Chapter 18

Tenets of Developing Standardized Test and Examination Items

Rabiu, M. B.,
Abdulrahman, M. A.
&
Wushishi, D. I.

Introduction

The pervasiveness of examinations at all levels of Nigeria's education system is a reflection of its usefulness in selection processes and also for placement purposes in our school systems. Developing test and examination items can be a very difficult and challenging task for even the most experienced teacher after completing the onerous task of instruction at any level of our education system. Test or examination should not be seen as the end of teaching and learning processes but rather, an instrument of determining whether already predetermined goals or objectives set at the beginning of the class were achieved or not. The test or examination must always evaluate students' performances so that the scores obtained can be used to represent their achievements in an accomplished task. The "standards" presented herein were developed by leading experts in test, measurement and evaluation with inputs from several sources acknowledged. Some of the procedures highlighted are known because they are being adopted by agencies charged with the responsibility of educational assessment like WAEC, NABTEB, NTI, COREN and NECO amongst others in Nigeria. Therefore, to adjudge a test or examination to be standard, it has to satisfy some other parameters. And, the procedures towards achieving that are: developing items blue print, test creation, review items, items validation, establishing the passing score, items distribution, items administration and reporting (Brown & Smith, 1997).

What then is the validity of a test? Validity of while a test is the ability of the test to measure that which it purports to measure, while the reliability of a test is the consistency of the test in measuring students' ability over a relative period

of time (Sambo, 2007). The question of usability of a test arises from the degree to which teachers and external assessors deploy the test instrument with minimum expenditure of time, energy and money. Developing standardized test or examination items requires strict observation of detailed processes by experts and educators working together in generating the items.

Learning Outcomes

It is hoped that at the completion of this text, you should be able to accomplish the following:

1. Identify the steps or procedures involved in writing standardized test or examination items;
2. Determine the role(s) of subject-matter experts (SME's) in the process of developing test items blue print.
3. Determine the purpose of generating test or examinations items, either for commercial public examination or for local school purposes of grading and placement.
4. Determine the role of a psychometrician in conducting review of items generated which is considered very critical for validation of the whole examination that would be conducted.
5. Determine the constituents of items validation in test construction and be able to undertake the process.
6. Determine grading procedure, distribution and custodians of questions before the conduct of an examination.

The procedure established hereafter are steps to proper generation of examinations items. They are not sacrosanct as they may differ from country to country and even perhaps within the same country or state depending on the purpose the examination is designed to achieve:

Procedure No. 1: Developing Test Items Blue-Print

Before developing test items blue-print, the purpose of developing them must be clearly defined. The first critical task is that of identifying knowledgeable individuals i.e. Subject-Matter Experts (SME's). Having people who are novice might create so many problems which eventually will undermine the purpose. That is why experts and highly recognized professionals with many years of on-the-job experience are most paramount. These individuals are selected from the various disciplines in locations/bodies where the examinations are intended for.

The number could range from 8-15. However, this may depend on the nature of items to be developed but not dogmatic. The essence could be to have many of them that may be required in other phases of items development.

The purpose of developing a test blueprint is to identify the attributes the examiner wants to measure. This will ensure that items to be assembled are consistent and, they will be statistically relevant. The examination blueprint should contain basically, the purpose of the examination, description of the target audience, total number of items on the examination, number of items measuring the domain or objectives earlier selected, content outline, examination format and item types (multiple choice, short answer, fill in the blank e.t.c). So, the format of examination to be used must be determined at this point, whether it is multiple choice questions (MCQ) alone that is the mostly preferred especially with electronic examination (e-exam) or in combination with essay (written/open response) format. Researchers can consult literature on rules of preparing MCQ's which are available in library shelves or online

Procedure No. 2: Test Creation

The moment the test blue print is finalized, the process of test or examination items creation starts which may vary depending on the beneficiaries or purpose for which it is created. Standardized items creation for commercial or public examinations is more tedious, rigorous and time consuming. Educational experts or experienced teachers, professionals in subject specialized fields are assembled to generate items from a prescribed curriculum syllabus. They must have content know-how on the subject they will develop the items. These specialists are usually from a diverse sections similar to the group of students it is intended for. It starts in an individualized stage leading to group stage to compare notes and make amendments/corrections especially in error statements that are grammatically ambiguous or unclear. Items writers should be knowledgeable about Bloom's Taxonomy (1956) cognitive levels and accompanying qualifying verbs. For instance,

Table 1: Bloom's Taxonomy of Cognitive Levels

Knowledge: it is the least level of cognition. Examples of words used are; identify, list, recall, specify, state, reproduce, select, state, enumerate, read, *et cetera.*
Comprehension: examples of words used are; classify, convert, describe, discuss, distinguish between, estimate, explain, *et cetera.*

Application: words such as apply, arrange, compute, construct, demonstrate, develop, discover, discover, modify, operate, *et cetera*. are implored for its determination.
Analysis: words such as analyse, associate, correlate, determine, diagram, differentiate, discriminate, distinguish, estimate, infer, *et cetera*. are used.
Synthesis: words such as combine, compile, compose, communicate, construct, create, design, develop, devise, express, formulate, generate, *et cetera* are employed.
Evaluation: words such as appraise, assess, compare, conclude, contrast, criticize, decide, defend, discriminate, evaluate, interpret, judge, justify, *et cetera*. are employed for its attainment.

Source: Brown, Race and Smith (1996).

Note: the action verbs listed are not exhaustive of the description for each of the cognitive levels.

Procedure No. 3: Items Review

This stage is critical before items validation which is mostly undertaken by a smaller team of experts or Subject Matter Specialists (SME's) who are psychometrics. They check whether items meet standards for which the test or examination is intended for. The process requires experts who will thoroughly review items for accuracy, relevance, quality, alignment, standards, clarity and removal of potential biasness or clues. The responses provided must include quality distractors that makes it difficult for a below average students or candidates to quickly select the correct option from the provided choices. It is after this process items are kept in items - pool prior to field-trial.

Procedure No. 4: Items Validation

Immediately after items are generated by SME's before field-trial, they are checked by validators to determine whether items are likely to perform the intended functions using appropriate statistical tools for item difficulty and item discrimination. Each item is exposed to statistical test first and then reviewed when consensus is arrived at from a psychometric perspective. The reliability index obtained is used for interpretation of whether the item generated is valid or not.

Table 2. Reliability Interpretation Indicators

Reliability Coefficient	Interpretation
0.90 and above	Excellent reliability; at the level of the best standardized tests.
0.80 – 0.90	Very good for a classroom test.
0.70 – 0.80	Good for a classroom test; in the range of most. There are probably a few items which could be improved.
0.60 – 0.70	Somewhat low. This test needs to be supplemented by other measures (e.g., more tests) to determine grades. There are probably some items which could be improved upon.
0.50 – 0.60	Suggests need for revision of test items, unless it is quite short (ten or fewer items). Otherwise, the test definitely needs to be supplemented by other measures (e.g., more tests) for grading.
0.50 or below	Questionable reliability. This test would not contribute greatly or heavily to the course grade, and it needs revision.

Source: University of Washington, Seattle, WA (2016).

The determination of item difficulty and discrimination indices provides more probabilistic measure of internal consistency for items generated and the instrument for determining the strength or quality of the test or examination is as follows:

Item Difficulty: is the procedure for determining the proportion of students who answer an item correctly. Item difficulty is supposed to provide a P-value notation which indicates how difficult (which is low P-value) or too easy (which is high P-value) is a question. Items with low P-value (0.03) are recommended for review for either the wordings are unclear/ambiguous while items with P-value greater than 0.75 may be too easy. An item that has P-value of 95% is too easy because it is unable to discriminate between knowledgeable students and those who picked the answers by chance and should be flagged up for review. At the end, items are listed according to their degrees of difficulty (easy, medium & hard).

Item Discrimination: is the predictive ability of an item to correlate with a student total score. Item discrimination indicates also how well a test differentiates amongst students on the basis of how well they know the material being tested. Negative discrimination arises when low performing students answer the items correctly. In some situations, both low and high performing students have same scores on items which is known as zero discrimination. Where high performing students answer correctly the items low performing students fail to answer is referred to as positive discrimination.

On the whole, items are categorized as good, fair and poor. Therefore, whenever the discrimination index is below 0.25 the item must be revisited, perhaps the wordings may be responsible for high performing students selecting the incorrect option.

It is important to note that, the P-value notation and item discrimination index are only requirements that serve the purpose of guidance to test developers in selecting items for test or examination purposes. Even though an item with excellent statistic like, P-value of 0.65 and discrimination index of 0.80 appears to be a good item, nevertheless, it may be as a result of guess work of selecting options A and B, because none or few students selected options C and D. Such an item can be flagged for further review.

Procedure No. 5: Establishing the Passing Score

Once items review and validation are completed, score or grade for pass/fail must be established as a criterion for passing or failure. This standard is categorized into normative and absolute. The normative standard is established on the basis of students' performance relative to available spaces/slots. This form of standardization is used by federal and state agencies for determination of placement or slot for students especially for scholarship awards, since the number needed would have been determined prior to the written (aptitude) examination. JAMB examination body uses the normative standard when admitting students into different programmes in the tertiary institutions of learning. While the absolute standard is more or less like the criterion referenced validation where the pass/fail score is determined based on students' attainment of certain minimum standard. This is like achievement examinations used for the purposes of certification to ensure satisfactory completion and attainment of certain competency level.

Procedure No. 6: Items Distribution

After establishing passing score or cut-off mark and grade as the case may be, the items are then distributed in a normal school setting for students to answer in an examination condition or to Centres and schools in commercial settings. This stage is complex in commercial examination settings which is similar to WAEC and NECO which requires strict monitoring to prevent sharp practices that may invalidate or jeopardize the whole examination exercise. At the end, the number and percentage of students who are chosen is alternative to each other.

Procedure No. 7: Items Administration and Reporting

Items administration is the point at which the students are kept under same examination physical condition. If the examination is taking place in different environments, it must be scheduled for the same time and date. Students should be informed well ahead of time the scheduled period for a proposed examination to allow for adequate preparation in order to remove test-anxiety. Reporting will include test or examination score recorded after administration and marking. With the advent of computers, scoring has almost become timeless using automatic scoring machines that give result almost instantly with Computer Based Test (CBT). Essay examinations and other formats require usually the use of manual marking and scoring. The results are reported in percentages that is in respect of those who passed and failed while pictorials and descriptive charts can be used for presentation for the purposes of easy comprehension and analysis by administrators to parents and other stakeholders.

Conclusion:

The procedures discussed are by no means exhaustive or static standard that are not reviewable. They are intended to provide some form of guide for generating valid and reliable test and examination items especially for multiple choice and essay examinations. Teachers and particularly lecturers in tertiary education levels may conclude that the procedures enunciated are didactic land inflexible therefore unrealistic for practice in the context of University for instance. What cannot be dispensed of, is that, it provides a step-wise approach to items generation and can be used where the objectives to achieve quality and realistic examination data is desired.

References

Bloom, B. (1956). Classification of Educational Objectives Handbook I: Cognitive Domain. New York Toronto: Longman, Green.

Brown, S., Race, P., & Smith, B. (1996). 500 Tips on Assessment. London, UK: Kogan Page.

Brown, S., & Smith, B. (1997). Getting to Grips with Assessment. Birmingham, UK: staff and Educational Development Association.

Reliability Interpretations (2016). University of Washington (Seattle, WA). Office of educational Assessment. http://www.washington.edu/boundless

Sambo, A.A. (2007). *Research Methods in Education*. Sterling-Hoden Publishers (Nig.) LTD, Lagos, Ibadan, Benin-city, Jattu-Uzairu