Introduction

The validity of a test according to Gay (1981) referred to the extent to which it measures what it was supposed to measure. Alonge (1985) defined validity as what the test measures, how well it does so and what can be inferred from it. Cronbach (1988) considered validity as the accuracy of a specific prediction or inference made from a test score. As described in the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA] and National Council on Measurement in Education [NCME], 1999) validity referred to the "degree to which evidence and theory support the interpretations of test scores" (p.9). The process of accumulating evidence to support the validity of test score interpretations starts prior to the development of an assessment (AERA, APA & NCME, 1999).

Gronlund (1981) defined validity as the extent to which the result of an evaluation procedure serves the particular purpose for which the results are intended to measure and nothing else.

> Thus, validity evidence is often defined in terms of (a) correlations between tests measuring the same construct or between a test and the criterion behaviour of interest. (b) tables of specification to determine whether the content of a test measures the breadth of content targeted and (c) using a range of strategies to build a logical case for the relationship between scores from the assessment and the construct the assessment is intended to measure. (p.76)

In the classroom context, validity is seen as multi-dimensional construct that resides, not in tests but in a relationship between any assessments, the context to measure and consequences of its interpretation and use (Obe, 1983). This means that validity encompasses (a) how assessments draw out learning, (b) how assessment fits with the educational context and instructional strategies used, and (c) what occurs as a result of assessments including the full range of outcomes from feedback, grading and placement, to students' self concepts and behaviours, to students' constructions about the subject disciplines (Messick, 1989).

Cronbach (1988), referred to validity as the ability of a test to do the job it is expected to do. A test may have high validity for one purpose but have low or moderate validity for another. This means that any assessment no matter how well designed is valid for some purposes and invalid for others.

> Stanley and Hopkins (1982) identified three main types of validity. The one that pertains to (a) future performance on some criterion are known as criterion – related validity (b) the extent of knowledge of universe of curricular content and processes is referred to as content validity (c) the degree to which certain psychological traits or constructs are actually represented by test performance (construct validity). Alonge (1985) posited validity to be a measure of how well it fulfilled the functions for which it was capable of achieving certain aims, while Messick (1989) explained it in terms of how well a performance on an assessment correlates with performance on another instrument. They concluded that validity is the extent to which the assessment method measures what is supposed to measure. Messick (1999) defined validity as the extent to which a claim about a student based on assessment data from that student was justified. The extent to which an assessment therefore, measures what it is intended to measure and permits appropriate generalizations about students' activities is referred to as validity, (Decaesteker, Coetsier & Lievens, 2000).

Establishing validity from the point of psychometric principles entails making the warrant explicit, examining the network of beliefs and theories on which it relies and testing its strength and credibility through various sources of backing. It requires determining conditions that weaken the warrant, exploring alternative explanation for good or poor performance and feeding them back into the system to reduce inferential errors (Mislery, Wilson, Ercikan & Chudowsky, 2001).

Validity is the extent to which how well an assessment fulfils the functions for which it is being used. An assessment, therefore, lacks validity for a particular task if the information it provides is of no value (Linn, 1986). In other words, an assessment may be valid for one specific purpose, a particular group, subject matter and instructional objective and not for others. For instance, an assessment that is valid for Technical Drawing achievement is invalid for personality test. This is because an assessment of Technical Drawing achievement does not measure any aspect of personality. When the concept of validity was applied to teaching, Hartman (2001) argued that a valid teacher was the one who taught what should be taught and whose students learnt what they were supposed to learn. A teacher therefore, might be valid for a particular purpose, of an area of instruction and a specific level (Mazzo, 2001). If the results are intended to assess students' achievement validity expects that it should be seen to be so and nothing else. That is, a valid secondary school technical drawing teacher for instance, may not serve as a building construction expert. Therefore, one should decide on what the assessment should measure before it can be done accurately.

Validity in this study pertains to the results of two independent assessments and not the instrument itself. In this context, it is erroneous to think that assessment is valid or invalid, rather, it is a matter of degree such that there can be high, moderate and low validities (Nicholls & Nohen, 1993; Rowe & Hill, 1996). Since the individual who draws inference from an assessment must determine the extent to which it serves its purpose, the establishment of validity invariably requires judgment or appraisal to determine its value, worth and effectiveness (AERA, 1998; Gillis, Griffin, Trembath, & Liang, 1999).

Ezewu and Okoye (1982) described validity as the extent to which a test or other measuring instruments fulfil the purpose for which it is used, or by a study of the relationship between test scores and other variables. Abdullahi (1983) stated that validity of an assessment as valid if it turned out to be a suitable measure. Education pundits (Yoloye, 1983 and Cronbach, 1988) agreed that the single most important consideration in evaluating an assessment was the degree of validity.

The steps involved in determining validity and reliability, according to Rowe & Hill (1996:3) are as follows:

1. Administer the new test to a defined group of individuals;
2. Administer a previously established valid test (or acquire such scores if already available) to the same group at the same time or shortly thereafter.
3. Correlate the two sets of scores
4. Evaluate the results.

In this study, the defined group is the undergraduate Vocational and Technical Education students in the Universities under study. The students' research projects, are scored by a team of assessors after the work has been criticized and necessary corrections effected. Thereafter, the same group of students are assessed by external examiners. The two independent scores are correlated to establish if there is any relationship between them so that certain crucial decisions can be taken based on the results.

The technical education teacher may want to measure the skills acquired by the students through a written test instead of a practical project. This is because a good practical test is often difficult and more expensive to design and set while a written test has high validity, it means students who perform well in the written test would have high practical skills and may do well on the job (Alonge, 1985). Similarly, some subjects examined by West African Examination Council (WAEC) for Senior Secondary School Certificate (SSCE) have alternative to practical. It is assumed that a student who performs well in the alternative to practical paper will also do well in the actual practical work and if that happens, a reliability and validity is established. This according to Okoro (1994) explained why West African Examination Council (WAEC) designed the alternative to practical papers in Science and Technical subjects for external candidates.

According to Satterley (2000), the validity of an assessment is the extent to which it actually measures what it is stated to measure. Accuracy of measurement resulting from an assessment and how likely it is that the same result will be produced in slightly different circumstance is what IBO (2003a) referred to as reliability. An assessment is reliable if a student will gain the same result when a test is repeated on different occasion, and assessment marked by different markers. Validity and reliability are essential characteristics of any assessment system especially where the outcome is of great importance to the student and the teacher. Both characteristics are in fact multi-modal.

Since assessments are designed for a variety of purposes, it is not surprising that there are different types of validity. Lievens (2002) identified different types of validity as Construct, Translation (face and content validity) and Criterion – related (predictive, concurrent, convergent and discriminant validity). He described validity from the point of operationalisation. That is, translating a concept into a functioning and operating reality.

Construct validity is defined as the appropriate truth of the conclusion that the operationalisation accurately reflects its construct. It is defined by Gronlund (1981) as the extent to which test performance can be interpreted in terms of certain psychological construct. A construct is a psychological trait which is assumed to exist to explain some aspects of human behaviour which include creativity and validation, the nature and strength of all factors that can influence such a performance test.

Translation validity is relatively a new concept given to both face and content validity. The focus is to find out whether the operationalisation is a good reflection of the construct. Face validity even though the weakest, focuses on operationalisation to find out "on its face" if the test is a good translation of the construct. The quality of face validity assessment can be improved considerably by making it more systematic. Lievens (2002) explained content validity as checking the operationalisation against the relevant content domain for the construct. That is, how well content covered in the course is reflected in the test items. This should include domain specification and the definition of the target group. This validity addresses the basic question: Is the measure of a test representative of the content of the property being measured?

Criterion – related validity is also known as empirical validity because it employs empirical techniques in studying the relationship between scores on a test and some outside criteria. Criterion related validity aims at prediction and generalization from one score to another. Predictive Validity is the operationalisation's ability to predict something using some valued measures of what is assessed in future activity. In other words, it is concerned with the usefulness of a test in prophesying some future performance. Prediction is an attempt to forecast an outcome on the basis of data or information considered relevant to the observed event. A high correlation, therefore, would provide evidence of predictive validity and vice-versa. Concurrent validity, according to Denga (1987), is the degree to which scores on a test are related to the scores on another, already established test administered at the same time or to some other valid criterion available. When a test is capable of doing the same job as some other tests, easier or faster the concurrent validity is established and in most cases the new test will be utilized instead of the other tests. According to Allen (1988) concurrent validity is determined by establishing a relationship or discrimination. The method involves determining the relationship between scores on the test and scores on some other established criterion. Okoro (1994) defined concurrent validity as the extent to which performance in one activity can be used to predict performance in another activity taking place not in the future but in the present.

Convergent validity is the opposite of discriminant validity. The former according to Lievens (2001)

> Examines the degree to which the operationalisation is similar to (converges on) other operationalisation that it theoretically should be similar to while the latter examines the degree to which operationalisation is not similar to (diverges from) other operationalisation's that it theoretically should not be similar to (p.5)

For instance, convergent validity of a test in research methods skills can be shown by correlating the scores on the test with scores in project writing, where high correlations will be evidence of convergent validity. In discriminant validity, the test scores on Technical Drawing skills may be correlated with the scores of research methods, where low correlations will be evidence of discriminant validity.

Since validity is a mark of truthfulness and usefulness there is need for reinforcement of students and public confidence in the grades awarded by assessors. Reliability is a sine – qua – non but does not guarantee validity. Generally, regardless of all other merits, if a test lacks validity for a particular task, the information it provides is useless. From this discussion, it is clear that there is a need to establish the reliability and validity of internal and external assessment of scores in Vocational and Technical Education Projects in Nigerian Universities.

Basically, reliability is the degree to which an assessment consistently measures whatever it measures. The more reliable an assessment is, the more useful the scores obtained from the administration of such assessment in decision making process. An unreliable assessment therefore, is essentially useless (Gay, 1981)

Reliability is usually expressed numerically as a coefficient. A high coefficient indicates high reliability with minimal error while a low correlation coefficient indicates low reliability with maximal error. A coefficient of 1.00 shows that a student's score is perfectly reflected in both internal and external assessment. However, no assessment is perfectly reliable. For instance, an assessment may yield consistent scores i.e. reliable but may not be the 'true' score i.e. invalidity (Wikimedia Foundation, 2006). In other words, if an assessment measures what is supposed to measure, it is reliable if it does so every time, but may consistently measure the wrong thing and become invalid. This illustrates an interesting relationship between validity and reliability.

Intimately related to validity is the concept of reliability. That is, how consistent test scores or results exist in form of measurement and another. Test scores provide a limited measure of behaviour obtained at a particular time, and unless the measurement can be shown to be reasonably consistent over different occasions or different samples of the same behaviour, little confidence can be placed in the results. Gronlund (1981) succinctly stated that reliability provides the consistency which makes validity possible. He further observed that although a highly reliable measure might have little or no validity, a measure which has been shown to have satisfactory degree of validity must of necessity possess sufficient reliability.

Reliability according to Cronbach (1970) is the accuracy or precision with which a measure based on one sample of test tasks at one point in time represents performance based on a different sample of the same kind of task or at different points in time or both. Accuracy may be expressed by a reliability coefficient or by the measure if it is reliable to the extent that an individual remains nearly the same in repeated scores as represented by a low standard error of measurement or by a high reliability coefficient (Cronbach, 1988). Okedara (1980) and Okoro

(1994) viewed reliability as the degree with which a test consistently measures what it was intended to measure. For the two education pundits, reliability is the degree of consistency between two measures of the same thing. According to Bello (1998), this consistency meant how effective one could generalize a measurement over different occasions and over different samples of the same behaviour.

According to Ohuche and Akeju in Chukwuemeka (1989), reliability is the accuracy, trustworthiness or consistency of a measuring instrument. Similarly, Jan (2001) perceived it as repeatability, stability, reproducibility and dependability. They opined that reliability is the degree of consistency or stability of the measure obtained from the instrument. If individuals tend to maintain the same order of merit on each of two administrations of a test, the instrument is said to be reliable since it is consistent standard error of measurement.

Stanley and Hopkins (1982) were of the view that repeated sets of scores of a series of objects or individuals would ordinarily show some degree of consistency. This tendency towards consistency from one set of scores to another is known as reliability. The degree of reliability of a set of scores is a very important consideration in educational evaluation, both in practical day to day use of tests and in empirical research. A score for an individual on a test is obtained to make some judgment about an individual and usually to take practical action based on the result. The implication of this is that an educator must always give priority attention to this instrument in selecting a test to see whether they are both reliable and valid for the decision and characterization that he proposes to make.

There are a number of ways in which reliability can be expressed or used as a precision of a set of scores or instrument. According to Oshkosh (2005), these are:

(i) Inter-Assessor or Inter-Observer Reliability
(ii) Test-Retest Reliability
(iii) Parallel Forms Reliability
(iv) Internal Consistency Reliability.

Inter-rater reliability according to Ayodele (1989) was used to assess the degree to which different assessors or observers give consistent estimates of the same phenomenon. Since human beings are notorious for inconsistencies, how can the reliability in the observation of two observers be determined? It means an inter-rater reliability outside the context of the study must be established. There are two major ways to actually estimate inter-assessor reliability (Abiri, 2006). If the rating consists of categories, the raters check by calculating the category each observation falls in the percentage of agreement between the raters. For instance, if 86 out of, 100 were checked in the same category by assessors, it means the percentage of agreement would be 86%. The other major way to estimate inter-rater reliability is appropriate when the measure is a continuous one is to calculate the correlation between the ratings of the two observers. For instance, there might be rating of the overall level of activity in a classroom on a 1-to-7 scale. Their ratings are conducted at regular intervals (e.g. every 30 seconds). The correlation between these ratings would give an estimate of the reliability.

Test-retest reliability is estimated when the same test is administered to the same sample on two different occasions (Salim, 2001). This approach assumes that there is no substantial change in the construct being measured between the two occasions. The amount of time allowed between measures is critical. It is obvious that whenever measures of the same thing are taken twice, the correlation between the two observations will depend in part by how much time elapses between the two assessments. The shorter the time gap, the higher the correlation, the longer the time gap, the lower the correlation. This is because the two observations are related over time. That is, the closer the time, the more similar the factors that contribute to error. Since correlation is the test-retest estimate of reliability, different estimates depending on the interval can be obtained (Abiri, 2006).

In parallel forms reliability, two similar instruments are created. One way to accomplish this is to create a large set of questions that address the same construct and randomly divide the questions into two sets. The two instruments are administered to the same sample. The correlations between the two parallel forms are the estimate of reliability. One major problem of this approach is that lots of items that reflect the same concept must be generated which is not an easy feat to achieve. This approach, according to Yoloye (1983), assumed that randomly divided halves are parallel or equivalent which might not necessarily be the case. The major difference is that parallel forms are constructed so that the two forms can be used independent of each other and considered equivalent measures.

Cronbach (1970) reported that consistency reliability estimate can be done when a single instrument is administered to a group of respondents on one occasion to estimate reliability. In effect, the reliability of the instrument is judged by estimating how well the items that reflect the same construct yield similar results. That is looking at how consistent the results are for different items of the same construct. There are different consistency measures that can be used. They are: - average inter-item correlation, average item total correlation, split-half and Cronbach's Alpha. In split-half reliability all the items that purpose to measure the same construct are divided into two sets. The entire instrument was administered to a sample of respondents and the total score for each randomly divided half was calculated. The split-half reliability estimate is simply the correlation between these two total scores (Williams and Bateman, 1996). Imagine the split-half reliability is computed and the items were randomly divided into another set of split halves and recomputed, the Cronbach's Alpha is mathematically equivalent to the average of all possible split-half estimates (American Education Research Association [AERA], 1998).

Each of the reliability estimators has certain advantages and disadvantages. Inter-rater reliability is one of the best ways to estimate reliability when there is one observation (Tuckman, 1985). However, since it requires multiple raters or observers, the correlation of ratings of the same single observer repeated on two different occasions, the test-retest approach can be used where there is a single rater and don't want to train any others. On the other hand, in some studies it is reasonable to do both to help establish the reliability of the raters or observers.

The parallel forms estimator is typically used in situations where you intend to use the two forms as alternate measures of the same thing. Both the parallel forms and all of the internal consistency estimators have one major constraint which is multiple items designed to measure the same construct. This is relatively easy to achieve in certain contexts like achievement testing. Cronbach's Alpha tends to be the most frequently used estimate of internal consistency.

The test-retest estimator is especially feasible in most experimental and quasi-experimental designs that use a no-treatment control group. In this design a control group that is measured on two occasions (pre – test and post – test) is established. The main problem with this approach is the absence of information about reliability until the post-test is collected and if the reliability estimate is low, it is not acceptable.

Each of the reliability estimators will give a different value for reliability. In general, the test-retest and inter-rater reliability estimates will be lower in value than the parallel forms and internal consistency ones because they involve measuring at different times or with different raters. Since reliability estimates are often used in statistical analyses of quasi-experimental designs, the fact that different estimates can differ considerably makes the analysis even more complex (Linn, 1986).

Although reliability of internal and external assessment scores of undergraduate vocational and technical education research projects forms the focus of this research, absolute precision in award of marks by assessors on every task undertaken by a student is not possible. It is for this reason that International Baccalaureate Organisation (2004) recommended at least 95% confidence that every grade is "correct" and a correct result in this sense is the one that would be confirmed by subsequent remarking of candidates work by experts.

Relationship between Reliability, Validity and Fair Assessment

Reliability and validity are important qualities of any fair assessment system. If a measuring instrument measures what is purports to measure, such instrument can be said to be valid because it has turned out to be a suitable measure. For instance, the students' research projects are scored by both internal and external examiners. When the two sets of scores are correlated and a relationship is established it can be said that the assessment is reliable and valid. It is reliable because the instrument used in the process of assessment gives consistent information about the abilities being measured and valid since it has measured what is designed to measure.

Reliability is a sine-qua-non but does not guarantee validity. The more reliable an assessment is, the higher the degree of process. An unreliable assessment is not valid and therefore may consistently measure the thing. In that case, it is reliable but invalid. An fair assessment cannot be said to be credible, dependable, truthful without being reliable and valid. An assessment is fair if it is reasonable, logical and acceptable by all stakeholders within an educational system.

Qualities of Fair Assessment

Assessment is derived from the Latin word *assidere* which means to sit beside. In educational context, the process of observing, describing, collecting, recording, scoring and interpreting information about a students learning is known as assessment.

Salvia and Ysseldyke (1985:12) described assessment as a process of collecting data for two purposes: (a) specification and verification of problems (b) making decisions for or about students. Dietel, Herman and Knuth (1991) conceptualized assessment as a method used to better understand what the current knowledge that a student possesses. This implies that an assessment can be a teacher's subjective judgment based on the observation of a student's performance. The current knowledge implies that what a student knows is always changing and that judgment about student's achievement through comparison over a period of time which may reflect decisions about grades, advancement, placement, instructional needs and curriculum are bound to change. Griffin, Gillis, Keating and Fennessy (2000) viewed assessment from it's characteristics which provided accurate estimates of students performance that enables teachers and other decision makers to make appropriate decisions.

The first vital quality of fair assessment is its validity which captures these essential characteristics and the extent that an assessment actually measures what it is intended to measure and permits appropriate generalisations about students' skills and abilities. If the assessment is valid, it is safe to generalise that a student will likely do as well on similar questions not included in the assessment. The results of a good assessment, in short, represents something beyond how students perform on a certain task or on a particular set of question; they represent how a student performs on the objective on which those tasks were intended to assesses.

A second important quality of fair assessment information is its consistency or reliability. Will assessment results for a student be similar if they are gathered at some other time or other different circumstances or if they are scored by different assessors? For example if a student's age is asked for, on three separate occasions and in three different locations and the answer is the same each time then, that information is considered reliable. In the context of vocational and technical education assessment, inter – rater reliability is essential because it requires that independent raters give the same score to a given students response (Allen, Matters, Dudley and Gordon, 1992). When viewed from monitoring, review and evaluation of programme, Hager, Athanason and Gonczi (1994) explained assessment as the process of identifying, understanding the problem and planning a series of action to deal with it. They further stressed that there are usually a number of different stages in this process, but the end result is always to have a clear and realistic plan of activities designed to achieve a set of clear aims and objectives.

The third essential quality of a fair assessment is practicality. That is, the instrument can easily be used by all because the process of using it and arriving at evidence is obvious. Closely linked with this is the ease with which the performance can be evaluated. Issues tied to practicality are the time it will take to complete the assessment process, how expensive the materials or equipment needed to carry out the task and whether it can be administered to a group of students or not.

Gillis and Bateman (1999) considered assessment as the purposeful process of gathering appropriate and sufficient evidence of competence and the interpretation of that evidence to enable the outcomes to be communicated to key stakeholders. The International Baccalaureate Organisation (2004) explained the term to cover all the various methods by which a student achievement can be evaluated with the assessment instruments such as tests, examination, extended practical work, projects and oral work carried out over a long period and sometimes marked by the student's teacher.

Finally, a fair assessment according to Alonge (1989) is a measure of what you are speaking about expressed in numbers. He further argued that when a particular material cannot be measured, the knowledge concerning the material is of meagre and unsatisfactory kind. Salvia (as cited in Medugu 2002) described fair assessment as a process which starts with decision making because different decisions require different types of information. According to IBO (2004), assessment should be able to accommodate the element of curriculum which is individual subject and the curriculum as a whole. The organisation went further to categorise that assessments into formative and summative. Summative assessment is concerned with the final judgment of performance while formative is concerned with the improvement of performance. In broad terms, marking and grading involve summative assessment while reviewing and giving feedback involve formative assessment.

Assessment does not operate in a vacuum. Since one of the important methods of verifying efficiency of an educational process and outcomes is fair assessment, it must be seen to be valid. Assessment scores may have perfect consistency but not correlated with any other variable and therefore, will be of no practical validity value. Alonge (1989) conceptualized assessment this way:

> When you measure what you are speaking about and express it in numbers, you know something about it; but when you cannot measure it and express it in numbers, your knowledge is of meagre and unsatisfactory kind and therefore cannot be said to be valid (p.100)

Whenever students' project works are assessed and scores assigned to them based on certain established parameters, it gives an idea about the scholastic standing of the individual. However, such assessments have significant difficulties such as unreliability, invalidity and unauthenticated. Assessment places an assessor in a difficult position both as a support and a judge of a student learning. This may be compounded by the strong element of subjectivity because of the relationship between the teacher and the students.

Recommendations

The following recommendations were made based on the findings of the study and their implications.

1. Since assessors have different training backgrounds there is need for every assessor to be given special training and be restricted to marking guidelines to reduce inconsistencies and enhance reliability. Universities' mark schemes should be detailed enough to reduce ambiguity and increase the level of reliability.
2. Sufficient information should be provided to help students identify areas of weakness and strengths so as to correct and improve on their assessments.
3. Reading through each section of every chapter of vocational and technical education research project and scoring should be encouraged rather than reading through and awarding scores at the end of every chapter or the entire work.
4. Since assessment criteria differ from one university to another external assessor should pay more attention to the guidelines of the institution they are moderating to enhance reliability.

5. Since education is dynamic, assessment criteria should be evaluated from time to time for efficacy and relevance.

Conclusion

A fair assessment must serve as motivators, mechanisms for review, influence cognitive processing and ensure feedback to promote the overall learning process.

References

Abdullahi. A. (1983). A study of correlation between examination scores in selected school subjects. *Journal of Nigerian Education Research Association* 3(1), 29-34

Abiri, J.O.O. (2006). *Elements of evaluation measurement and statistical Techniques in education*. Ilorin: Library Publications Committee, University of Ilorin.

Allen, J.R. (1988). Internal and external assessment scores: A focus on gender differences. *Retrieved on May 4, 2005 from on line.net.*

Allen, J.R.; Matters, G.N.; Dudley, R.P.; & Gordon, P.R. (1992). *A report on the assessment of the Queensland Senior Curriculum to identify the common elements.* Brisbane: Queensland Board of Senior Secondary School Studies

Alonge, M.F. (1985). The place of continuous assessment in guidance and counseling service. *Journal of School of Pure Science*, 1(3), 160-167.

Alonge, M.F (1989) *Measurement and evaluation in education and psychology,* Ado-Ekiti: Adebayo Printing Nigeria Ltd.

American Education Research Association (1998). *Standards for educational and Psychological Testing*. Washington, D.C: Author.

Ayodele, S.O. (1985). *Assessment practices and the english teacher*. In J.D. Medugu An investigation of the assessment practices of industry technology teachers in Adamawa State. An unpublished M.Ed. Thesis. Abubakar Tafawa Balewa University, Bauchi.

Bello, B.O. (1998). Continuous assessment in primary schools: A Crucial Measurement index of the 6-3-3-4 education policy in Nigeria, *Nigerian Journal of Curriculum Studies Special edition*, 122-125.

Chukwuemeka, D. C. (1989). *Validity and reliability of teacher made biology that questions used in Niger state secondary schools*. An Unpublished M.Ed Thesis, University of Nigeria Nsukka.

Cronbach, L.J. (1970). *Essentials of psychological testing*. New York: Harper and Row. In J.O.O. Abiri (author). Elements of evaluation measurement and statistical techniques in education, Ilorin: Library and Publication committee.

Cronbach, L.J. (1988). Five perspectives on validity argument. Hillsdale, N.J: *File://D:\Education Policy Analysis Archieves Vol.4 No.17 Taylor & No.htm.*

Decaesteker, C. Coetsier, P. & Leivens, F. (2000). *Validity*, *acceptability and fairness of tests used in student selection.* A paper presented at the International Congress of Psychology. Stockholm, Sweden. 11th April, 2000.

Denga, D.I. (1987). *Criterion-referenced measurement, continuous assessment psychological testing*. Calabar: Rapid Education Publishers Ltd.

Ezewu, E. & Okoye, N.W. (1982). *Principles of Assessment Ibadan*: Evans Brothers Ltd.

Gay, L.R. (1981). *Educational research competencies for analysis and application,* Torants: Charles E. Merril Publishing Company.

Gillis, S., & Bateman, A. (1999). *Assessing in vocational education training: Issues of reliability and validity, a reviw of research*, Adelaide: National Centre for Vocational Education Research

Gillis, S., Griffin, P., Trembath, R., & Liang, P. (1996). *The examination of the theoretical underpinning of assessment. A research report,* Melbourne: Australian National Training Authority.

Gronlund N.E. (1981). *Measurement and evaluation in teaching*. New York: Macmillan Publishing Co. Inc:

Griffin, P; Gills, S; Keeling, J & Fennessy, D. (2000). Assessment and senior secondary school certificates: Interim report, Australian national training authority funded research project, Sydney: Research Centre for vocational education and training.

Hager, P., Athanason, J., & Gonczi, A. (1994). Assessment technical manual, In M. Williams & Bateman A. (Ed) *Graded assessment in vocational education and training: An analysis of national practice, drivers and areas of policy development.* Australia: National Centre for Vocational Education Research *w.w.w.never.edu.au.*

Hartman, R.S. (2001). Validity studies of the hartman profile model file://O:/ *Validitystudies of the Hartman profile model.htm*

International Baccalaureate Organisation (2004). *Diploma programme assessment, principles and practice.* Cardiff: International Baccalaureate Organisation [IBO]

Lievens, F. (2001). Assessors and use of assessment center dimensions: A fresh look at a troubling issue. *Journal of Organizational Behaviour* 22, 203-221.

Lievens F. (2002). Factors which improved the construct validity of assessment centers: *A review of International Journal of Selection and Assessment* 6, 141-152

Linn, R.L. (1986). Educating validity assessments: The consequences of use. *Educational measurement and evaluation Issues and Practice* 16(2), 14-16

Mazzo, C. (2001). Frameworks of State: Assessment policy in historical perspective. *Teachers College Record,* 103(3), 367-397.

Messick, S. (1989). Validity. In R. Linn (Ed.) educational measurement, *website:* www.ibo.org.

Mislery, R.J., Wilson, M.R; Ercikan, K; & Chudowsky, N. (2001). Psychometric Principles in student assessment. Doodrecht, the Netherlands: *Kluwer Academic Press.*

Nicholls, F. & Nohen, S.B., (1993). Uses and abuses of achievement test scores. *Educational Measurement Issues and Practices*, 11, 9-15.

Obe, E.O. (1983). *A survey of the attitudes of some lagos state secondary school teachers attitude toward continuous assessmen*t. In J.D. Medugu An investigation of the assessment practices of introductory technology teachers in adamawa state. An unpublished M.Ed. Thesis. Abubakar Tafawa Balewa University, Bauchi.

Okedara, C.A. (1980). Teacher made tests as predictors of academic achievement. *African Journal of Educational research* 3(1) 15-16

Okoro, O.M. (1994). *Measurement and evaluation in education*. Uruowulu-Obosi: Pacific Publishers.

Rowe, K.J. & Hill, P.W. (1996). Assessing, recording and reporting students' Educational progress: *The case for 'subject profiles'* file://E:\ Assessment in education.htm

Salim, B.A. (2001). *Assessment bodies and development of modern techniques In assessment and examination.* A paper presented at the 16th congress of the Nigeria Academy of Education held at University of Jos, Jos between 12th – 16th November,

Salvia, J., & Ysseldyke, E.J. (1985) *Assessment in special and remedial education.* Boston*:* Honghton Miftlin Co.

Satterley, D. (2000). The quality of external assessment. In W. Harlen (Ed.) Enhancing quality in assessment. file://D:\quality.of.external assessments.htm

Stanley, J.C., & Hopkins, K.D. (1982). *Educational and psychological measurement and evaluation.* Boston: Allyn and Bacon.

Tuckman, B.W. (1985). *Measuring Educational Outcomes.* Hondras: Brace Jovanorich International.

Wikimedia Foundation (2006). Qualification and assessment. *Retrieved on August, 25, 2006 from* http://en.wikibooks.org.//wiki/SA_NCS:qualification_assessment

Williams, M., & Bateman, A. (1996). Graded assessment in vocational education and training: An analysis of national practice, drivers and areas for policy development. *Retrieved on August, 25, 2005 from* www.never.ed.au.

Yoloye, E.A. (1983). The validity of longe thorndike intelligence tests for achievement in Nigeria grammar schools. *The West African Journal of Educational and Vocational Measurement* (1) 48-58.