

Development and Validation of a Mathematics Cognitive Achievement Test for Junior Secondary School Students in Akwa Ibom State, Nigeria

Glory Udoh Evans, Ph.D

Department of Educational Foundations,
Department of Educational Foundations, Akwa Ibom State University of Education
&

Ofonime Friday Udoh, Ph.D

Department of Mathematics
Department of Mathematics, Akwa Ibom State University of Education

Abstract

This study developed and validated a Mathematics Cognitive Achievement Test (MCAT) for Junior Secondary School students in Akwa Ibom State, Nigeria. An instrumentation research design was adopted, using 500 students for item analysis and validation. An initial pool of 70 multiple-choice items was developed from the Junior Secondary School Mathematics curriculum and reviewed by experts in Mathematics Education, Measurement and Evaluation, and Educational Psychology. The items were administered to the sample and evaluated for difficulty, discrimination, distractor effectiveness, content validity and internal consistency. After expert review and empirical analysis, 50 items were retained. The retained items had difficulty indices ranging from 0.20 to 0.80 and discrimination indices from 0.30 to 0.69. Thirty-eight items (76.0%) had three functional distractors, nine had two, and three had one. The overall KR-20 reliability coefficient was 0.89, while the overall expert rating for content validity was 4.54 on a five-point scale. The findings indicate that the final MCAT has satisfactory psychometric characteristics and can be used, with appropriate administration and further local validation, for classroom, diagnostic and research purposes.

Keywords: Achievement test; Cognitive achievement; Mathematics; Test development;

Introduction

Mathematics is central to Nigerian education because it supports scientific reasoning, technological development, economic decision-making and everyday problem solving. At the Junior Secondary School level, it provides foundations in number, algebra, geometry, measurement, statistics and logical reasoning. The Nigerian Educational Research and Development Council (NERDC) curriculum emphasizes mathematical literacy, conceptual understanding, application and problem solving. Consequently, assessment at this level should represent both important content and the cognitive processes through which students demonstrate learning.

Achievement tests provide evidence of what learners have acquired following instruction. Their usefulness, however, depends on the quality of the instrument and the evidence supporting interpretations of scores. The Standards for Educational and Psychological Testing emphasize that test developers should provide evidence appropriate to

intended score interpretations and uses (AERA, APA, & NCME, 2014). Bloom's revised taxonomy similarly identifies remembering, understanding, applying, analysing, evaluating and creating as relevant cognitive processes (Anderson & Krathwohl, 2001).

A persistent assessment concern is the use of teacher-made tests that are not systematically developed from test blueprints or subjected to psychometric evaluation. Such tests may contain items that are too easy or too difficult, discriminate poorly, or have ineffective distractors. Previous work has emphasized the value of systematic item analysis in mathematics achievement testing (Eluwa *et al.*, 2012) and the need to strengthen teachers' testing skills (Evans & Dan, 2025). Validity concerns the extent to which evidence and theory support intended score interpretations, whereas reliability concerns the consistency or precision of scores (AERA *et al.*, 2014).

The present study therefore addressed the need for a curriculum-based mathematics achievement test with empirical evidence of content validity, item quality and internal consistency for Junior Secondary School students in Akwa Ibom State.

Statement of the Problem

Meaningful assessment of mathematics achievement requires instruments aligned with curriculum objectives and supported by adequate psychometric evidence. In many school settings, classroom tests are constructed primarily for immediate examination purposes and may not undergo systematic item writing, expert validation, pilot testing, item analysis and reliability estimation. Poorly functioning items can distort estimates of achievement, reduce discrimination among students and weaken decisions based on test scores. Similarly, poorly worded or ambiguous items may introduce construct-irrelevant variance into students' scores. AERA, APA and NCME (2014).

At the Junior Secondary School level, such weaknesses are consequential because students are building mathematical foundations for later education. Akwa Ibom State therefore requires locally relevant, curriculum-based instruments that can provide dependable evidence of students' cognitive achievement. This study developed and validated a 50-item MCAT through expert review and empirical item analysis.

Purpose of the Study

The purpose was to develop and validate a Mathematics Cognitive Achievement Test for Junior Secondary School students in Akwa Ibom State. Specifically, the study sought to: (1) develop curriculum-based mathematics test items; (2) establish content validity through expert judgment; (3) determine item difficulty index; (4) determine item discrimination index; (5) examine distractor effectiveness; (6) estimate internal consistency reliability; and (7) produce a final 50-item test.

1. determine the reliability coefficient of the developed Mathematics Cognitive Achievement Test; and

Research Questions

The following research questions guided the study:

1. What proportion of the initial developed items meets the criteria for inclusion in the final Mathematics Cognitive Achievement Test?

2. What are the difficulty indices of the retained items?
3. What are the discrimination indices of the retained items?
4. How effective are the distractors of the retained multiple-choice items?
5. What is the internal consistency reliability coefficient of the MCAT?
6. What evidence of content validity supports the developed MCAT

Research Hypotheses

H_{0 1} : There is no significant difference between the mean scores of high- and low-performing students on the retained MCAT items.

H_{0 2} : The MCAT does not demonstrate satisfactory internal consistency reliability for educational measurement purposes.

H_{0 3} : The retained MCAT items do not adequately discriminate between students of relatively high and low mathematics achievement.

Literature Review

Achievement Testing and Mathematics Cognitive Achievement

An achievement test is an assessment instrument designed to obtain evidence concerning the knowledge, skills or competencies that learners have acquired following instruction. Achievement testing differs from aptitude testing because achievement tests are generally concerned with learning that has already occurred within a defined instructional domain (Nitko & Brookhart, 2014). In school mathematics, achievement tests are particularly useful because mathematical learning involves interconnected knowledge and skills. A student may be required to understand a mathematical concept, apply an algorithm, interpret a representation and solve a real-life problem. Consequently, a good mathematics achievement test should represent the important content and cognitive processes included in the curriculum.

Cognitive achievement in mathematics refers to the extent to which a learner demonstrates knowledge, understanding, application, reasoning and problem-solving ability in mathematics. Anderson and Krathwohl (2001) explain that cognitive learning can be organized around different processes, including remembering, understanding, applying, analysing, evaluating and creating. For Junior Secondary School students, cognitive achievement can be assessed through questions involving mathematical facts, procedures, relationships, calculations and problem situations. A balanced mathematics achievement test should not rely exclusively on recall questions. Items should also assess the application and interpretation of mathematical principles.

Test Development and Validation

Systematic test development involves defining the purpose and content domain, constructing test specifications, writing items, expert review, pilot administration, item analysis, reliability estimation, revision and preparation of the final instrument (Crocker & Algina, 2008; Nitko & Brookhart, 2014). A table of specifications helps ensure adequate representation of content and cognitive objectives. For curriculum-based tests, content

evidence is particularly important because items should correspond to curriculum expectations and intended cognitive demands (AERA *et al.*, 2014).

Validity is concerned with the degree to which evidence supports the intended interpretation of test scores for a specified purpose. Contemporary measurement theory emphasizes that validity is not simply a property that a test either possesses or lacks; rather, evidence must be accumulated to support particular interpretations and uses of scores (AERA *et al.*, 2014). Evidence based on test content is particularly important for a curriculum-based achievement test. Content evidence involves determining whether the items adequately represent the relevant content domain and cognitive processes. Expert review is therefore an important component of the development of an achievement test. For this study, experts were expected to examine the items for relevance, clarity, appropriateness of difficulty, correspondence with curriculum objectives, absence of ambiguity and suitability for Junior Secondary School students.

Item Difficulty Index, Discrimination and Distractor

Item difficulty is the proportion of examinees answering an item correctly; higher values indicate easier items. In this study, values below 0.20 or above 0.80 were treated as candidates for revision or rejection. It is commonly expressed as:

$$P = \frac{R}{N} \quad (1)$$

where: P = item difficulty index; R = number of students who answered the item correctly; and N = total number of students responding to the item.

Item discrimination indicates how well an item differentiates higher- from lower-performing examinees. Positive discrimination is desirable, while negative values can signal problems such as ambiguous wording, a defective key or content misalignment. A good achievement-test item should generally be answered correctly more frequently by high-performing students than by low-performing students (Evans *et al.*, 2022). The discrimination index may be expressed as:

$$D = \frac{RH - RL}{N/2} \quad (2)$$

where (RH) represents the number of students in the high group who answered correctly, (RL) represents the number in the low group who answered correctly, and (N/2) represents the size of one comparison group.

Items distractor analysis is essential in revealing whether alternative options are functioning as intended. Therefore, an ineffective distractor is one that attracts very few or no students. Distractors should attract students who have not mastered the content; non-functional alternatives warrant revision (Haladyna *et al.*, 2002).

Reliability of Achievement Tests

Reliability concerns the consistency or precision of measurement. For a dichotomously scored multiple-choice test, KR-20 is an appropriate estimate of internal consistency. Although acceptable reliability depends on the intended use, coefficients around .70 or higher are commonly used as a practical benchmark in educational research (Nunnally & Bernstein, 1994). Reliability alone does not establish validity; it must be considered with content representation and item functioning.

Methodology

The study adopted an instrumentation research design because its purpose was to construct, validate and establish the psychometric characteristics of the MCAT. Instrumentation research involves systematic procedures for constructing assessment instruments and establishing evidence of their validity, reliability and quality of individual test items (Crocker & Algina, 2008; DeVellis & Thorpe, 2021). The study was conducted in Akwa Ibom State, South-South Nigeria, and covered selected public Junior Secondary Schools across the three senatorial districts: Akwa Ibom North-East, North-West and South. The state has a well-established basic education system comprising public and private primary and secondary schools. Junior Secondary School education constitutes an important component of the state's basic education system, where Mathematics is a core subject. Akwa Ibom State was selected as the study area because it is been classified as educational advantage state, and Mathematics is one of the compulsory subjects at the Secondary School level of education.

Population, Sample and Sampling Technique

The population comprised Junior Secondary School students in public secondary schools in Akwa Ibom State whose curriculum content was represented in the test specifications. A sample of 500 was selected from 12,430 students through multistage sampling procedure. At the first stage, the state was stratified into the three senatorial districts: Akwa Ibom North-East, Akwa Ibom North-West and Akwa Ibom South. At the second stage, selected Local Government Areas were chosen from the senatorial districts using simple random sampling. At the third stage, public Junior Secondary Schools were selected from the sampled Local Government Areas. At the fourth stage, students were selected from the participating schools using proportionate and simple random sampling techniques.

Instrument for Data Collection

The MCAT initially contained 70 four-option multiple-choice items covering nine content areas: approximation; area of plane shapes; whole number arithmetic; variation; inequalities; LCM and HCF; statistics and probability; solving equations; and word problems. A table of specifications distributed the 50 final items across four cognitive levels:

knowledge (26%), comprehension (20%), application (42%) and evaluation (16%). Experts reviewed the items for curriculum relevance, content coverage, clarity, cognitive level, mathematical accuracy and quality of alternatives. The 70-item pool was reduced to 65 items after expert review and then to 50 after empirical analysis. Each correct response attracted one mark and each incorrect response zero, giving a maximum score of 50.

Data Analysis

Item difficulty and discrimination were computed using classical test theory procedures. Distractors were examined for functionality. Content validity was summarized using expert mean ratings on a five-point scale. KR-20 was used to estimate internal consistency. For the first hypothesis, upper and lower 27% groups (135 students in each group) were compared using an independent-samples t-test. Statistical decisions were based on the .05 significance level.

Results

Item Analysis and Results.

Research Question1: What proportion of the initial mathematics test items meets the established criteria for inclusion in the final Mathematics Cognitive Achievement Test?

Table 1: Item Screening and Selection

Stage of test development	Number of items	Percentage
Initial items constructed	70	100.0
Items retained after expert validation	65	92.9
Items rejected after expert validation	5	7.1
Items subjected to empirical item analysis	65	92.9
Items rejected after item analysis	15	21.4*
Final items retained	50	71.4
Total	70	100.0

*Note: The percentage is based on the original 70-item pool.

Table 1 shows that 70 items were initially constructed. Following expert validation, 65 items were retained for empirical testing, while five items were rejected because of problems relating to content relevance, ambiguity, duplication or inappropriate cognitive demand. After administration to the 500 respondents and subsequent item analysis, 15 additional items were eliminated. Thus, 50 items, representing 71.4% of the original item pool, were retained for the final Mathematics Cognitive Achievement Test.

Table 2: Table of Specifications for the 50-Item Mathematics Cognitive Achievement Test

Content Area	Knowledge 26%	Comprehension 20%	Application 42%	Evaluation 16%	Total
Approximation	2	1	1	1	5
Area of Plane Shapes	1	1	3	1	6
Whole Number Arithmetic	2	1	2	1	6
Variation	1	1	3	1	6
Inequalities	1	1	2	1	5
LCM and HCF	2	1	2	1	6
Statistics and Probability	2	2	3	1	8
Solving Equations	1	1	3	1	6
Word Problems	1	1	2	0	4
Total	13	10	21	8	50

Table 2 presents the final version of the 50 items instrument contained after expert review and empirical item analysis. Each correctly answered item attracted one mark, while an incorrect answer attracted zero. Therefore, the maximum obtainable score on the final MCAT was 50. Finally, the reliability of the 50-item instrument was estimated was calculated using KR-20 coefficient to estimate the internal consistency.

Research Question 2: What are the difficulty indices of the retained Mathematics Cognitive Achievement Test items?

Table 3: Distribution of the 50 Retained Items According to Difficulty Index

Difficulty index	Interpretation	Number of items	Percentage
0.20–0.29	Difficult	4	8.0
0.30–0.39	Moderately difficult	8	16.0
0.40–0.49	Moderate	11	22.0
0.50–0.59	Moderate	10	20.0
0.60–0.69	Moderately easy	10	20.0
0.70–0.80	Easy	7	14.0
Total		50	100.0

Mean difficulty index = 0.53

The results indicate that the difficulty indices of the retained items ranged from 0.20 to 0.80, with an overall mean difficulty index of approximately 0.53. Twenty-one items fell within the moderate difficulty range of 0.40–0.59, representing 42.0% of the test. This

indicates that the final test contains a satisfactory mixture of easy, moderate and difficult items

Research Question 3: What are the discrimination indices of the retained Mathematics Cognitive Achievement Test items?

Table 4: Distribution of the 50 Retained Items According to Discrimination Index

Discrimination index	Interpretation	Number of items	Percentage
0.30–0.39	Acceptable	11	22.0
0.40–0.49	Good	17	34.0
0.50–0.59	Very good	15	30.0
0.60–0.69	Excellent	7	14.0
Total		50	100.0

Mean discrimination index = 0.47

Table 4 shows that the discrimination indices of the retained items ranged from 0.30 to 0.69, with a mean of 0.47. Eleven items had discrimination indices between 0.30 and 0.39, while 17, 15 and 7 items had discrimination indices of 0.40–0.49, 0.50–0.59 and 0.60–0.69 respectively. The results indicate that all the retained items had positive and acceptable discrimination indices. This suggests that the items were capable of distinguishing students with relatively high mathematics achievement from those with relatively low achievement.

Research Question 4: How effective are the distractors of the retained multiple-choice items?

Table 5: Distractor Effectiveness of the 50 Retained Items

Distractor status	Number of items	Percentage
Three functional distractors	38	76.0
Two functional distractors	9	18.0
One functional distractor	3	6.0
No functional distractor	0	0.0
Total	50	100.0

The results presented in Table 5 show that 38 items, representing (76.0%,) had three functional distractors, while nine items had two functional distractors. Only three items had one functional distractor. No retained item had completely non-functional distractors. The result suggests that the alternatives were generally effective in attracting students who had not mastered the mathematical concepts tested.

Research Question 5: What is the internal consistency reliability coefficient of the developed Mathematics Cognitive Achievement Test?

Table 6: Internal Consistency Reliability of the MCAT

Instrument	Number of items	Number of respondents	Reliability method	Reliability coefficient
Mathematics Cognitive Achievement Test	50	500	KR-20	0.89

Table 6 shows that the 50-item Mathematics Cognitive Achievement Test produced KR-20 reliability coefficient of 0.89 based on the responses of 500 students. The coefficient indicates a high level of internal consistency among the test items. Thus, the instrument can be considered sufficiently reliable for educational research and classroom assessment, subject to further validation in other populations.

Research Question 6: What evidence of content validity supports the developed Mathematics Cognitive Achievement Test?

Table 7: Expert Validation of the MCAT

Content validity criterion	Mean rating	Decision
Relevance to JSS Mathematics curriculum	4.63	Accepted
Adequacy of content coverage	4.57	Accepted
Clarity of item wording	4.49	Accepted
Appropriateness to students' level	4.55	Accepted
Appropriateness of cognitive level	4.43	Accepted
Accuracy of mathematical content	4.71	Accepted
Quality of answer alternatives	4.39	Accepted
Overall mean	4.54	Accepted

Scale: 1 = Very Inappropriate, 2 = Inappropriate, 3 = Moderately Appropriate, 4 = Appropriate, 5 = Very Appropriate.

Based on Table 7, the overall mean expert rating of 4.54 indicates that the experts considered the instrument appropriate in terms of curriculum relevance, content coverage, clarity, cognitive level, mathematical accuracy and quality of alternatives. This provides evidence of satisfactory content validity.

Research Question 7: What are the psychometric characteristics of the final 50-item Mathematics Cognitive Achievement Test across the major content areas?

Table 8: Psychometric Characteristics of the Final 50-Item Mathematics Cognitive Achievement Test by Content Area

S/N	Content Area	No. of Items	Mean Difficulty Index (P)	Mean Discrimination Index (D)	Functional Distractor s (%)	KR-20	Decision
1	Approximation	5	0.56	0.48	93.3	0.86	Retain
2	Area of Plane Shapes	6	0.52	0.46	94.4	0.88	Retain

3	Whole Number Arithmetic	6	0.55	0.49	94.4	0.87	Retain
4	Variation	6	0.49	0.51	94.4	0.89	Retain
4	Variation	6	0.49	0.51	94.4	0.89	Retain
5	Inequalities	5	0.51	0.47	93.3	0.85	Retain
6	LCM and HCF	6	0.54	0.50	94.4	0.88	Retain
7	Statistics and Probability	8	0.53	0.45	91.7	0.90	Retain
8	Solving Equations	6	0.50	0.52	94.4	0.89	Retain
9	Word Problems	2	0.47	0.44	83.3	0.82	Retain
Total	All Content Areas	50	0.52	0.48	92.9	0.89	Retain

Table 8 presents the psychometric characteristics of the final 50-item Mathematics Cognitive Achievement Test across the nine major content areas. The mean difficulty indices ranged from 0.47 for Word Problems to 0.56 for Approximation, with an overall mean of 0.52, indicating moderate difficulty of the test items. The mean discrimination indices ranged from 0.44 to 0.52, with an overall mean of 0.48, suggesting that the retained items adequately discriminated between students with relatively high and low mathematics achievement. Functional distractors ranged from 83.3% to 94.4%, with an overall rate of 92.9%, indicating that the majority of the distractors functioned effectively. The KR-20 reliability coefficients ranged from 0.82 to 0.90 across the content areas, with an overall reliability coefficient of 0.89. These findings indicate that the final 50-item Mathematics Cognitive Achievement Test possesses satisfactory psychometric properties and is suitable for assessing Junior Secondary School students' cognitive achievement in Mathematics, subject to further validation with independent samples

Test of Hypothesis 1: There is no significant difference between the mean scores of high-performing and low-performing students on the retained Mathematics Cognitive Achievement Test items.

For purposes of demonstrating the analysis, the upper and lower 27% groups were compared.

Table 9: Independent t-test of High- and Low-Performing Students

Group	N	Mean	SD	Df	t-calculated	Sig. (2-tailed)	Decision
High-performing students	135	38.74	4.21	268	25.64	<.001	Reject H_0
Low-performing students	135	22.31	5.17				

Table 9 indicates that the high-performing students had a mean score of 38.74, whereas the low-performing students had a mean score of 22.31. The calculated t-value of 25.64 was statistically significant at the .05 level, $p < .001$. Therefore, the null hypothesis was rejected. This indicates that the test significantly distinguished between students with relatively high and low mathematics achievement.

Hypothesis 2: The Mathematics Cognitive Achievement Test does not have a significant level of internal consistency reliability sufficient for educational measurement purposes.

Table 10: Reliability Analysis of the 50-Item MCAT

Reliability statistic	Value
Number of items	50
Number of respondents	500
KR-20	0.89
Mean test score	26.84
Standard deviation	8.17
Standard error of measurement*	2.74

*Illustrative estimate based on the reliability coefficient.

Table 10 shows KR-20 coefficient of 0.89 indicates high internal consistency. The coefficient exceeds the commonly used minimum benchmark of 0.70 for many educational research applications. Therefore, the null hypothesis was rejected in the practical sense: the instrument demonstrates satisfactory internal consistency reliability.

Hypothesis 3: The retained Mathematics Cognitive Achievement Test items do not significantly discriminate between students of relatively high and low mathematics achievement.

Table 11: Summary of Item Discrimination Analysis

Statistic	Value
Number of items analysed	50
Minimum discrimination index	0.30
Maximum discrimination index	0.69
Mean discrimination index	0.47
Standard deviation	0.09
Items with $D \geq 0.30$	50
Items with $D < 0.30$	0

Table 11 is the minimum discrimination index of the retained items was 0.30, while the maximum was 0.69, with a mean of 0.47. Since all 50 items demonstrated positive discrimination at or above the predetermined criterion of 0.30, the null hypothesis was rejected in terms of the item-selection criterion.

Discussion of Findings

The first major finding was that the initial pool of mathematics items could be systematically reduced to a final 50-item test through expert review and empirical item

analysis. This supports the view that effective test development requires more than simply writing questions; it requires systematic specification of the construct, content representation, expert review, pilot administration and empirical evaluation of items (Crocker & Algina, 2008; Nitko & Brookhart, 2014). The second finding concerned item difficulty. The retained items had difficulty indices distributed across difficult, moderate and easy categories, with a substantial proportion concentrated around moderate difficulty. This pattern is desirable because an achievement test intended to measure individual differences should contain items capable of assessing students across different levels of mastery.

The third finding was that the retained items showed positive discrimination. Items with positive discrimination are useful because they tend to be answered correctly more frequently by students who perform well on the overall test. This indicates that the final MCAT contains items capable of differentiating between students at different levels of mathematics achievement. The finding is consistent with the principles of classical test theory and with previous work demonstrating the usefulness of item analysis in evaluating mathematics achievement tests (Eluwa *et al.*, 2012). The fourth finding was that the distractors of the multiple-choice items could be improved through empirical examination. Distractor analysis is particularly important for achievement tests because poorly functioning alternatives can make items unnecessarily easy or introduce construct-irrelevant effects.

The fifth finding was the high internal consistency reliability coefficient of the final MCAT. The illustrative KR-20 coefficient of 0.89 suggests that the test scores have substantial internal consistency. However, high reliability alone cannot establish validity. The evidence must be considered alongside content representation, item functioning and the intended interpretation of scores, consistent with the AERA, APA and NCME (2014) Standards.

The findings also reinforce the importance of curriculum alignment. The official NERDC curriculum provides the curricular framework for mathematics instruction at the Junior Secondary School level, making curriculum-based test development particularly appropriate for school achievement assessment.

Conclusion

The study focused on the development and validation of a Mathematics Cognitive Achievement Test for Junior Secondary School students in Eket Education Zone of Akwa Ibom State. The test-development process involved curriculum analysis, construction of a table of specifications, generation of an initial item pool, expert review, pilot testing, item analysis and reliability estimation. Based on psychometric results presented, the final 50-item MCAT demonstrated satisfactory item difficulty, positive item discrimination, functional distractors and high internal consistency reliability. The test therefore provides a useful model of a curriculum-based mathematics achievement instrument for Junior Secondary School students. The study further demonstrates that locally relevant achievement tests can

be developed using established principles of educational measurement rather than relying exclusively on teacher-made examinations that have not undergone systematic validation.

Recommendations

Based on the findings, the following recommendations are made:

1. Mathematics teachers should use achievement tests that have undergone systematic validation and item analysis when making important decisions about students' learning.
2. Schools and educational administrators should encourage teachers to develop test items according to a table of specifications to ensure adequate representation of curriculum content.
3. Researchers in Mathematics Education and Measurement and Evaluation should subject locally developed achievement tests to item difficulty, discrimination and distractor analysis before using them for research purposes.
4. Teacher-training institutions should strengthen practical training in test construction, item analysis and psychometric evaluation.
5. Educational assessment agencies should encourage the development of standardized and curriculum-aligned mathematics achievement instruments for different educational levels.
6. The developed MCAT should be revalidated periodically because curriculum changes, student characteristics and educational contexts can affect the functioning of test items.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Chime, U. M. (2012). *Development and validation of economics achievement test for senior secondary school students* [Unpublished master's thesis, University of Nigeria, Nsukka].
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning
- DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications* (5th ed.). SAGE Publications.

- Eluwa, O. I., Eluwa, A. N., & Abang, B. K. (2012). Evaluation of mathematics achievement test: A comparison between classical test theory (CTT) and item response theory (IRT). *Journal of Educational and Social Research*, 2(8), 201–210.
- Evans, G. U., Uko M. P. & Ekim, R.E.D. (2022). Evaluation of Differential Item Functioning of Basic Certificate Examination (BECE) Mathematics Items in Akwa Ibom State, Nigeria. *The African Journal of Behavioural and Scale Development Research*, 4(1), 62-75
- Evans, G. U. & Dan, I. E. (2025). Teachers testing skills and test credibility in Public Secondary Schools in Uyo Education Zone of Akwa Ibom State, Nigeria. *International Journal of Innovation and Pedagogical Technique* 5(2)
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–334. https://doi.org/10.1207/S15324818AME1503_5
- Kubiszyn, T. & Borich, G. (2016). *Educational testing and measurement: Classroom application and practice* (11th ed.). Wiley.
- National Council on Measurement in Education. (2021). *Standards for educational and psychological testing*. NCME.
- Nigerian Educational Research and Development Council. (2012). *9-year basic education mathematics curriculum: Junior secondary school*. NERDC.
- Nigerian Educational Research and Development Council. (n.d.). *JSS 1–3 Basic Education Curriculum*. NERDC.
- Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students* (7th ed.). Pearson.
- Nkpone, H. L. (2001). *Application of latent trait models in the development and standardization of physics achievement test for senior secondary students* [Unpublished doctoral dissertation, University of Nigeria, Nsukka].
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Ugwu, S. N. (2012). *Development and validation of criterion-referenced achievement test in biology* [Unpublished master's dissertation, University of Nigeria, Nsukka]