Assuring Quality Through Adherence to Due Process in Test Development In Nigeria

By

Prof. B. A. Salim (OFR); P. I. Odor Ph.D and T. N. Kolo Ph.D
Joint Admissions and Matriculation Board (JAMB),
Nigeria.

A paper presented at the 32nd Annual Conference of IAEA held in Singapore from 21st – 26th May, 2006.

Assuring Quality Through Adherence to Due Process in Test Development In Nigeria

By

Prof. B. A. Salim (OFR); P. I. Odor Ph.D and T. N. Kolo Ph.D
Joint Admissions and Matriculation Board (JAMB),
Nigeria.

Abstract

The paper attempts to define quality of a test in terms of properties that are intrinsic to a test item as well as those that are extrinsic to the test as an instrument. It stresses the need for laying a solid foundation in test development such as capacity building and following **due process** in the actual test development process as essential steps in assuring a quality test.

1.0 What is Quality?

The word "quality" has been variously defined. Dictionary definitions of quality include a high degree of goodness, worth or excellence (Oxford, 1974; Chambers 1983; Webster 1992). Bowerman and O'Connell (1997), defined quality from the point of view of the commercial/industrial world. To them, quality can be seen from two perspectives as:

(i) "fitness for use" in terms of the product performing as desired/determined by the consumer; and
(ii) the extent to which consumers feel that a product/service exceeds their needs and expectations.

Furthermore, Bowerman and O'Connell (1997) identified **three focal points** in the assessment of quality. These includes the **design;** i.e. the raw materials and

the manufacturing specifications that must be met if the unit is to operate acceptably, **conformance to the design**; i.e. the ability of a process (the development process) to meet the specifications set by the design and **level of performance;** i.e. how well the product or service actually performs in the market place.

The essence of quality of design and conformance to the level of design is further captured by Ogunniyi (1986). In his opinion, a good quality product does not come out of a vacuum. In his words, "It never did and it never will". Like Bowerman and O'Connel (1997), he also noted that the quality of a product is described by its design and conditioned by the production process.

A number of things can be discerned from the above. One is that quality is assessed from the point of view of the consumer. The customer determines the extent to which a product/service can be described as "fit for use" and whether the product or service exceeds his/her needs or expectations. Customer expectations per say often appear to exist in the realm of the unknown. The product/service developer has to task his intellect and professional genius to guess or arrive at what could be appropriate. The "guessing" activity is never carried out in a vacuum, there is always the need to fall back on the essential substrate, which is the underlying professional skills, talents and competences. Thus, developing a product/service is never the activity of the unschooled in the area of expertise concerned – if quality is being sought.

Secondly, quality is a product of fore-thought rather than an after thought or happenstance. It is a product of a well thought out plan/design. Achieving the plan or the design requires following some processes in which every step is identified and followed meticulously to ensure that the resultant test is fit for the purpose for which it is designed.

Taking a cue from the foregoing, this paper looks at quality of a test from two perspectives: **intrinsic and extrinsic**. Intrinsic refers to properties innate to the materials that form the major components of the product/service. Extrinsic refers to the aesthetic properties of the final product. These two perspectives can be demonstrated using so many products. Take a piece of upholstery chair for example, on the surface, the cover materials may be shining, pretty and attractive. To an unschooled eye, the chair may be considered to be of super quality. Are many of us not deceived today into buying such flashy chairs? But the quality of the chair is defined more by aspects of the chair that may not be visible to the ordinary eye. For example, our perception of the chair as a quality product may breakdown if we begin to examine such aspects of the chair as the quality of the wood, the cover, the thickness of the nails etc, it was made of. Thus, the actual quality of the chair becomes determined by the sum of the qualities of the critical component parts. The qualities of these critical parts thus become the defining factors in the final determination of the quality of the product. These are the indices of quality of the chair.

1.2 Quality In Testing

In the context of educational assessment, Moahi (1997) sees quality assurance as a process and product - based concept in which every stage in the manufacture of a product is identified and fine – tuned to the highest possible level to ensure that the resultant examination is fit for the purpose for which it were designed. Kpodo (1997) defines quality in educational assessment in terms of performance of the instrument or the predictive functions of the test. She observed that assessment may be narrowed down to the degree to which a given assessment shows consistency with the actual future performance of the person assessed.

In a testing situation, **intrinsic quality** refers to properties that are internal to the test. It refers to those properties that relate to both the designer and the design and the actual process of developing the test items. They may include the test design in terms of the test specifications and the processes adopted to ensure that the final product reflects or approximate to a considerable degree the test specifications. **Extrinsic quality** relates to properties that may be external to the test but may affect perception of the test and the products by stakeholders. It also relates to effort aimed at reducing the influence of extraneous factors on performance in the test.

1.3 What Intrinsic Quality in Testing Entails

Achieving intrinsic quality in testing relies on two factors: the **man** and his **predisposition** to follow the due process of test construction. By the man, we mean the people in charge of the test development process as well as the actual item writers. For any test development activity, the bedrock is knowledge. The individuals involved in the process must have the requisite knowledge, skills and competencies that form the focus of the examination. They must have adequate knowledge of the academic skills and objectives that form the basic frame work on which the items are written and evaluated, as well as be able to identify the task being assessed by a given individual test item.

In addition, the individuals must have adequate (not smattering) knowledge of the content of the subject matter on which the test is being developed. Of particular importance is the individual's ability to operate beyond the specific content and begin to recognize the multi-disciplinary nature of knowledge i.e that extraordinary realm where the test developer begins to see knowledge not as a specific content or subject matter area but as one whole which students should be helped to master in order to make better success in life. For the problems of life

do not come as specific content or subject matter but as issues which may require a student to achieve a right mix of knowledge to arrive at a solution.

The foregoing emphasize that, in a test development environment, the actors should have adequate foundation in both the principle of testing and the subject matter that is the target of the test. There is no gainsaying that an adequate foundation is the basis of a solid building. The converse is also true that a no foundation or a weak one at that is a recipe for collapse of or crack in the structure.

The question then is whether test development agencies especially those in a developing nation have the aforementioned actors. The answer may not be very encouraging. It is worth acknowledging that test development to a third world nation, is something that has come to be embraced without a thorough understanding of the various aspects that form the basic framework. For example, in a Nigerian university, students would hear so much about the Bloom's Taxonomy of Educational Objectives in a measurement class. Very few however, come in contact with these educational objectives beyond excerpts from the main book, which may be presented to them by the lecturers in charge. In other words, they do not come to grips with the actual definitions or descriptions of the six levels of the cognitive components. Relatedly, not much is done with explaining the meaning of each component and sub-components. Thus, most pre-service teachers leave school with nothing more than shallow knowledge of the objectives of education as contained in the Bloom's Taxonomy. Yet, when they come out, they are expected to perform as experts in test construction. One then wonders, what such teachers can do when they lack understanding of the basic knowledge that guides thoughts during test construction.

This is complicated by the fact that understanding of the various educational objectives as contained in the Bloom's Taxonomy is not a simple task. The present approach where the examination bodies gather only subjects specialists

from the universities for purposes of curriculum development, test construction and moderation misses the mark. These subject specialists may be tempted to test only content without due recognition of the cognitive tasks that the educational process is expected to achieve in the first instance.

Coupled with inadequate pre-service education in testing, our initial contact with testing as an organized human endeavour is another source of problem. In most developing nations, initial contact with examination is through external examination bodies in London and Cambridge. With time, these examination institutions metamorphosed into regional examination bodies e.g. the West African Examination Council (WAEC) for the West African Region. Staff of these early examination bodies learnt the practice of testing firsthand from the colonial officers. What they actually developed was a 'cook-book' method or routine of activities that detailed what should be done in order to have a test. Of worth noting is that this transfer of knowledge/methods was not accompanied with development of theoretical basis for what is practised.

As each nation began to create her own national examination bodies, these new bodies often poached foundation staff from these regional examination bodies. Such foundation staff who lacked the theoretical knowledge of testing had no option than to transfer the earlier learnt routinized activities. Often, when questions are asked, on why certain things are done or why it should not be done in a different way, the response is always 'this is how we have been doing it from the beginning'.

From the above, what comes across is that the practitioner often lacks the knowledge required to operate effectively in a testing environment, hence they need to develop remedial programmes. In developing such programme(s), special note should be taken of the need to achieve adequate knowledge of the objectives as well as a common language and agreement among the test developers on the definitions of these objectives as a necessary first step to build quality tests.

These can be done through creating periodic forum where experts in the subject matter areas, experts in the educational psychology, sociology, tests and measurement and practising teachers can come together and brainstorm on the definitions. The interactions at such periodic fora, would create a common understanding among the participants on what could be regarded as acceptable behavioural expressions of each component and sub-component of the Bloom's taxonomy of education. Such fora too, would build up a body of knowledge that can be used at various levels of (pre-service and in-service) of prospective test developers. Armed with the knowledge therefrom and the confidence instilled in them, through the interactions, products of such meetings will be more predisposed to write more focused and thus, more quality items.

The Bloom's taxonomy has been used in this case to demonstrate what can be achieved through focused discussions. The interest may also be on more abstract qualities, such as aptitudes, skills and competencies. The ultimate is that, we have a growing crop of people, experts, teachers, etc who understand what to focus the items on and who can easily determine when another author of an item has not done it well.

Another aspect related to the test developers is that they often practice the art without being provided any rules within which to operate. What is being said in essence is that there is a danger if each item developer is left to operate with his/her own criteria. It will be better if the owners of an examination define the parameters of operation for the test item writers. In this case, the test developers would know ahead of time the criteria for assessing the test items and thus, work within them. Furthermore, peer review of the items could attract greater agreement among the reviewers. This guide for item writing and review can be provided in the form of **item writers manual.**

2.0 Following Due Process in Test Development

This requires following religiously and honestly the psychometric processes required in testing. According to Moahi (1997) this process includes determination of the purpose of the test, development of test blue print and item writer's manual, pilot testing, item analysis to determine validity and reliability, modification of non-functional items, selection of items for the test, production and administration of the test, scoring and analyzing the results. Pidgeon (1997) also listed 8 steps which ought to be followed when an examination is being prepared. They are: identification of purposes; development of test plan or blueprint; development of detailed specifications; preparation of appropriate items; review and editing of items; preliminary trial – testing and item anlysis; final selection of items; organization of questions and production of question papers.

In this paper, effort is made to discuss these processes with special emphasis on the areas considered very pertinent to assuring the quality of a test. These include:

(i) Develop the Syllabus

To ensure quality, examination agencies review their examination syllabus periodically. The objective of the review is to update the contents of the existing syllabus. In reviewing the syllabus, new inputs are drawn from relevant stakeholders including subject specialists from secondary and tertiary institutions and experts in curriculum development. Mehrens and Lehmann (1991) describe what an ideal review section should look like. They note that it is a meeting of curriculum and subject matter specialists. After a thorough study and analysis of syllabi, textbooks and programmes throughout the country, a list of objectives is

prepared i.e. information candidates/students should have, principles they should understand and skills they should possess.

Syllabus development is another form of curriculum development. It should contain the goals to be achieved by each subject and the specific objectives that should be achieved through the teaching of each content. In this wise therefore, participants at the syllabus development should be a mixture of subject specialists and specialists in education. This is to ensure that the knowledge of both groups complement each other in arriving at both credible content and related objectives. Over-emphasis on subject specialization may deprive the syllabus review effort of the wise counsel of the educationists, especially in terms of sharpening the objectives. Also, the in-house staff responsible for a subject, should be both a subject specialist and an educationist. This is to ensure that the subject officer brings to bear both aspects of his training in collating the products of the syllabus review effort.

(ii) Develop the Test Blue Print

The test blue print is simply a matrix, showing the objectives to be measured and the content where these objectives would be measured. At this phase, the objectives to be sampled by the test are then reduced to a test outline or table of specifications. This is based on the judgments of the various experts involved in the test planning. It guides the test maker in constructing the test (Mehren and Lehman, 1991). On the surface, the development of this important guide may look simple but in operation it is very complex. This arises from the fact that, this is the point where the professional knowledge (i.e. those in the subject areas of interest and those from education) comes to bear. Failure to achieve the right mix may set the test in a very wrong path. It is worthy of note that, because of this

complexity, some test development agencies avoid this phase to the detriment of the quality of the test.

(iii) Write the Test Items

As stated earlier, the item writers' manual is a necessary guide for item generation. It should contain such essential components as definition of various skill levels, provide an overview or checklist for assessing the value of each item and finally give the item writer a proforma that would guide the format of the item. This proforma should contain essential information as the name of the author, the content area to be tested, the skills to be tested, the essential reference materials relevant to the content, the item stem, alternative responses and the keyed response. With these, both the item writer and peer reviewers would have clear idea on what the author of the item has written as well as criteria for assessing the quality of the items.

Often times, some examination bodies just commission item writers without the essential guidelines mentioned above. To assure quality therefore, the examination bodies need to settle down and articulate this essential guide if only to get the best out of the item writers.

(iv) Validate the Test Items

This stage involves the preliminary screening of each test item for relevance. The essence is to determine whether an item submitted by an item writer conforms to the objective or skill level indicated for it by the writer. Operationally, experts check the stem and make necessary corrections taking into consideration the objectives being tested. They also look at the alternative responses to see whether they are plausible detractors. Again, they look at the keyed response to ensure that it is the

best among the alternatives and that it is the most correct answer to the question. Ideally, a proforma or review guidelines are provided and individual raters are supposed to achieve a considerable level of agreement among themselves on the checklist provided.

In practice however, some examination agencies while conducting this, fail in two areas. The first is that they may not provide any checklist and as a result, deny the process, the essential rigour and the desired output. Secondly, it may be an all comers activity where expertise is limited to knowledge of the subject matter area.

What takes place at this stage is called moderation by some examination agencies. It may be regarded as checking for face validity. To some of the examination agencies, this may be the only check for validity of the test items. This however, should not be so, as there is need to go further. For according to Mehrens and Lehman (1991), face validity is not really validity at all in the technical sense of the word. It simply refers to whether the test looks valid on the face of it. It is a desirable feature of a test in the sense that it is useful from a public acceptance standpoint.

(v) Pilot or Trial Test the Items

As stated earlier, stopping at the moderation phase does not assure complete quality of the test. There is still the need to go a little bit further to trial test the items. At this phase, the items that passed the test of the moderation phase are put together into a test form and presented under an ideal testing condition to the potential candidates for the examination. Their responses are scored and each item is checked for index of difficulty and index of discrimination to make sure that each item is operating at the required level. The response alternatives are checked to ensure that the keyed response is not a give away. Also, the items in the test as a group is

subjected to relevant statistical analysis to determine those that hang-together, the task levels they represent and the reliability of the items in each dimension identified as well as the reliability of the whole items in a test. Items and group of items that pass these final tests then form candidates for the item bank. Those that can be redeemed are then modified and scheduled for future trial testing.

This phase should be a joint activity of the subject specialists, measurement experts, educationists, statisticians/computer specialists, etc,. Not having a complete team may deprive this exercise of complete expert opinion. The rigour involved at this stage appears frightening to many examination bodies who thus, try to circumvent this phase to the detriment of the quality of examination and its credibility.

Mehrens and Lehman (1991) pointed out that after careful review (moderation) and editing, the tryout or experimental items are arranged in a test booklet. Then, the instructions to both administrators and pupils (candidates) are written and the tryout tests are given to a sample of pupils (candidates) for whom the test is designed. After the answer sheets have been scored, an item analysis is made to identify the poor items. In addition, comments from test administrators pertaining to timing and clarity of instructions for both administrators and pupils (candidates) are noted. Further editing is then done on the basis of the item analysis (or more items are written if needed and the content validity is rechecked). The test is then ready to be standardized. After a representative sample of pupils (candidates) has been selected, the refined test is administered and scored. Reliability and validity evidence is obtained and norms are prepared for the standardization sample. Throughout the item writing stage, attention is given to preventing racial, ethnic and gender bias.

(vi) Develop the Test

This requires determining the format of the test, which should include the total number of items that will form the test, total number of items that will represent each skill level of interest (as may have been determined in the test blue-print), the arrangement of the items (may be from simple to the more complex, etc) and the test instructions. All these require expert input and consideration and should be agreed upon before implementation. With these done, the credible items in the bank can now be pulled to form the test based on the agreed format.

3.0 What Extrinsic Quality in Testing Entails

As said earlier, this refers to efforts aimed at reducing the effect of extraneous factors on the performance of candidates in the test. It borders mainly on the integrity with which the test development process is handled, that is, from the initial test construction to its actual presentation at the test administration phase. It also includes the way the test papers and instructions and the answer sheets are structured to ensure that candidates are not confused, that all forms of malpractice are curtailed to the barest minimum and resulting scores are not compromised or manipulated.

Some examination bodies concentrate only on ensuring the integrity of the testing materials and the resulting test scores. These they do by ensuring that, right from the construction of test papers to the scoring of answered scripts, adequate security, vigilance, etc measures are put in place to maintain integrity of its examinations. For example, constructed test items are well guarded and secured to avoid any possible leakage. Since this is an area most examination bodies concentrate on, it will not be dealt with in any further detail.

4.0 Conclusion

This paper arose from observations and discussions on the practice of test development especially among external examination agencies. The finding is that oftentimes, the major emphasis of some of examination agencies appears to be that of convincing themselves and the general public that the examination is of high integrity. It appears that, as long as the integrity of the examination is not violated to a noticeable degree, the examination body can beat its chest that it has done a praiseworthy job. Often, the essential indices of quality are never brought to the public domain thus, denying the consumer the opportunity to evaluate the quality of the examinations.

Failure to follow the essential steps (due process) in test development may be attributed to a number of factors. The factors include: lack of requisite staff to carry out the work; lack of resources to go the whole hog; lack of will to carry through the due process of test development and the pressure of time to meet each target examination date.

With respect to lack of requisite staff, there is need to admit that there is absence of skilled manpower within the testing industry. We need to know that we have critical skill shortages in the testing enterprise. Most testing agencies have a high pool of staff but when it comes to skill, there is a serious deficiency in the system. To make up therefore, examination bodies have to:

(i) appreciate the need to develop and install programmes that would promote in-house training of staff in measurement and evaluation;

(ii) appreciate the need to develop and install such a staff training policy that would encourage staff to participate in post-graduate studies in educational measurement and evaluation in recognized institutions and

(iii) make conscious effort to recruit specialist Nigerians in diaspora in order to boost present specialist composition of the examination bodies.

To assure that examination bodies meet deadlines for the examination while not violating the due process of test development, examination bodies should take advantage of emergent automation systems to develop effective and secure item banks.

References

Bowerman, B. and O'Connel, R. T. (1997): Applied Statistics, Improving Business Processes, The McGraw – Hill Companies, Inc., IRWIN, Chicago U.S.A., 1997.

Hornby, S; Cowie, A. P; and Gimson, A. C: Oxford Advanced Learner's Dictionary of Current English, Oxford University Press, London, 1974.

Kirkpatrick, E. M; Richard Clay: Chambers 20th Century Dictionary, The Chancer Press Limited, Suffork, England, 1983.

Kpodo, L. (1997): "Quality in Educational Assessment and its Indicators", Proceedings of 15th Annual Conference of AEAA held at Sheraton Hotel and Towers, Abuja, 21-27 September, 1997.

Mehrens, W. A. and Lehman, I. J. (1991): Measurement and Evaluation in Education and Psychology (4th ed.), Wadsworth/Thomson Learning, Canada, 1991.

Moahi, Serara (1997): "Improving the Quality of the Primary School Leaving Examination in Botswana: Why criterion Referenced Testing?" Proceedings of 15th Annual Conference of AEAA held at Sheraton Hotel and Towers, Abuja, Nigeria, 21 – 27 September, 1997.

Ogunniyi M. B. (1986): "The Inevitability of Mass Failure in our Public Examinations", Proceedings at the National Conference on Mass Failure in Public Examinations in Nigeria, 21 – 25 April, 1986.

Pidgeon, D. A. (1967) "Modern Look at Examinations". Physics Education Vol. 2, 1967.

The New Webster's Dictionary of the English Language, International Edition, Lexicon Publications Inc., New York, U.S.A., 1992.