Improving Validity of Tests through Improved Test Development Procedures

by

Anyanwu, I. E., Ph. D

&

Onwuakpa, F. I. Williams, Ph. D
Quality Assurance Department,
National Examinations Council (NECO)
Minna, Niger State, Nigeria

Being a Paper Presented at the Annual International Association for Educational Assessment (IAEA) Conference Held at the University of Kansas, Lawrence, Kansas, USA from October 11 to 15, 2015

August, 2015

TITLE: Improving Validity of Tests through Improved Test Development Procedures

by

Anyanwu, I. E., ph. D

&

Onwuakpa, F. I. Williams, Ph. D
Quality Assurance Department,
National Examinations Council (NECO)
Minna, Niger State, Nigeria

ABSTRACT

Psychometricians over the world are of the view that validity and reliability of test items are very critical in Quality Assurance of test items. However, validity is the most important between the two attributes of a good test because it gives the true scores, relevance and appropriateness of test items.

This paper presents the definition of validity of test items, its types and importance in quality assurance of testing procedures.

Effort was made to identify some test development procedures in order to improve upon validity measures of tests among which are providing clear instructions in the tests, avoiding the use of difficult vocabularies, appropriate arrangement of items and improving upon the length of tests.

The paper opines that when these test development procedures are adopted, then, there is an assurance in the improvement of the validity of the test.

1.0 INTRODUCTION

Schools and other educational centres are established for the purposes of enabling learners (students or pupils) acquire certain desirable learning experiences and competencies. These learning experiences and competencies include knowledge, abilities and skills. This can be acquired as a result of teaching and learning activities, going on in the classroom through the interactions of the learners with teachers and materials respectively.

It is expected of every teacher to assess the extent to which these learners have acquired the desirable learning experiences and skills over a given period of time. This is done in order to know how much the learners have learnt as well as how well the teacher has taught. The assessment of the teachers' teaching and learners' amount of learning can be done through the use of class work, assignment, quiz and projects.

However, every assessment instrument such as tests must possess certain critical attributes or characteristics which must include validity, reliability and usability [Nwana (1997), Okpala et al (1993), Nenty (1997)]. Among these three major attributes of tests, the validity component has been regarded as the most special and important of all (Nwana 1997). This is because validity of a test is a measure of the test's appropriateness, meaningfulness and usefulness of any inference made from its scores. Also, assessment is very important as an integral part of every teaching-learning process by providing results and at the same time addressing their learning difficulties. Against this premise, it is very necessary that test developed by teachers to assess the learners' learning in the classroom as well as their performance at the end of the term must be valid and reliable. It is expected that tests and other assessment procedures used in public examinations ought to be valid and reliable as well. In the light of the above, there is the need to identify certain best practices in test development that will help in improving the validity of tests used by our teachers in the school system as well as in public examination bodies. This formed the basis of this paper which seeks to identify ways of improving the validity of test instruments used in educational assessment activities.

1.1 Definition of Validity of a Test

There are several definitions of validity in the measurement procedure. Validity refers to the appropriateness, meaningfulness, and usefulness of any inference made from a test score [American Psychological Association (APA), American Educational Research Association (AERA), and National Council on Measurement in Education (NCME), 1985]. Shimbery (1990) defines it as the level of confidence with which an examinee's test score could be used to infer the ability under measurement possessed by the examinee. A more common definition of validity says that is the extent to which a test measures what it was designed to measure appropriately without contamination from other characteristics. For example, a test of Reading Comprehension should not require Mathematical ability.

1.2 Relationship Between Validity And Reliability of A Test

There is a very important relationship between validity and reliability of a test. A test with very low reliability index will also have a low validity index. Clearly, every measurement with very poor accuracy or consistency (reliability) is very unlikely to be fit for its purpose.

But, by the same token, the things required to achieve a very high degree of reliability can impact negatively on validity. For example, consistency in assessment conditions leads to greater reliability because it reduces 'noise' (variability) in the results. On the other hand, one of the things that can improve validity is flexibility in assessment tasks and conditions. Such flexibility allows assessment to be set appropriate to the learning context and to be made relevant to particular groups of students. Insisting on highly consistent assessment conditions to attain high reliability will result in little flexibility and might therefore limit validity.

1.3 Types of Validity of A Test

Okpala et al (1993) identified four types of validity: face validity, content validity, construct validity and criterion-related validity.

1.3.1 *Face Validity*

The face validity of a test assesses whether the items appear to be appropriate, reasonable, challenging and feasible. It has no statistical index as to the degree of the appropriateness of the items. It's purely a qualitative assessment of whether the items look good or bad to the eyes of a tester or other user. This could be achieved by giving the test items to an expert(s) or someone who is experienced in teaching and perhaps testing. He or She looks at the items one by one to know whether the questions at face value can be attempted by the testees for which it was designed in terms of difficulty or easiness of the terms, time to attempt the questions and relevance to the level of the testees and topic(s) being taught or learnt by the testees.

Some authorities in educational assessment have argued that there is nothing like face validity in a technical sense of validity. They argued that just because a test has face validity does not mean that it will be valid in the technical sense of the word. The argument is that since no statistics are involved, the name "face validity" or "on the face of it" has no empirical justification. In the light of this argument, Anikweze (2009) suggested that a better strategy is to secure rational or logical validity based on the quantified consensus of experts considering the test or instrument in terms of appropriateness for the objectives it is expected to measure.

1.3.2 Content Validity

Every testing instrument has two major components: the objectives of instruction (in terms of knowledge, comprehension, application, analysis, synthesis and evaluation) and the topics of instruction. In this regard, a test has content validity if it measures knowledge of the content domain of which it was designed to measure. In another sense of it, content validity concerns primarily on the adequacy with which the test items comprehensively and representatively sample the content areas to be measured. For instance, good achievement test in Mathematics for the Basic Educational

Certificate Examination (BECE) in Nigeria should cover items on Number and Numerations, Algebra, Mensuration, Trigonometry Statistics and Probability appropriately. No area(s) of these content domains should be left behind. It must also cover these content areas across the behavioral objectives as determined by the tester.

Experts judgment (with no statistics) is the primary method used to determine content validity. The expert judgment is based on properly prepared test blue print i.e operational chart called table of specification showing the distribution of the test items by behavioural objectives and by content areas (Bloom et al, 1956). The table of specification ensures that all aspects of the syllabus or curriculum are adequately represented in the body of test. The table of specification provides functional content validity for the test.

Another strategy for determining the content validity of a test is to ensure that the test has a high correlation with other tests that purport to sample the same content domain.

1.3.3 ***Criterion-Related Validity***

A test is always designed to measure some behaviours of testees on present or future performance. The present or future behavior of the testees (i.e candidates or students) is described as an independent measure called Criteria. For instance, a child's National Common Entrance Examination score in Nigeria (for admission into secondary school) could be used to measure his or her performance at the end of the Basic or Senior Secondary education (future performance). In another sense, the same National Common Entrance Examination score could be used to estimate the testees or pupils present level of intelligence (present performance).

However, Criterion-related validity is studied by comparing test or scale scores with one or more external variables or criteria known or believed to measure the attribute under study. For example, when one predicts success or failure of students/pupils from academic aptitude measures, one is

concerned with criterion-related validity. In fact, in criterion-related validation, the basic interest is usually more in the criterion and some practical outcomes than in the predicators. (Kerlinger, 1973).

In a sense, all tests are predictive i.e they predict a certain kind of outcome, some present or future state of affairs. For example, Aptitude tests predict future achievement; Achievement tests predict present and future achievement and competence; and intelligence tests predict present and future ability to learn and to solve problems. (Thorndike, R and Hagen, E, 1969). The single greatest difficulty of criterion-related validation is the criterion. Obtaining a criterion may even be very difficult. For instance, what criterion can be used to validate a measure of teacher effectiveness?
Astin (1964) maintains that every criterion must have the following desirable qualities such as relevance, freedom from bias, reliability and availability. Against this background, Criterion-Related Validity is the extent to which scores obtained from a test or any evaluation instrument are in agreement with Current Criterion (concurrent validity) or predict future criterion measures (predictive validity). The basic distinction between concurrent and predictive validities is the time interval when the criterion data are gathered (Okpala et al, 1993).

1.3.3.1 ***Concurrent Validity:*** Concurrent Validity of a test is that which measures or determines the testees present level of performance and typical behavior of the testee. (ie students or learner) (Okpala et al, 1993). For instance, a test which measures whether the testee has attained the minimum prescribed level of competence at the end of instruction will be subjected to concurrent validity. It is interested in measuring the

current level of performance and not the future level of performance.

1.3.3.2 ***Predictive Validity:*** Predictive Validity refers to the accuracy with which the results of a test for instance, test scores of an aptitude test forecast future behavioural change in students. For instance, one would use the scores of candidates in the National Common Entrance Examination (NCEE) taken in 2015 to forecast the candidates' grades in their Basic Education Certificate Examination (BECE) in 2018 (after 3 years). In doing this, we correlate the Mathematics scores of the candidates in 2015 NCEE with their scores or grade in 2018 BECE Mathematics using Pearson-Product Moment Correlation Coefficient or Spearman (Rho) rank-order correlation coefficient.

1.3.4 Construct Validity

A construct or trait is a psychological attribute which underlines explanations of a universe of behaviour such as intelligence, creativity, aptitude, perception, reasoning ability, study habits, scientific attitude etc. Hence, construct validity may be defined as the extent to which a test measures a specific trait or psychological construct. For instance, if a test was designed to measure students' study skills, the tester will administer the test to a sample of students, collect and analyse their responses. In this process, the tester will attempt to find out whether the test measures attitudes to school work, verbal reasoning, attitude toward teacher and peers or whether it actually measures study skills.

However, construct validity is complex and difficult to determine as some of the constructs are not measured directly but through indirect means. Okpala et al (1993) identified

seven methods of obtaining evidence of construct validity which are as follows:

- Experimental Interventions;
- analysis of mental processes required by the test item;
- Correlation with other instruments;
- factor analysis;
- internal consistency;
- multi-trait multimethod matrix method;
- appeal to logic.

In estimating construct validity of a test, it is very necessary that we should precisely define what the construct is about before embarking on developing the test. Specific definition of the construct implies a thorough explanation of its meaning. For instance, developing a test that measures creativity of Junior Secondary School students requires the tester to define creativity in words or by means of a list of activities or behaviours that can be demonstrated by a creative child. The description should be clear and not to be confused with other related constructed such as intelligence.

2.0 STRATEGIES FOR IMPROVING TEST VALIDITY THROUGH IMPROVED TEST DEVELOPMENT PROCEDURES

Validity of any test is a very important characteristic that is a must for any test be it for public or school examination. Caution must be exercised in the test development procedures in order to ensure validity and fairness in every assessment activity.

However, the following guidelines for Best Test Development Procedures or Practices must be adhered to if we want achieve a good level of validity in our testing activities:

2.1 Thorough Planning and Development of a Test

The planning and development of a test starts with identifying the purpose of the test. The purposes of the test must be clearly specified in order for valid interpretations to be made on the basis of the scores

from the test. Such purposes like promotion, diagnosis of learning/teaching, admission, and selection must be stated clearly. The purpose helps the test developer to identify the skills, abilities and knowledge to be tested. It also helps to determine the level of the skills, abilities and knowledge to be covered in the test appropriately. For instance, a Mathematics test designed to promote students from Junior Secondary Class Three (JS 3) to Senior Secondary Class One (SS 1) must develop items at all components of JS 3 Mathematics (Geometry, Algebra, Statistics/Probability, Mensuration, Trigonometry and Number & Numeration) at reasonable proportions. This must also be linked to the cognitive levels (knowledge, comprehension, application etc) also at reasonable proportion using a carefully prepared and standardized Table of Specification (ToS).

Within the planning and development framework, it is also essential to develop a precise and explicit definition of the construct the test is intended to measure. The underlying theoretical rationale for the existence of the construct should be well articulated (John W. Young et al 2013). A test that is built on a strong theoretical foundation is one that is more likely to lead to valid interpretation of the test scores. In addition, a clear definition of the construct in a test being measured can help to clarify the skills associated with that construct. This enables test developers to create tasks for a test that will best engage the testees skills and reflect the construct of interest. For instance, a test in Mathematics designed to measure the testees' level of intelligence at JS 3 levels will elicit their skills in Algebraic expression, logic, qualitative aptitude, number and numeration at high level of difficulty (eg between 0.1 to 0.3 difficulty levels)

2.2 Provision of Clear Instructions and Use of Unambiguous Items

Test developers must ensure that clearly stated instructions that will guide the testees are provided for the testees. This is because instructions give directions as to the number of items to attempt and the response mode. If instructions are vague, the testees may be confused resulting to a low validity. Moreso, ambiguous statements

in test items may lead to different interpretations thereby reducing t validity (Okpala et al, 1993).

2.3 Use of less Difficult Vocabulary

The use of appropriate vocabulary is an important component of test items construction. When a test in Mathematics contains difficult words beyond the level of the testees, it is more or less measuring the testees' knowledge in comprehension in English Language and not their cognitive knowledge of Mathematics.

For instance, in a Junior Secondary Class I (JS I) Mathematics test, the students were asked this question: What is the Quotient of 3 and 4? Many students gave answers such as -1, +1, 7, 12 etc. The term quotient means division and the right answer is supposed to be ¾. In effect, 95% of the students failed this item because of their lack of the knowledge of the term quotient which means division.

2.4 Use of Appropriate Level of Difficulty of Test Items

In building up a test, the difficulty level of the items must be considered. The selection of the difficulty level of the items depends on the purpose of the test as well as the class level of the testees. An achievement test must have an average level of difficulty of items between 0.4 to 0.7 whereas an intelligence test must have an average difficulty level between 0.2 to 0.3. However, the age and class level of the testees must be taken into consideration so that the test will be valid enough.

2.5 Appropriate Arrangement of Test Items

Test developers should try to arrange the test items (for instance multiple choices) in increasing order of item difficulty. When they are arranged from difficulty level to easiness level, the testees tend to spend much time on difficulty items such that majority of them may not attempt all the items. This reversed arrangement (from high difficulty to low difficulty level) affects the testees motivation to continue with the test and by so doing reduces the validity of the test.

2.6 Use of Poorly Constructed Items Must be Discarded

Test developers should be well trained in item generation. They must be conversant with school curriculum, examination syllabus and a good knowledge of the subject matter/content.

When items are poorly constructed, they may contain some clues to the testees which give way to the answer. If they are constructed at the appropriate cognitive level of the testees, the testees will attempt the questions optimally there by having good scores which are their original abilities in the subject. By so doing, the test must be of good level of validity.

2.7 Altering the Pattern of Answers or Keys

In arranging test items to be of good level of validity, items should be so arranged in such a way that the testees will not identify the pattern of answers or keys. For instance, in a multiple choice test in Mathematics with given answers such as 1A, 2B, 3C, 4A, 5B, 6C … etc. Smart testees may easily identify the pattern of arrangement of A, B, and C and use this to the last item. This leads to guessing which lowers the validity of test results. If the pattern is altered across the test length, only testees who know the answers to the items will attempt it and those who don't know it will fail the items. This is a systematic way of improving validity of a test.

2.8 Length of a Test Must be Considered

A test is supposed to appropriately represent the subject matter content and behavioral objectives. If the test is too short; there is the tendency of not adequately covering the contents and objectives of interest thereby sacrificing the validity of the test (content validity). When the tests item are many (i.e making the test length long), there is the high possibility of capturing every important component of the content and behavioral objectives. For instance, a multiple choice test items for Senior Secondary Class three (SS 3) level meant for certification of the students after three years of Senior Secondary education ought to have between 50 to 60 items. This will help the test developer to achieve a high level of content validity in the test.

2.9 The Length of time or Duration of a Test

Earlier in this paper, it was mentioned that validity of a test is very likely to be improved and ensured when clear cut instructions are clearly stated. It is also important that instructions such as the length of time or duration f a test be made known to the testees. The length of time also depends on the difficulty level of the test items, subject, and class level of the testees as well as the length of the test. A test in Mathematics with 50 items at JS 3 level is not expected to be at the same duration when the test length is 30 items. The duration is an index of how fast they can finish the items in the test if all things being equal. In every examination be it school or public examination, the duration of the examination matters a lot so that the testees will be properly guided to finish at the expected time and cover enough items.

3.0 CONCLUSION AND RECOMMENDATION

This paper has tried to present vividly the importance of tests in every teaching-learning process. It also highlighted the importance of validity as a critical characteristic of every test because it gives meaning and value to any test. Specific kinds of validity of a test where outlined among which are content validity, face validity, predictive validity, concurrent validity and construct validity. Means and methods of determining and estimating all of them were dully outlined to aid those who are new in the practice of testing and measurement of students' learning. The paper also identified several strategies which must be considered in the improvement of test validity during the development stage of any test be it school-based or public examination such as use of clear instructions and unambiguous terms in the body of every test item, use of less difficult or non-confusing terms/words beyond the level of the testees, increasing the length of a test, allowing the duration of the test to be within the level of the testees psychological conditions and length of the test and appropriate arrangement of test items according to the increasing order of their difficulty indices.

Against this basis, the following suggestions and recommendations are therefore proffered:

- Public examination bodies should ensure that their test items are trial-tested before they are composed into test for use in their examinations;
- Test items should be arranged according to the increasing order of their difficulty indices in the body of every test;
- Table of Specification must be developed and carefully used during test development stage;
- Experts in test development should be used in editing; vetting and composing test items into test forms before they are used.

4.0 REFERENCES

- Becker, L. A (1999): Reliability and Validity Part II in Anikweze C. M. (2009): Simplified Approach to Educational Research.

- Bloom, B. S et al (1956): Taxonomyof Educational Objectives, Handbook I: Cognitive Domain, London: Longman Company Limited.
- John, W. Young, Youngsoon S. O. & Gary J. Ockey (2013): Guidelines for BestTest Development Practices to Ensure Validity and Fairness for International English Language Proficiency Assessments, Educational Testing Service.

- Okpala, P. N, Onocha C. O & Oyedeji O. A (1993): Measurement and Evaluation in Education; Stirling-Hordaen Publishers (Nig.) Ltd.