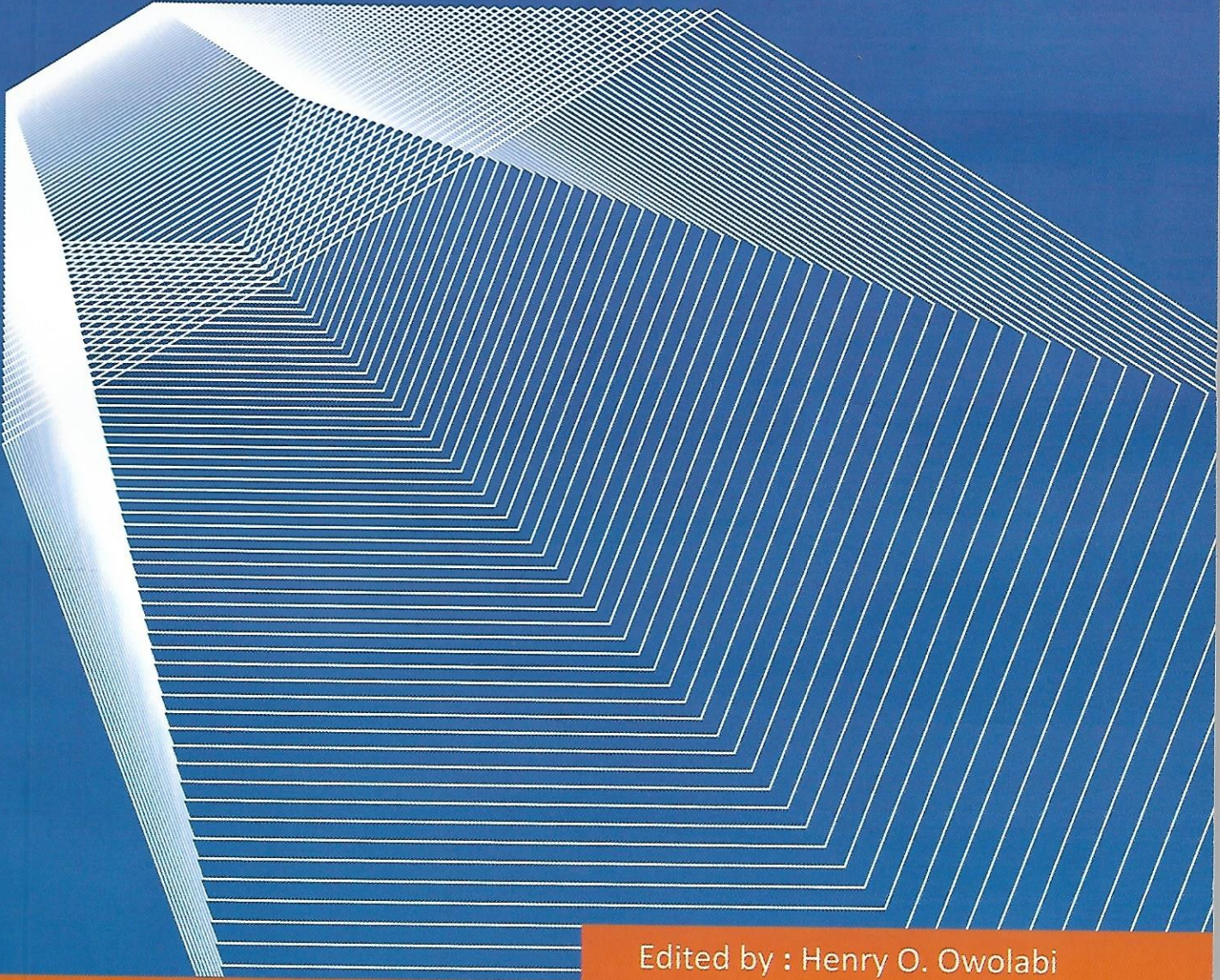


INTRODUCTION TO EDUCATIONAL MEASUREMENT AND EVALUATION

An abstract geometric design consisting of numerous white lines of varying lengths and orientations, creating a sense of depth and movement. The lines are arranged in a way that suggests a three-dimensional structure, possibly a series of nested planes or a complex crystalline form. The design is set against a solid blue background.

Edited by : Henry O. Owolabi

INTRODUCTION TO EDUCATIONAL MEASUREMENT & EVALUATION

**EDITED BY:
HENRY O. OWOLABI**

Copyright ©

Department of Social Sciences Education,
University of Ilorin.

January, 2019

All right reserved

No part of this publication may be reproduced or transmitted in any form or by means of electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without permission in writing from the publisher

ISBN: 978 - 978 - 972 - 067 - 5

Printed in Nigeria by
Wils & Bankleys Publisher (Nig) Ltd
15, Police Road, Ilorin

FOREWORD

A major concern of the 21st Century is how to develop the right kind of workforce to meet present day and future challenges. Education is globally recognized as the means by which those challenges could be met. Resting in the care of the system, structure and process of education are the tasks of assessment and evaluation. For teacher trainees, practicing teachers and all educators at all levels of education, the knowledge and skills of assessment and evaluation is indispensable. This empowers them to take good and appropriate decisions on educational issues and national development as a whole.

This book is written to provide the introductory knowledge and skills required by trainees and practitioners in the areas of educational measurement and evaluation. The book offers a simplified presentation of assessment techniques and measurement instruments such as test, questionnaire, observation, checklist, interview, rating scale and others. In addition, educators with knowledge of data interpretation can benefit from the simplicity of the presentation on the basic statistical concepts, data representation, measure of central tendency, measure of variability and measure of relationship. It is a good resource material for test development, administration and interpretation.

This book is an effort that provides useful key for the educators in the evaluation of educational achievement and related learning outcomes. It is a text that provides complete coverage of the basics of educational assessment. The text is written from the viewpoint that assessment is part of good teaching practice that helps educators improve students' learning. The contents of this book will provide the educators with a balanced and well structured knowledge and understanding of assessment practices.

Henry. O. Owolabi
Unilorin

TABLE OF CONTENTS

Basic Concepts in Educational Evaluation <i>Owolabi, Henry O.</i>	1
Purpose of Evaluation in Education <i>Daramola, Dorcas S.</i>	8
The Questionnaire <i>Bamidele, Esther. O.</i>	11
Observation Technique <i>Effrim, Paul. K.</i>	22
Checklist <i>Olutola, Adekunle. T.</i>	28
Interview <i>Mahmud, Jamiu. A.</i>	34
Rating Scale <i>Adewuni, David. A.</i>	38
Projective Techniques <i>Abdusalam-Nuhu, Rasheedat</i>	41
Inventories <i>Sheu Adaramaja .Lukman.</i>	46
Test Development Techniques <i>Ogunjimi, Mayowa. O.</i>	49
Concept of Test Validity <i>Ogunjimi, Mayowa. O.</i>	70
Test Reliability <i>Bamidele, Esther. O. & Oladele, Jumoke. I.</i>	75
Test Administration and Interpretation <i>Cobbinah, Andrews.</i>	92
Basic Statistical Concepts <i>Abdusalam-Nuhu, Rasheedat</i>	100
Descriptive Statistics in Education <i>Jimoh, Mohammed. I. & Akinboboye, Joseph. T.</i>	104
Measures of Variability/dispersion <i>Daramola, Dorcas. S.</i>	120
Correlation Statistics <i>Jimoh, Mohammed. I.</i>	133

TEST DEVELOPMENT TECHNIQUES

OGUNJIMI, M. O.

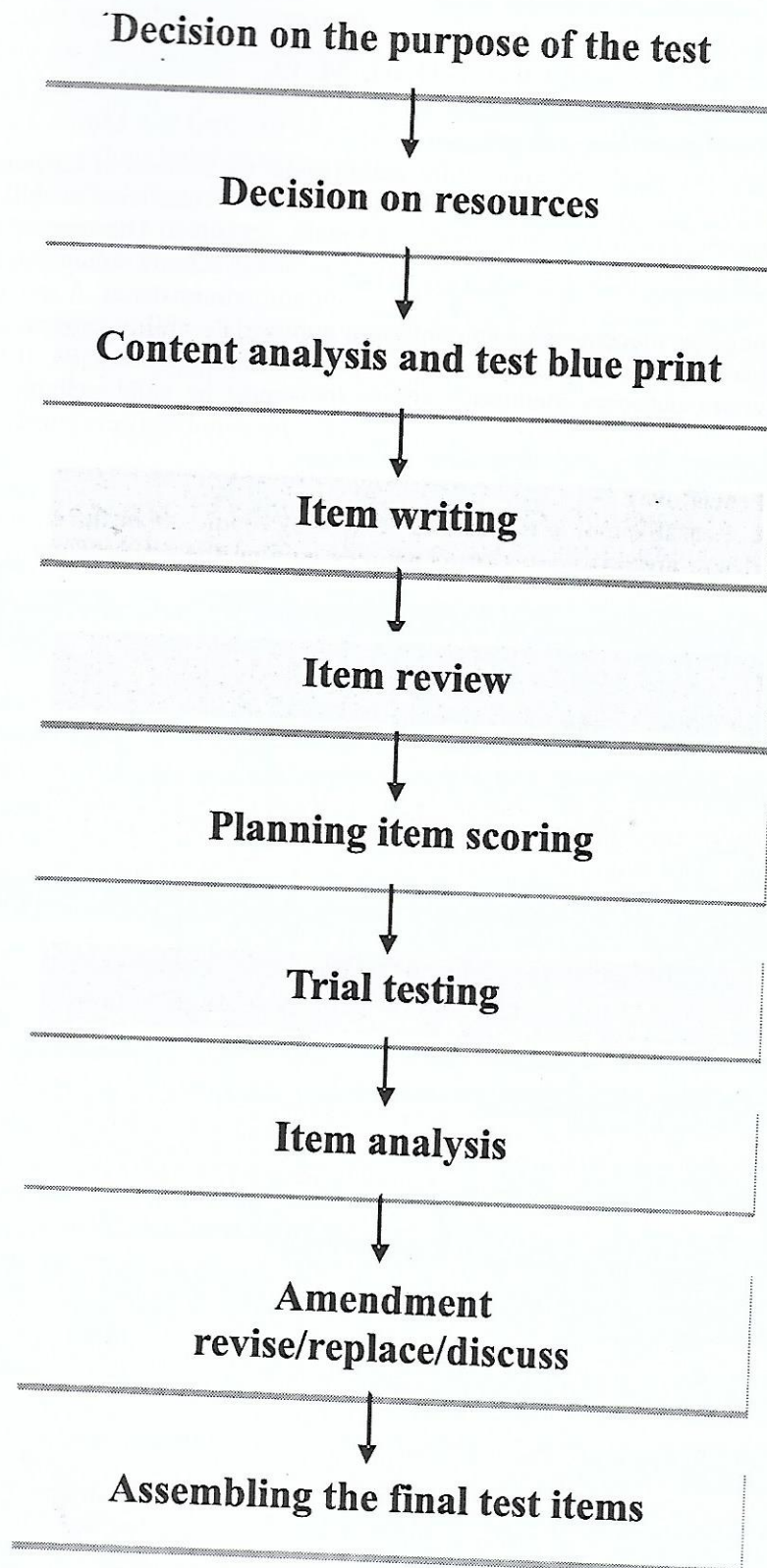
Introduction

Educational decisions which require information about the success of learning programme, or which require information about which students have reached particular level of skill and knowledge, depend upon valid measurement to inform those who make decisions. The type of information will depend upon whether the decisions are being made at the personal, school, regional or national level. In order to measure students' progress, tests are constructed and administered. A test is used to obtain information in guiding the placement of students into appropriate ability groups, to determine the retention level, evaluate the teaching effectiveness and to evaluate the curricula. If tests are to serve effectively the important purposes mentioned above, they must be valid, reliable and practicable. Consequently, to possess these qualities test must be carefully prepared and developed.

Test Construction Procedures

Tests are indispensable tool in evaluation; hence they should not be developed casually or in a haphazard manner. There are procedures or techniques guiding the construction of any test. These procedures are represented in the figure 1 below:

Figure 1: Steps in Test Construction



1. Decision on the purpose of the test

This entails determining the objectives of the test which may include curriculum evaluation, motivation of students, placement and selection, diagnosis of remedial work, formative and summative evaluation. Also, it is necessary to consider the likely actions that are to be taken according to the results of the test.

2. Decision on resources

It is also important to consider the resources needed in the test development and administration. These resources may include: the cost involve in providing human resources, word processing and computing facilities, materials and equipments for photocopying and printing.

3. Content analysis and test blue print

A content analysis provides a summary of the intentions of the curriculum expressed in term of the contents. The content is supposed to provide the learning experiences that will enable the test to achieve its stated objectives. This analysis of the content helps the test developers to determine the relative importance of the various aspects of the content and the emphasis to be placed on the specific content. Based on this, the table of specification or test blue print can be constructed.

A test blue print is a tabulation table of what the test is expected to cover rather than a description of what the curriculum covers. It is also called Table of Specification (TOS) which is a practical word given to the plan for scripting items for a test. Table of specification is a two-dimensional table relating instructional objectives and course content that specifies proportions of test items to be sampled. The table of specification enables a teacher to test the various concepts and gauge them over knowledge, skill and attitude. It provides the operational blueprint which guides the test builder and ensures that he builds a test that will do what it sets out to do.

Preparing a Table of Specification

The preparation of a table of specification requires the following procedures:

- a. Decide on the total number of items that will constitute the test
(e.g. 60 items)
- b. Decide on the proportion of items to be developed on each of the content areas, depending on the emphasis placed on it during teaching and the amount of time spent.

This is illustrated below:

Content

- Frequency Distribution	10%
- Measure of Central Tendency	15%
- Measure of variability	15%
- Measure of Relationship	40%
- Relative Standing	20%

- Decide on the proportion of items on each process objectives depending on the level of the cognitive behavioural objectives. The illustration is given below:

- Knowledge (recall of identification and information) 20%
 - Comprehension (focus on use of facts, rules, and principles) 30%
 - Application (applying facts or principles) 15%
 - Analysis (focus on separation of whole into components parts) 15%
 - Synthesis (focus on combining ideas to form a new whole) 10%
 - Evaluation (focus on developing opinions, judgments or decisions) 10%
- Determine the actual number of items to be constructed on each of the content area by finding out the respective percentage of the total number of items (illustration is shown in table I below).
- Determine the actual number of items to be written on each content area of the various dimensions of the cognitive behavioural objectives (illustration is also given in table I). It is good to ensure that the sum of the approximated numbers of items should be equal to the total number of the items desired on each of the content area. The respective result are presented in the table below:

Table1: Test Blue Print for A Test in Education Measurement.

Content Areas	Cognitive Behavioural Dimensions						Total
	Knowledge (20%)	Comprehension (30%)	Application (15%)	Analysis (15%)	Synthesis (10%)	Evaluation (10%)	
Frequency Distribution (10%)	1	1	1	1	1	1	6
Measure of Central Tendency (15%)	2	3	1	1	1	1	9
Measure of Variability (15%)	2	3	1	1	1	1	9
Measure of Relationship (40%)	5	7	4	4	2	2	24
Relative Standing (20%)	2	4	2	2	1	1	12
Total	12	18	9	9	6	6	60

The above table illustrates test blue print that will be used to develop an achievement test for units of study in educational measurement. This table provides a framework for organizing information about the instructional activities experienced by the students. The corner stone of classroom assessment practices is the validity of the judgments about students' learning and knowledge. Hence, table of specification is one tool that

teachers can use to support their professional judgment when creating or selecting test for use with their student.

1. **Item writing**

Once the test specification has been finalized, the next procedure is the construction of test items according to this specification. Item writing is the preparation of assessment task which can reveal the knowledge and skill of students when their responses to these tasks are inspected. Tasks which confuse, which do not engage the students or which offend, always obscure important evidence by either failing to gather appropriate information or by distracting the students from the intended task. Sound assessment tasks will be those which students want to tackle, those which make clear what is required of the students, and those which provide evidence of the intellectual capabilities of the students.

Item writing can be done by teachers for use within their own classroom (Teacher-made test) or by test specialists working with curriculum experts and examination bodies (Standardized test).

Teacher-made Test: These are test developed by teachers for a particular classroom use. Such test may not be used far and wide but are often designed to meet the particular learning needs of the students. The test is constructed by classroom teacher to measure the specific learning targets of the curriculum framework emphasizes.

Strengths of teachers made test includes, it:

- Provides teachers with the means to gather evidence about what their students have achieved.
- Helps instructors to identify student's strength and weakness.
- Keeps tabs on students learning and progress.
- Helps teachers plan and conduct future instruction.
- Guides students towards improving their own performances.
- Content may come from any area of the curriculum, items may be added, eliminated or modified as desired.
- Motivates and shape learning instructions.
- Construction procedures are very easy.

Weaknesses of Teachers Made Test

- Conclusions based on such tests are only limited to the members of a particular group and cannot be validly used for comparison with other groups.
- Scoring procedure can be biased and hence provide misleading information.
- Items may not be valid and reliable.
- Tests may not be used far and wide.
- Test items might be too difficult and may not have good discriminating power.
- The tests vary from teachers to teachers and from institution to institution.

- Teacher made test are not robust or rigorous enough to enable the comparison needed to be made objectively.
- No norms are provided, but may be developed by the teacher for his own class:
- The rules for administration and scoring are determined by the teachers.

Standardized Test: are constructed by the test specialists working with curriculum experts and teachers. They are constructed to have consistent questions, administration procedures and scoring procedures. This means all test takers are required to answer the same questions, or a selection of questions from common bank of questions in the same way, and that the test is scored in a standard or consistent manner, which makes it possible to compare the relative performance of individual students or groups of students. Linn and Gronlund (2000) note that standardized tests have the following qualities:

- Provide a systematic procedure for describing behaviours, whether in terms of numbers or categories.
- Include specified procedures for administration and scoring.
- Have an established format and set of materials.
- Present the same tasks and require the same response modes from all test takers.
- Provide tables of norms to which the scores of test takers can be compared in order to ascertain their relative standing.

Standardized tests are perceived as being fairer than teacher made tests because everyone get the same test and the same grading system.

Strengths of standardized tests

- Elimination of bias in assessment of individual testees.
- Provides data that can be aggregated, permitting comparisons of groups to a standard.
- Provide uniform test administration scoring and interpretation.
- Widely used for large-scale testing situations.
- Removes the potential for favoritism, bias and subjective evaluations.
- Provide valuable information that educators and school leaders can use to improve instructional quality.
- More reliable and valid than teacher made tests.
- Provides some type of "standard score" or norm which can help interpret how far an individual score ranges from the averages.

Weaknesses of Standardized test

- Items are fixed and are not modifiable.
- Tests can cause anxiety and encourage teachers to teach to the test instead of providing a rich and broad curriculum.

- It is more likely to become an object of debate and controversy when test scores are released and used to make consequential decision about educational policies.
- Test assumes a homogenous culture and a common curriculum.
- Test compares students rather than measuring the objectives of innovative instructional programmers.
- Repeated use may lead to loss of validity.

As earlier pointed out that item writing processes involve the preparation of assessment tasks by the classroom teacher or test specialists. The tasks expected from the students may require constructed or composed responses (Essay test) or tasks requiring provision of alternative responses and the candidate's task is to identify the correct alternative (Objective test). Thus, the two broad types of test items that can be developed are discussed below with specific precaution which must be taken in preparing each type of test item.

Essay Test: is a test that consist test items which requires a response composed by the examinee usually in the form of one or more sentences. No single response or pattern of responses can be listed as correct. The accuracy and quality of which can be judged subjectively only by one skilled or informed in the subject. This type of tests emphasizes freedom of expression and creativity as well as depth and scope of knowledge of the examinee in a given situation. Essay tests are of two types based on the extent of response expected:

- Extended response
- Restricted response

The extended response items are otherwise known as long essay items. They are items that require the examinee to supply long and comprehensive written answer in a manner that display his personal insights and the depth and scope of his knowledge of the subject matter. For example: Discuss the role of testing in educational decision making). The restricted response items require the examinee to supply written answers of restricted length or limited area. For example Discuss five roles of testing in educational decision making.

When to Use Essay Test

Essay tests are recommended for measuring educational achievement when:

- The group to be tested is small and the test items will not be reused with another group.
- The instructor wants to encourage the development of students' skill in written expression.
- The instructor is more interested in exploring students' attitude than in measuring achievements.
- The instructor is more confident of his or her proficiency as a critical reader than as an imaginative writer of good objective test items.
- Time available for test preparation is shorter than time available for the test scoring.

Guidelines for Writing Essay Test Items

In order to ensure that essay test items serve the purpose for which it is best suited, the following guidelines will help in the writing of essay test items:

- Clearly define the intended learning outcomes to be assessed.
- Ask questions or established tasks that will require the students to demonstrate command and skills of essential knowledge.
- Ask questions that are determinate, in the sense that colleagues in the field could agree that one answer is better than another.
- Define the examinee's tasks as completely and specifically as possible without interfering with the measurement process itself.
- Generally give preference to specific questions that can be answered briefly.
- Use enough items to sample the relevant content domain adequately but not so many that students do not have sufficient time to plan, develop, and also review their responses.
- Avoid giving examinees a choice among optional questions unless special circumstances make such options necessary.
- Divide a question into separate components when there are obvious multiple questions or pieces to the intended responses.

Strengths of Essay Test

- Provide effective way of assessing complex learning outcomes because it allows the ability to synthesis ideas and evaluate the worth of ideas.
- It allows students to demonstrate their reasoning.
- It provides authentic experience because constructed responses are closer to real life than selected responses.
- It is easier to construct than objectives test items because teachers don't have to create effective distracters.
- The used of essay questions eliminate the problems of guessing.
- Pose a more realistic task than multiple choice and other objective test items.
- Help find out how ideas are related to one another.
- It increases security because it is very difficult to plagiarize during an essay item examination.

Weaknesses of Essay Test

- It necessitates testing a limited sample of the subject matter, thereby reducing content validity.

- Scores may be influenced by factors extraneous to the content being tested (e.g. handwriting, spelling and grammar) and scorer's impression of the student, thereby having limitation of reliability.
- It requires more time for scoring student response.
- It emphasizes writing skills over content.
- It has low predictive validity because of low scoring reliability and limited content sampling.

Objective Test: is a test which requires task responses which can be scored accurately and fairly from the score key without having knowledge of the content of the test. This is test items that require the testee to either arrange concepts in a certain order, supply a missing word to an incomplete sentence, or select the correct answer from given alternatives. The objectivity of the test refers to the method of scoring and not necessarily to its item format. Test items are objectively scored to the extent that competent scorers or rater can agree on how responses should be scored.

Types of objective test items

- Alternative Response type
- Matching type
- Completion or Supply items
- Multiple-choice type

Alternative Response: This is the one that offers two options to choose from. It consists of a declarative statement that the examinee is asked to mark true or false, right or wrong, yes or no, agree or disagree. In each case, there are only two possible answers. The most common form it takes True/False questions. The test items is used in measuring the examinee's ability to identify the correctness of statements of fact, definition of terms, statements of principles, and also to distinguish fact from opinion. For example, Abuja is the Federal Capital of Nigeria. True () False ()

Strengths of Alternative Response Item

- True/False items have utility for measuring a broad range of verbal knowledge.
- A wide sampling of course content and materials can be obtained.
- Efficient method of testing a wide range of course content and material in a short period of time.

Weaknesses of Alternative Response Item

- Susceptibility to guessing. It has 50% guessing rate.
- Tendency for examinee to follow a certain pattern in responding to test items (Student response set).

Suggestions for writing Good matching items

The teacher/developer must:

- Provide clear directions and indicate the basis for matching the responses and premises.
- Keep the information in each column as homogeneous as possible.
- Include an unequal number of responses and premises.
- Keep the list of items to be matched brief and place the shorter responses on the right.
- Arrange the list of responses in logical order.
- Place all of the items for one matching exercise on the same page.
- Make all premises and options plural or singular to avoid giving unintentional clues.

Completion or Supply items: This requires the examinee to respond with a word, short phrase, number or symbol. It is suitable for assessing various abilities which include: simple recalling technology, facts, symbols and classifications. Completion test items require the students to supply answer rather than to recognize a correct alternative, but the responses tend to be quite short in comparison to essays.

Strengths of Completion or Supply items

- Relatively easy to construct and can be scored objectively.
- Lowers the probability of getting the answer correct by random guessing.
- Increase the level of item sampling.

Weaknesses of Completion or Supply items

- It is not suitable for measuring complex learning outcomes.
- It is difficult to score because many answers of varying degrees of correctness may be considered for total or partial credit.
- Spelling errors, grammatical errors, and legibility tend to complicate the scoring process further.

Suggestions for writing completion items

The following suggestions will help to avoid possible pitfalls and provide greater assurance that the items will function as intended:

- Word the item so that the required answer is both brief and specific.
- Do not include too many blanks.
- State the item so that only a single, brief answer is possible.
- Place the blanks at the end of the statement.
- Indefinite or open completion items should be avoided. That is, only

specific and important facts should be omitted.

- If the answer is to be expressed in numerical units, indicate the type of answer needed.

Multiple Choice Items: Is the one in which a student chooses one answer from a number of choices supplied. A multiple choice question consists of:

- A stem: This is the text of the question. It gives the problem scenario in one or more sentences.
- Options: The choices provided after the stem. It is also called alternatives or responses which are the possible solution to the problem presented within the stem.
- The key: Is the correct answer in the list of options.
- Distracters: Are the various plausible incorrect answers in the list of options. The distracters act to divert those examinees that are not sure of the correct answer. On the other hand, the distracters should not attract the attention of those examinees that are knowledgeable of the correct answer.

Multiple choice items are the most flexible of the objective type items. It provides the most useful format for measuring achievement at various level of learning. It is characterized by desirable difficulty level, ability to discriminate between high and low achievers and effective distracters.

Strengths of Multiple Choice Items

- Learning outcomes from simple to complex can be measured.
- Wide sampling of content of the curriculum.
- The problem of the task is well structured or clearly defined.
- Incorrect alternatives provide diagnostic information.
- Scoring is easy, objective and reliable.
- Flexible difficulty level.
- Provide scores easily understood and transformed as needed.
- Item analysis can reveal how difficult each item is and how well it discriminates between strong and weaker students.
- A broad sample of achievement or performance can be measured.
- Highly structured and clear tasks are provided.

Weaknesses of Multiple Choice Items

- Constructing good items is time consuming.

- It is frequently difficult to find plausible distracters.
- Susceptible to guessing.
- Item is ineffective for measuring some types of problem solving and the ability to organize and express ideas.
- Scores can be influenced by reading ability.
- It is difficult to determine why individual students select incorrect responses.
- Students can sometimes read more into the question than was intended.
- Sometimes there is more than one defensible correct answer.
- Does not provide a measure of writing ability.

Multiple-Choice Item Writing Guidelines

There are several essential guiding principles that should be adhered to during the process of writing multiple-choice items. Haladyna and Downing (1989) present a detailed list of rules for constructing good quality multiple-choice items. These are highlighted below:

Procedural rules for constructing Multiple-Choice Item

- Use either the best answer or the correct answer format.
- Format the item vertically, not horizontally.
- Allow time for editing and other type of item revisions.
- Use good grammar, punctuation and spelling consistently.
- Minimize the time required to read each item.
- Avoid tricky items, those which will mislead or deceive examinees into answering incorrectly.
- Avoid giving unintended cues-such as making the correct answer longer in length than the distracters.

Content-related rules for constructing Multiple-Choice Item

- Based each item on an educational or instructional objective.
- Test for important or significant information.
- Focus on a single problem.
- Keep the vocabulary consistent with the examinees' level of understanding.
- Avoid cueing one item with another: keep items independent of one another.
- Avoid copying textbook verbatim, phrasing when developing the items.
- Avoid items based on opinions.
- Be sensitive to cultural and gender issues.

Stem construction rules for Multiple-Choice Item:

- Present a single, definite statement to be completed or answered by one of the given choices.
- When using the completion form, don't leave a blank for completion in the beginning or middle of the stem.
- Ensure that directions in the stem are clear, and that wording lets the examinee know exactly what is being asked.
- Avoid excessive verbiage in the stem.
- Word the stem positively, avoid negative phrasing such as not or except. If this cannot be avoided, the negative words should be highlighted by underlining or capitalization.
- Avoid giving clues such as linking the stem to the answer.

General option development rules for Multiple-Choice Item

- Place options in logical or numerical order.
- Keep options independent, options should not be overlapping.
- Keep all options in an item homogeneous in content.
- Keep the length of options fairly consistent.
- Avoid or use sparingly, the phrase "all of the above" or "none of the above".
- Phrase options positively, not negatively.
- Avoid distracters that can clue test-wise examinees; for example, avoid clang associations, absurd options, formal prompts, or semantic (overly specific or overly general) clues.
- Avoid specific determinates, such as never and always.

Key Option Development Rule

- Position the correct option so that it appears about the same number of times in each possible position for a set of items.
- Ensure that there is one and only one correct option.

Distracter Development Rules:

- Use plausible distracters, avoid illogical distracters.
- Incorporate common errors of students in distracters.
- Avoid technically phrased distracters.
- Use familiar yet incorrect phrases as distracters.
- Use true statements that do not correctly answer the item.
- Avoid the use of humor when developing options.

Oral test: is a practice where an examiner poses questions to an examinee in spoken form. It involves two persons, examiner and examinee, face to face. The examiner asks questions, the examinee has to answer the question in such a way as to demonstrate sufficient knowledge of the subject in order to pass the test. Sometimes to improve the objectivity of the test, more than one examiner is present, or to improve the efficiency of the process and fairness of the judgments, several examinees might be interviewed simultaneously, giving each a chance to respond to the same questions and to comment on answers given by the other examinees.

Strengths of Oral Test

- Involve direct contact and interaction between examiner and examinee.
- It is less threatening, more personal and more humane than written test.
- Personal characteristics that would be impossible to assess on a written test can be evaluated in the face-to-face situation, such as candidate's appearance, manner, alertness, stress tolerance, speech pattern and quality.
- It permits flexibility in assessing examinee's knowledge and intellectual abilities than written test.
- Minimize the problem of cheating

Weaknesses of Oral Test

- It is easy for oral test enthusiasts to claim too much.
- Does not measure writing skills.
- It is essentially an individual process, hence it is time consuming.
- It opens the door to prejudice, partiality and discrimination on grounds other than the relevant traits and abilities.
- Low test validity as a result of stress of the confrontation.
- Difficulty in obtaining reasonable reliable scores in reasonable amount of time.

Practical Test: are nonverbal test that require examinees to perform a task rather than answer questions. They measure application of learning experiences in actual performance. The tests are designed to measure some motor or manual responses of an examinee that is expected to manipulate some concrete equipments or materials. The tests are interested in measuring the degree of skills an examinee exhibits in applying certain knowledge. They include practical tasks in trade subjects (such as woodwork, metalwork, leather craft e.t.c), in musical and dramatic performance, in skills such as swimming or gymnastics, or may refer to the skills required to carry out laboratory or field tasks in science, agriculture, geography, environmental health or physical education.

General suggestions in writing test items

- Begin writing items well ahead of the time when they will be used, allow time for revision.
- Match items to intended outcomes at the proper difficulty level to provide a valid measure of the instructional objectives.
- Be sure each item deals with an important aspect of the content area and not with trivia.
- Be sure that the problem posed is clear and unambiguous.
- Be sure that each item is independent of all other items.
- Be sure the item has one correct or best answer on which experts would agree.
- Prevent unintended clues to the answer in the statement or question.
- On a test with different question formats, one should group all items of similar format together.
- Questions should follow an easy to difficult progression.
- Space the items to eliminate overcrowding.
- Have diagrams and tables above the item using the information not below.

2. Item Review

Writing assessment tasks for use in tests requires skill. Sometimes the item seems clear to the person who wrote it but may not necessarily be clear to others. Before empirical trial, test items need to be review by other experts in the field to spot any ambiguities, grammatical faults and language difficulties, unintended clues to the correct answer, redundant words and distracters which are not plausible. The review should also identify items with no correct answer and items with multiple correct answers.

3. Planning Item Scoring

Arrangement should be made on how the test items should be scored. Concerning objective test items where the candidate's task is to provide or identify the correct answers, judgments of experts are needed to establish which option or which response is the correct answer for each item. Once these correct answers have been decided, the score key can then be used. The correct score key is crucial because errors in score keys will create interpretation problems. In such a case the total score obtained will not be the best measure of what the test is measuring.

On the other hand, there are potential difficulties in scoring prose, oral and manipulative responses. An expert judge is required because each response will require interpretation to be scored. The scorers may vary in their experience, vary over time in the way they score responses (due to fatigue or by giving varying credit for some correct responses over other correct responses), and also vary in the notice they take of handwriting, neatness, grammatical usage and spelling. In order to avoid or minimize such problems, a team of scorers could be trained. Such training will involve a discussion of the key issues that have to be identified by a candidate. The scorers can then apply what they

have learned by scoring the real samples of responses. The aim is to improve the agreement between scorers about the quality of each response.

4. Trial Testing

Item trial is sometimes called pilot testing. It is a means of subjecting the proposed test items to empirical trial with students similar to those who are going to use the final form of the test. Since items involve communication with students, an evaluation of this quality is required before the set of tasks can be used with a larger group. Conducting trial testing requires sound planning. The numbers of trial candidates, the balance between males and females, the diversity of age levels or schooling levels required for the trials, and the availability of test supervisors are all issues that need to be considered.

5. Item Analysis

Item analysis is a general term that encompasses a variety of methods for summarizing and analyzing the responses of students to test items. It is a systematic evaluation of the effectiveness of each item in a test. Certain patterns of responses can indicate desirable and undesirable features of the item or of the scoring procedures employed. Often these methods suggest why an item has not functioned effectively and how it might be improved. The item analysis is aimed at determining:

- Item difficulty
- Item discrimination power
- Effectiveness of the distracters in multiple choice items.

Procedures for Item Analysis: The following steps outline a simple procedure for analyzing 60 test items that was administered to forty-eight students:

- (i.) Arrange the test papers of the students (all 48 papers) in order of their scores from the highest to the lowest score.
- (ii) Select approximately one-third of the paper with the highest scores which should be identified as the upper group (16 papers). Select the same number of papers with the lowest scores which should be identified as the lower group (16 papers). Set the middle group of papers aside. The papers of the upper group (16 papers) and the lower group are to be used for the analysis.
- (iii) For each item, count the number of students in the upper groups who selected each alternative. Make the same count for the lower group.
- (iv) Record the count from step (iii) above, on the item card or on a separate sheet, tabulating the number of students choosing each alternative in the upper and lower groups as follows:

		Options					
		A	B	C	*D	E	TOTAL
Item 1	Upper Group	1	1	0	14	0	16
	Lower Group	3	4	2	5	2	16

*Key = D (correct option)

(v) Calculate the difficult index for each item.

(vi) Calculate the discrimination index for each item.

(vii) Use the result of step (iv), check each item to identify weak distracters, ambiguous alternatives, miskeying and indicators of random guessing.

Item Difficulty: can be referred to as the proportion of students responding correctly to an item (i.e. the percentage of students that correctly answered the item). It is an indication of the extent to which an item is difficult for examinees. The range is from 0% to 100% or more typically written as a proportion as 0.00 to 1.00. The higher the value, the easier the item. In order to establish difficulty index of an item, the number of students in the upper group who got the item correct is added to the number of students in the lower group who got the item correct (using step (iv) as example: $14+5=19$). The sum is then converted to percentage (e.g. $\frac{19}{32} \times \frac{100}{1} = 59\%$).

The index of item difficulty can be established using the following formula:

$$P = \frac{CR_U + CR_L}{N_U + N_L}$$

Where P = Item difficulty index

CR_U = No of persons in the upper group who got the item correct.

CR_L = No of persons in the lower group who got the item correct.

N_U = No of persons in the upper group.

N_L = No of persons in the lower group.

Therefore, the index of item difficulty for item 1 above is 0.59 (59%). The item could be adjudged as moderately difficulty. If the index falls 0.20, the item could be regarded as very difficult and should be reviewed for possible confusing language or highlighted for an area for re-instruction. Assuming the difficulty index (P-value) is above 0.90, this signifies that the item is very easy and hence, should not be reused again for subsequent test. Consequently, the lower the index, the more difficult the item. A good test should consist of items having a difficulty index ranging from 0.20 to 0.90. Items outside this range need to be reconstructed or discarded.

Item Discrimination: Discrimination may be thought of as the degree to which students with varying levels of achievement perform differently on a question. That is, do high achieving students usually produce better responses than lower achieving students? If so, we say the item is functioning properly in differentiating between high and low achieving students. Hence, item discrimination refers to the percentage difference in correct responses between the higher achievers and the lower achievers. The item discrimination index is used to indicate the extent to which response to an item could distinguish between the strong and weak students. It measures the extent to which items are capable of measuring individual differences. It ranges from -1.00 to 1.00. The higher the value, the more discriminating the item. A highly discriminating item indicates that the students who had high test scores got the item correct whereas students, who had low test score, got the item incorrect.

The discriminating index of an item (D) is established by subtracting the number of student in the lower group who got the item correct from the number of those in the upper group who also got the item correct, and dividing this difference by half the total number of student in

the two groups. The formula is presented thus:

$$D = \frac{CR_u - CR_l}{\frac{1}{2}T}$$

Where: CR_u = Number of students in the upper group who got the item correct.

CR_l = Number of students in the lower group who got the item correct.

T = Total number of students in the upper and lower groups.

For the example of item 1 given above, that 14 out of 16 students in the upper group got the item correct, and 5 out of 16 students in the lower group got the item right, the index of discrimination for the item is determined as follow:

$$D = \frac{14 - 5}{\frac{1}{2}(16 + 16)} = \frac{9}{16} = 0.56$$

Since the index ranges from -1.00 to 1.00, hence the positive index of 0.59 obtained in this example is moderate and satisfactory. Any negative index signifies that students in the lower group did better on the item than students in the upper group, hence the item should be removed from the test. A positive index which falls below 0.20 would be regarded as possessing a low discriminating power. One major reason for the desirability of high, positive discrimination index is that reliability increases as the average value of discrimination increases.

Effectiveness of distracters: This relate specifically to multiple choice items. Normally, all the options in each multiple choice item should be chosen by at least one person in the upper or lower group, and each distracter should be chosen more frequently by those in the lower group than in the upper group, but not with the case of correct option. Reviewing the options can reveal potential errors of distracters. The poor distracters can be revised, replaced or removed. Example is presented below for item 2:

		Options					
		*A	B	C	D	E	TOTAL
ITEM 2	Upper group	11	0	1	3	1	16
	Lower group	7	0	3	2	4	16

*Key = A (correct option)

In the above examples, option A, C, D, and E are functional because each of them was chosen by at least one person in the upper or lower group, except for option B which is completely not functional since it attracted no one, hence, should be replaced. However, while options C and E can be regarded as effective distracters because they discriminate in the right direction, since they were chosen by more persons in the lower group than in the upper group, option D on the other hand is ineffective because it was chosen by more people in the upper group than in the lower group. Option D should therefore be inspected and reviewed with a view to remove any ambiguities or other defects. Alternatives are ambiguous if higher group students are unable to distinguish between the correct answer and one or more of the incorrect alternatives. The alternatives should also be checked for inadvertent miskeying if a large number of upper group student choose a particular wrong response. You should suspect that students are engaging in considerable random guessing if an approximately equal number of upper group students are

choosing each alternative. An analysis such as this is useful in evaluating a test item, and when combined with an inspection of the item itself, it provides helpful information for improving the item.

1. Amending the Test

On the basis of the result of item analysis, item which do not perform as expected can be discarded or revised. However, discarding the items where there is a shortage of replacement items can lead to distortions of the achieved test specification. Thus, test writers may prepare more items that they think will be needed for each cell in the grid.

2. Assembling the Final Test Items

After amending those that need to be amended and discarded the unsuitable items, the test writer compiles the selected items for final use. It is important to consider the following principles when compiling the test items:

- Arrange the items in order of difficulty. Usually the easiest questions are presented first to ensure that weaker candidates are not discouraged.
- Arrange the items on the basis of format or type, so that it will be possible to write directions relating to each item type.
- Ensure that the position of the correct options in multiple choice items does not follow a definite pattern but randomized.
- The final test should be consistent with the test blueprint.
- The final test should have an attractive layout and suitable spacing of the items to facilitate easy reading and responding.
- The test instructions must be prepared. It must be cleared and unambiguous to facilitate test administration and promote test validity. Time allowed for candidate to complete test items should also be indicated.
- Prepare the scoring guide for each item in the test.

Preparing the final test is not the end of the work. The data from the use of the final test should be monitored as a quality control check on their performance. Such analysis can also be used to fix a standard by which the performance of future candidate may be compared, or may be used to detect whether there has been a breach of test security.

References

- Abiri, J .O. O.(2006). *Elements of Evaluation Measurement and Statistical Techniques in Education*. Ilorin: Unilorin Press.
- Atjaz, A.,Aziz, N. &Tayyaba, Z.(2013). Table of Specification Development and Uneager: An Overview. *Interdisciplinary Journal of Contemporary Research in Business*, 4(12), 354- 361.
- Haladyna, T.M. &Downing, S.M. (1989). A Taxonomy of multiple –choice item - writing rules. *Applied Measurement in Education*. 2(1), 37-50.
- Izard, J. (2005). Overview of test construction.In K.N. Ross(Eds), *Quantitative research methods in educational planning*. Paris: UNESCO/International Institute for Educational Planning.
- Linn, R.L. &Gronlund, N.E. (2000). *Measurement and assessment in teaching*. Columbus, OH: Merrill.
- Ntiko, A.J. &Brookhart, S.M. (2007). *Educational assessment of students*. (5th ed.)Upper Saddle River, New Jersey: Pearson Education Inc.
- Ogunjimi, M. O. (2009). *Essential of educational test, measurement and evaluation*. Ibadan: Joytal Publications.



W&B
Wils & Bankleys
Publishers(Nig)Ltd



978-978-972-067-5