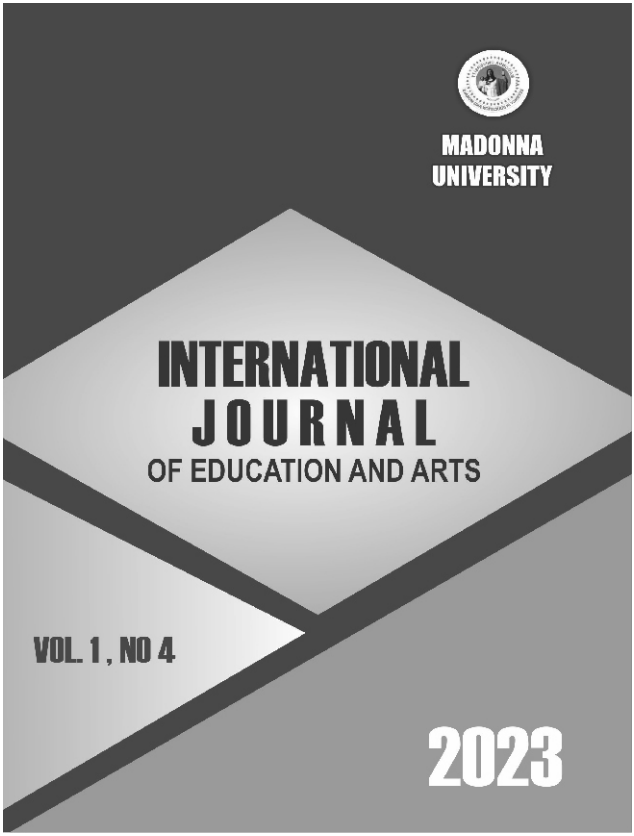
MADONNA
UNIVERSITY
INTERNATIONAL
JOURNAL
OF EDUCATION AND ARTS
VOL. 1 , NO 4
2023

International Journal of Education and Arts Vol.1 , N0 4 Nov. 2023

Developing Standardized Tests and Marking Guide for Effective Learners' Assessment

Dr. Caroline Ifeyinwa OKORIE
Department of Computer Science Education
Faculty of Education and Arts
Madonna University Nigeria.
Tel.: +2348067943728
Email : okoriecarol12@gmail.com

Abstract

The aim of this paper is for instructors to learn how to showcase their skills in preparing valid, reliable and useful items including test planning, test preparation, test scoring/grading based on defined marking guide for effective learners' assessment. This is to promote Fairness, objectivity and consistency in evaluating learner's performance. The paper discusses the different types of tests, basic considerations in classroom test development, and explanation of different steps in test construction and how to prepare a defined standard marking guide. The paper also emphasizes the importance of transparency and continuous improvement in the learner's assessment process. The paper creates awareness that a good test well-constructed and the marking guide well prepared reflects the goals of the instructor. It will equally make instructors to avoid subjectivity and bias. This paper also shows clarity and accuracy of the assessment material crucial in achieving reliable results which will enhance and facilitate tests development and marking. This paper will make instructors to learn how to construct good items and marking guide. I recommend that instructors should attend seminars and workshops to upgrade their knowledge because they are the

driving force in any educational system. They are also the fulcrum of the Education Act

Keywords: Developing, Standardized Tests, Marking Guide, Effective Learners' and Assessment.

Introduction

Assessment plays a crucial role in evaluating learner's understanding and knowledge acquisition. To ensure the validity and reliability of the evaluation process, setting standardized tests and creating a well-defined marking guide are essential. This paper provides a comprehensive guide for educators to develop effective assessment.

An assessment test is a measuring instrument designed for assigning figures to educational and psychological attributes for passing judgment and drawing conclusion. Tests are carried out to know the mental and intellectual abilities of a testee or a learner in order to expunge, evaluate and pass judgment. A doctor cannot give treatment without tests; a tailor cannot sew a cloth without using a tape to measure you.

In educational world, an instructor cannot draw references about a learner's intelligent quotient without testing a person with a particular type of test. When test items are gathered, organized, prepared, constructed and scrutinized by experienced test developers, it is called a standardized test.According to Isaac and Ibe (2017), being fully grounded with testing skills, as it expected, instructors are to determine what to be learned and then define same so precisely that test items constructed by them should show the desired performance and serve useful purposes. This is more so because, appropriate assessment of learning is an essential aspect of teaching learning process. It is important instructors harness the various testing skills, during testing, if the results of their test must be objective in revealing teaching learning effectiveness. These skills, which are however steps in preparing valid reliable and useful test items, include; test planning, test preparation, test administration,

test scoring/grading, test interpretation and item analysis.

A standardized test is a test that is administered and scored in a consistent or "standard" manner. Standardized test is designed in such a way that the conditions for administering, scoring procedures and interpretations are consistent and done in a predetermined manner. Any test in which the same test is given in the same manner to all test takers, and graded in the same manner for everyone is a standardized test.Standardized test is designed to permit reliable comparison of outlines across all test takers because everyone is taking the same test. There are two major categories of human attributes; those that can be physically seen and those which could be visualized. Among those that can be physically seen are height or distances measured with a ruler, tape marked out in units of meters, or weight marked with the use of balance-scale marked out in kilogram units. Those that cannot be seen include aptitude, achievement, intelligence, interest and other aspects of personality. Similarly, in education such attributes are measured with the use of test items to which numbers or scores have been assigned.

We use visualized type of test like achievement test, aptitude test, personality standardized test and interest standardized test to measure them. For instance, a child's mathematics aptitude is not seen in him, rather it is deduced from manifestation of certain behaviors or characteristics. Such attributes are known as constructs.To measure an educational construct therefore, one has to describe the characteristics associated with such construct in behavioral terms. For example, one can describe mathematical aptitude as ability to add, subtract, divide and multiply given figures or perform any other mathematical operations as specified. Such mathematical operations are put down in form of test items. Thus, when a group of learners take such a test, their scores would vary according to the number of test items each learner gets right.

These test scores give the teacher or interested person a "quantitative information about the existence of the attributes or constructs mathematical aptitude in each child". A set of test scores therefore, indicated degrees of existence of specific attributes in a

group of learners in terms of numbers.Now we have achievement test which is designed to measure how well an individual has mastered a specific knowledge or skill to which the individual has been exposed (criterion- referenced test) or rank the pupils in order of their performance (i.e. Norm- referenced test).

There are two types of achievement test. Teacher made test and Standardized test.Teacher made tests according to Nworgu (2015) are those tests constructed by classroom teachers individually or in groups for the assessment of their pupil's achievement in specific school subjects. Such tests assume various labels such as quizzes, mid-term test, end of term/year examinations and the like.While Standardized tests according to Nworgu (2015) are constructed by experts. Elaborate procedures and high degree of precision are adopted in the development and standardization of the tests. In other words, standardized tests are modified teacher made tests by experienced personnel. This paper concentrates on standardized tests. Below are tables showing different types of standardized tests and the way they are been developed and used.

Basic Types of Standardized Classroom Testing/Evaluation

Type of Standardized Test	Purpose	Timing	Characteristics	Examples
Educational Placement Test	Determine readiness for a specific educational level or program.	Administered at the beginning or before placement in courses.	Places individuals in appropriate educational programs based on current abilities.	Advanced Placement (AP) exams, Graduate Record Examination (GRE) subject tests.
Formative Test	Provide ongoing feedback during the learning process to inform instruction and student learning.	Occurs during instruction and learning activities.	Informal, low-stakes, not usually graded; helps adapt teaching in real-time.	Quizzes, polls, class discussions, homework, teacher observations.
Diagnostic Test	Identify students' strengths, weaknesses, and prior knowledge before instruction.	Conducted before or at the beginning of a course or unit.	Aims to diagnose areas of difficulty or gaps in knowledge; helps tailor instruction.	Pre-tests, readiness assessments, placement tests.
Summative Test	Measure what students have learned at the end of a course, unit, or instructional period.	Occurs at the end of a specific instructional period or course.	Formal, high-stakes, graded; typically determines grades or course completion.	Final exams, end-of-term projects, standardized tests, midterm assessments.
Achievement Test	Assess knowledge or skills in a specific subject.	Usually summative, at the end of instruction.	Measures individual performance relative to standards or learning objectives.	SAT, ACT, state - mandated assessments.

Aptitude Test	Measure potential to develop specific skills or excel in certain areas.	Typically administered before career or education planning.	Predicts future potential or capacity to learn.	Graduate Management Admission Test (GMAT), Differential Aptitude Test (DAT).
Cognitive Ability Test	Assess general cognitive abilities like reasoning and memory.	Often administered independently of specific events.	Measures general cognitive skills, not tied to content.	IQ tests (e.g., Stanford-Binet).
Personality Test	Examine personality traits, behavior patterns, or preferences.	May be used in various contexts, such as employment or clinical assessment.	Measures personality traits or tendencies.	Minnesota Multiphasic Personality Inventory (MMPI), Myers-Briggs Type Indicator (MBTI).
Employment Test	Evaluate job applicants' qualifications, skills, and aptitude for specific positions.	Part of the employment application and selection process.	Assesses job-related skills, knowledge, and suitability for specific roles.	Cognitive ability tests, personality assessments, job-specific skill assessments.
Performance Assessment	Assess skills or knowledge through practical tasks or demonstrations.	May be administered during or at the end of instruction.	Measures practical application of skills or understanding of specific content.	National Assessment of Educational Progress (NAEP), job skills assessments.

Table 1.1

From the table given above, the various types of standardized tests are given with illustrations clearly stated.

Basic Considerations in Classroom Test Development

These are the steps you take in class testing to ensure that you set standardized questions (tests) and prepare a standard marking guide.

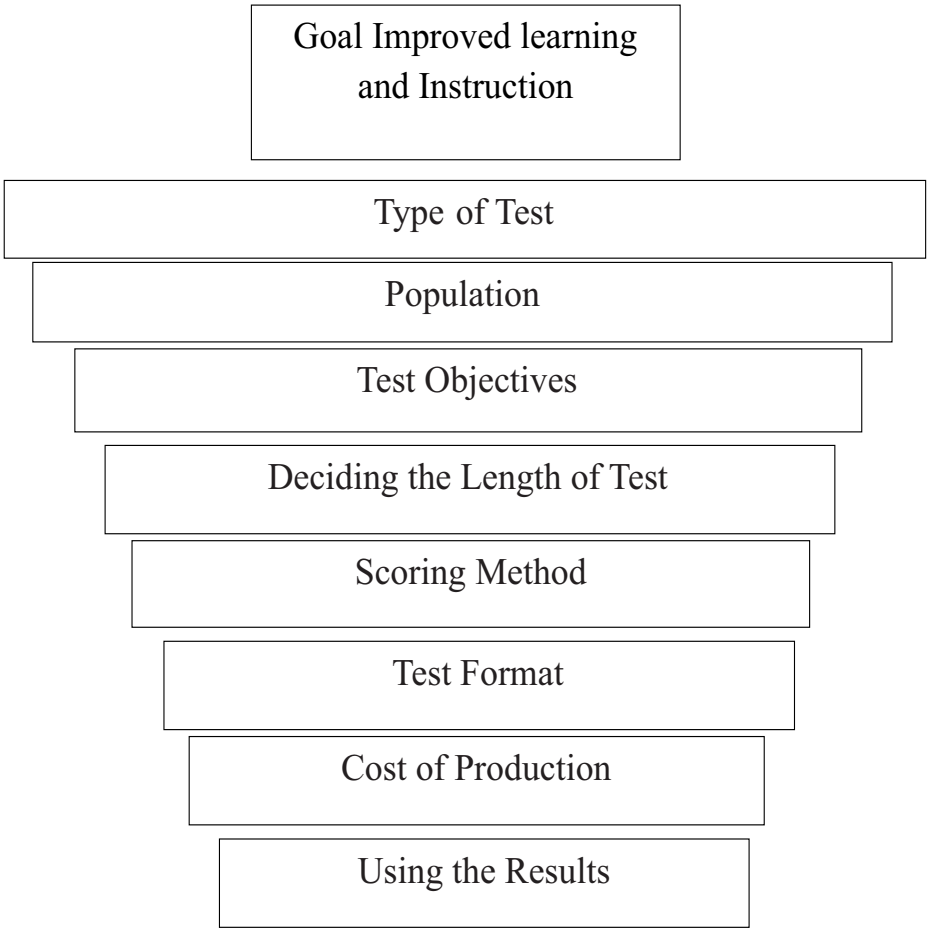

Table 1.2

There are number of basic issues which must be settled by a test developer before embarking on the actual process of test construction:

1. Type of Test: The test developer should know various types of tests such as achievement test, aptitude test, personality test and know which one he/she wants to develop

2. Target Population: The target population should be clearly defined in a way that would help the test developer to make decision

on the type of items, test length and reading level.

3. Test Objective: In developing a test, the test developer must carefully state the 3 levels of educational taxonomy. The test should measure the cognitive, affective and psychomotor domain.

4. Deciding the length of the test: The test developer should make sure that the length of the test is neither too long nor too short. A long test is boring and might not be reliable and valid. Also, a short test may not cover enough content and educational objectives.

5. Scoring method: It should state earlier whether the test will be scored manually or with machine (electronic marking)

6. Test format: It should be stated whether the format of the test could be essay or objective type. If the developer selects essay type, he/she must decide on whether it will be extended-response or restricted-response type. Like those for whom the test is made, they are made by the experts who follow rigorous steps in their construction.

7. Cost of production: The test developer should determine the amount of resources (finance, materials and time) available to him/her for test production and administration.

Any test and indeed any evaluation instrument must satisfy the criteria of reliability, validity as well as objectivity. In the construction of test therefore, a number of steps are involved.

Steps in Test Development Chart

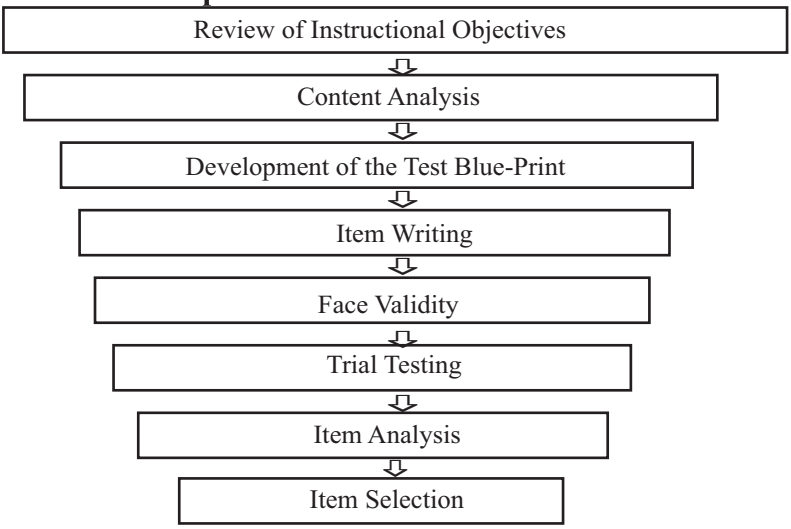

Table 1.3

1. Review of Instructional Objectives

The 1st step in the construction of a test is the review of instructional objectives. Instructional objectives are those behavioral changes which an instructor expects to notice in his learners after they have been exposed to a particular topic. The test developer must consult his unit plan, or his lesson notes in other to obtain the list of instructional objectives. Remember to represent the levels of intellectual functioning such as knowledge, comprehension, application and higher order thinking processes which include analysis, synthesis and evaluation. This will take the test developer to the next step in test development which is content analysis.

2. Content Analysis

Content is derived from the curriculum. According to Nikolopoulou (2023), Content validity evaluates how well an instrument (test) covers all relevant parts of the construct it aims to measure. An analysis of the content should address whether the content meets the current and long-term needs of the learners. What constitutes the long-term needs of the learners is a value judgement based on what one sees as the proper goals or objectives of a curriculum. The content could also be obtained from an officially recommended textbook for the subject. This leads us to the next step which is relating the content and instructional objectives.

3. Development of Test Blue-Print (Table of Specification)

A test blue-print is a two-way grid of table which specifies the level of objectives as they relate to the content of the subject. It is a means of ensuring content validity of the test. Usually, the objectives are written at the horizontal part of the table while the contents or topics are written on the vertical part of the table. By the listing of the objectives, the test developer can assign weights to each cell or column in accordance with the importance attached to each objective or topic. The weight in each cell indicates the number of test items to be devoted in each objective and each topic.

By constructing test items which fits these specifications, it is possible to build a test which will sample both the objectives and subject matter in a representative manner. For this to be so, the weight in each cell should reflect the emphasis placed on each objective and topic during the teaching and learning. Unless such a test blue-print or table of specifications used as a guide in the test construction, there is the tendency to overload the test with items that cover only a limited content area or limited levels of objectives. The use of such device therefore, prevents the construction of tests that are biased. Hence it is used to build content validity into the tests.

Table of Specification (Test Blue-Print)

Contents	Objective Level						
	Knowledge 40%	Compr 25%	Application 20%	Analysis 5%	Synthesis 5%	Eval-uation 5%	Total 100%
Topic A 35%	8	5	4	1	1	1	20
Topic B 10%	2	2	1	0	0	0	5
Topic C 25%	6	4	3	1	1	1	16
Topic D 20%	5	3	2	1	1	1	13
Topic E 10%	3	2	1	0	0	0	6
Total 100%	24	16	11	3	3	3	60

Table 1.4

Table 1.4 shows that for the test being constructed, 5 content areas were covered and that each of the six objectives of the cognitive domain is being tested. The test has 60 questions. Having decided on the content objective levels and number of questions to set, the next step is to decide on how many of the 60 items will be assigned to each of the cells that relate objectives to content. This decision is taken by assigning weights according to emphasis or importance attached to each topic and objective can be determined through the analysis of the

relevant syllabus, curriculum and textbooks.

The weightings are for topic A 35%, topic B 10% etc. and for the objectives, the weightings are knowledge 40%, comprehension 25% etc. To work out the actual number of questions to be assigned to each cell, use the formula.

For example, for Topic "A" knowledge where 60 questions are to be constructed, the number will be $\frac{35}{100} \times \frac{40}{100} \times 60 = 8.4 \cong 8$

For that of comprehension is $\frac{35}{100} \times \frac{25}{100} \times 60 = 5.25 \cong 5$

With this, the number of questions for each cell is worked out and this serves as the guide for constructing the test items. After the development of the test blue-print, the next step is to write out the items based on the item writing.

4. Item Writing

The format may be essay-type, objective-type, multiple choice or short answer type. The following guidelines if adhered to will enhance the quality of the items.

- Much items should be constructed so that enough items will survive the item analysis
- The items should be clearly written so that the task is absolutely clear to the testee
- The test developer must not show any clue to the right answer
- Items that are too difficult or too easy must not be constructed
- Enough time should be given to complete the task to ensure reliability. The test developer must build in a good scoring guide and must adhere strictly to it.

If a good scoring guide is provided, then there is likelihood of having an objective assessment. After writing the items, they are sent for face validation.

5.Face Validation

Face validity, specifically refers to whether a test really "looks" valid to the examiner or any other person looking at it. So, to establish the face validity of an instrument is to subject that instrument to the scrutiny of relevant experts. The relevant experts here may range from Measurement and Evaluation experts to experts in the subject matter area. A representativesample of the testees for whom the instrument is being prepared should also be given the instrument for face validation.

6.Trial Testing

The items are then trial-tested. This is done by administering the test on an equivalent sample of the group for which the test is developed. The scripts are scored and then used for item analysis.

7.Item Analysis

This involves the analysis of individual itemsthat are in the test, after the test items have been subjected to statistical analysis, those that passed the analysis are selected for the final form of the test while those that failed are either discarded or modified and tried out again. Some new items may be constructed and retested for the actual analysis of the items during item analysis; this procedure is to compare the responses of learners in the upper one-third and the lower one-third continuum on the bases of total test score. The responses of the learners in the middle one-third are not included in the analysis. The responses that have been made to each test item by testees in the two references are tabulated. The effectiveness of each test item can then be determined by calculating the following indices from the learner's responses to the items, according to Nworgu (2015).

a) Item facility (difficult index or easiness index): It provides answer to the question. How hard is the item? It is represented by 'P' referred to as P-value.

$$P = \frac{U + L}{2N}$$

U = the number of testees in the upper one-third of the group who passed the item.
N = the total number of testees in either the upper or lower one third of the group who took the test. An ideal item will have P-value of 0.50 but realistically, it can range from 0.30 to 0.70.

b) Discrimination Index: This index answers the question; 'Does the item distinguish between the bright learners and dull learners?' This is the measure of the extent to which the item discriminates between the bright and the dull learners $D = \frac{U - L}{N}$

An ideal item should have a d-value of +1.00 but realistically, it could range from +0.30 to +1.00.

c) Distractor Index: It answers the question; 'Do all the options attract responses or are there some that are so unattractivethat they might as well not be included? $D.I = \frac{L - U}{N}$

In order to estimate the item indices, a hundred (100) item tests was given to 45 learners in a class. The lowest and highest scores were 90% and 20% arranged in an orderly form.
The upper one-third (1/3) of 45 should be 15 learners or papers, given to the upper group. The last 15 papers were selected to represent the lower group.
Let's use the 1st item in the test paper as an example.
Eg1. Change this number to standard form 175,000

a. 1.75×10^4 b. 17.5×10^5 c. 1.75×10^5 d. 175.000×10^4

Let us suppose further that the response pattern of the candidates in the two reference groups to this first item are as follows;

	Alternatives			**Options**	**Total**
	A	**B**	**C***	**D**	Total
Upper Group (U)	3	3	9	0	15
Lower Group (L)	5	6	4	0	15
Total	8	9	13	0	32

Table 1.5

The correct option to this item is B. From the above tabulation, we can calculate the item indices as follows;

Item facility (difficulty index)

$$= P = \frac{U + L}{2N+} = \frac{9 + 4}{2 \times 15} = \frac{13}{30} = 0.43 \therefore P = 0.43$$

The above analysis shows that the item is good in terms of difficulty index.

Discrimination Index for C.

$$d = \frac{U - L}{N} = \frac{9 - 4}{15} = \frac{5}{15} = 0.33$$

The above analysis shows that the item is also good in terms of discrimination index.

Distractor index (D.I) of option

$$A = \frac{L - U}{N} = \frac{5 - 3}{15} = \frac{2}{15} = 0.13$$

D.I. of option *B*

$$= \frac{6 - 3}{15} = 0.2$$

D.I. of option *D*

$$= \frac{0 - 0}{15} = 0.00$$

Options A & B have positive distractor indices and therefore are good. Option D has zero distractor index. This means that it did not appeal to any of the learners as a plausible option, hence no learner chose it. This shows that option D is a bad option because it did not follow the rule of standard form formula: *(a* x *x)*. After analysis, only items that scaled through will feature in the final tests and the bad ones will be discarded or reconstructed.

8. Item Selection

The items that have satisfactory statistical qualities are selected for inclusion on the final form of the test.

9. Test Assembly

Having selected the good items, the next step is to assemble the tests in the form it should be. Thus, some general rules to observe during test assembly are;

- Group items of the same type together
- Number all the items consecutively from the first to the last
- Arrange each sub-division of the test so that easier

onescome before the more difficult ones

- State the time and directions for answering the questions clearly.

10. Final Testing/Norming

Having selected and assembled the test items, the next thing is to go for final testing by administering the test on a fairly large sample of learners similar to those for whom the test is intended. This final testing will give an indication of the general performance of this group on the test. On the basis of the data collected from this group. Norms are established for the test. After this, the test can go for final production which will involve printing and production.

How to Prepare a Standard Marking Guide

Marking guide is a system for awarding points for correct answers or for proficiency in an examination or competition. Learners can view their exam papers to see how the marking guide was applied; it is a plan or guidelines used in the marking of learners' written workby instructors.

Whether you are marking exam answers or learners' assignments (for example posters, presentations and practical work), the time spent making a good marking guide can save you hours when it comes to marking a pile of scripts.

It can also help you to know (And show) that you are doing everything possible to be uniformly fair to all learners, as you may be required to show your model answers on some programs to people including external examiners and quality reviewers. It is important to design guides in the first place so that they will stand up to such scrutiny.

The following suggestions should help:

1. Write a model answer for each question if the subject matter permits. This can be a useful first step towards identifying the mark-bearing ingredients of a good answer. It also helps you to see when what you thought was going to be a 30 minutes' question turns out to take an hour if you have difficulties in answering the questions. The chances are that your learners will too be making model answers and marking guides for course work assignments.

2. Make each decision straight forward as possible. Try to allocate each mark so that it is associated with something that is either present or absent, or right or wrong in learners' answers.
3. Aim to make your marking guide usable by a non-expert inthe subject. This can help your marking guides be usefulresources for learners' themselves, perhaps in the next year's programs.
4. Aim to make it so that anyone can mark given answers, and agree on the scores within a mark or two.
5. Allow for consequential marks for example, when a candidate makes an early mistake, but then proceeds correctly thereafter (especially in problems and calculations, allow for some marks to be given for ensuring correct steps even when the final answer is quite wrong).
6. Pilot your marking guide by showing it to others. It's worth even showing marking guides to people who are not closely associated with your subject area. If they can't see exactly what you are looking for, it may be that the guide is not yet sufficiently self-explanatory. Extra detail you add at this stage may help you clarify your own thinking and will certainly assist fellow markers.
7. Look at what others have done in the past. If it is your first time writing a marking guide, looking at other people's ways of doing them will help you focus your efforts and offer you guidance and examples
8. Learn from your own mistakes. No marking guide is perfect. When you start applying it to a pile of scripts, you will soon start to adjust it. Keep a note of any difficulty you experienced in adhering to your guide and take account of this next time.

Marking is as important as planning and teaching. Marking is best seen as dialogue between you and the learners, knowing what to mark comes with experience. Marking is important if guides of learning are well planned. Try to mark learner's books properly by carefully preparing a good marking guide.The success of curriculum is hinged upon its effective implementation at classroom level and the teacher is central to this (Amadi, 2012). It is therefore, imperative to explore the understanding of developing standardized test by instructors for the purpose of achieving educational objectives

Conclusion

Developing standardized tests and preparing standard marking guide are critical process in conducting effective assessment. It promotes fairness, transparency, and reliability in evaluating learners' understanding. Tests developers must follow all the steps in this paper to construct their tests to make it to be standardized type. The tests items will base on intellectual activity ranging from simple recall of facts to problem solving, critical thinking and reasoning. A good test reflects the goals of the instructor. The instructor's true goal is to use the prepared standardized tests and marking guide to grade learner's responses consistently and objectively. It is the purpose of a test that determines the subject matter contents to be considered, the behavioral objectives to be measured, as well as the nature & difficulty level of the test items. Test construction process goes with the particular purpose for which the test is intended. It is therefore important that the test designers adhere strictly to a formal taxonomy specifying the objectives to be measured by a test. The author recommends that instructors should be attending conferences, seminars, In-service training/ workshop and counselling to improve their potentials.

References

Amadi, V.C. (2012). *Evaluation of Implementation of Continous Assessment in Secondary Schools Owerri*. Edition.zone I & II of Imo State: An Unpublished Master's Thesis.

Isaac, O.U. &Ibe, O. (2017), An Assessment of the application of Testing Skills Among Secondary School Teachers. In R.P.IUkwuije, P.Uzoma, & P.U Osadebe, (Eds). *Nigerian Paper of Educational Research and Evaluation*(pp. 145-149).A publication of the Association of Educational Researchers and Evaluators of Nigeria (ASSEREN).

Nworgu, B.G. (2015). Test Development Process. *Educational measurement and Evaluation Theory and Practice* (pp. 123-136). (2nd Edition). University of Nigeria Nsukka: University Trust Publishers Enugu.

Nikolopoulou, K. (2023). *What Is Content Validity? | Definition & Examples*. Scribbr. Retrieved October 6, 2023, from https://www.scribbr.com/methodology/content-validity/.