E-ISSN: 1115 - 8425 P-ISSN: 3027-012X
MPRI-JAEE | INTERNATIONAL JOURNAL OF ASSESSMENT & EVALUATION IN EDU.
VOL. 08 NO. 8 – APRIL. 2025

DEVELOPMENT AND VALIDATION OF BASIC SCIENCE CRITERION REFERENCED TEST FOR JUNIOR SECONDARY SCHOOL STUDENTS IN MAIDUGURI METROPOLIS, BORNO STATE, NIGERIA

MOHAMMED ALI; DR. FATI S. ABANI; & PROF. ALI MBURZA

Department of Education, University of Maiduguri, Borno State, Nigeria.

Corresponding Author:
mujaheedmali95@gmail.com

DOI: https://doi.org/10.70382/mejaee.v8i8.053

Abstract

The study developed and validated Basic Science criterion referenced test for junior secondary school students based on: alpha reliability of the test, relationship between item's score and total scores of the test and item's discrimination power of the test. The study adopted descriptive survey research design. The population for the study consists of all students of twenty junior secondary three in Maiduguri Metropolis that were purposively selected from a population of 6,950 students. Out of 6,950 students, 362 male and female students were sampled using stratified sampling techniques. Basic Science Criterion Referenced Test (B-SCRT) was the instrument for Data collection. The study found that the Cronbach's alpha reliability of the objectives and essay test were .97 and .90. The study also found positive correlation coefficient of .14 to .73 for the relationship between item's score and total scores of item 1 to 60. The study further revealed

Keyword:
Development, Validation, Basic Science, Criterion Referenced Test, Junior Secondary School Students, Maiduguri Metropolis, Borno State.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

positive discrimination power of between 16 and 56% Based on the findings, it was concluded that, the constructed Basic Science criterion referenced test is highly consistent, positive relationship exists between item's scores and total scores. The study also found that, all the discrimination power of the test were appropriate. Based on the findings, it was recommended that: criterion referenced test with at least reliability index of 0.90 should be considered by test developers when developing reliable test to ensure the internal consistency of the test. Item's of criterion referenced test with positive correlation between items scores and total scores should be developed by test developers to support the validity of the test. When developing objectives or essay, item of criterion referenced test by the test developers, there is need to develop items with positive discrimination power that could identify poor average and good performing test takers.

Introduction

Criterion referenced test is an instrument used to measure students' performance. It compares students' performance to predetermined criteria. Put otherwise, it refers to a predetermined set of standards used to evaluate test takers' performance. Determining the degree to which an examinee has met the established requirements is the purpose of the criterion-referenced test. A criterion referenced test evaluates students' performance in accordance with a predetermined benchmark or criterion. In order to pass test, students must perform at this level and students' scores are evaluated in relation to the criterion score rather than the scores of other test takers (Poham, 2015).

Development of criterion referenced test requires a systematic and detailed-oriented approach based on sound theoretical educational measurement principles to ensure sufficient validity evidence and support the proposed inferences from the test scores. Details and issues, both large and small, comprises the enterprise usually associated with the terms test development. All of these details must be well executed to produce a test that estimates students' achievement and ability fairly and consistently in the content domain purported to be measured by the test and to provide documented evidence in support of the proposed test score inferences.

Assembling criterion-referenced test necessitates a meticulous methodology grounded on reliable theoretical principles of educational assessment in order to guarantee adequate validity evidence and bolster the suggested conclusion drawn

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

from the test results. Large and small details and problems make up the enterprise that is typically connected to the phrases test development. In order to create a test that measures students' success and ability in the content domain that the test is intended to measure fairly and consistently, as well as to offer recorded evidence in support of the suggested test result inferences, all of these things must be completed with precision items (Mehrens & Lehmann**,** 2021)

Another crucial stage in creating a test with high-quality items in terms of accuracy was validation, which comes after a test has been created, administered and scored. It is one of the most crucial steps in determining the caliber of the items. The key components of a criterion-referenced test, according to Haladyna and Downing (2017), include reliability, relationship between items score and total score and items discrimination power.

Test reliability is crucial when it comes to the criteria domain. Measurement's consistency or the degree of consistency between test scores or other evaluation findings across measurements, is referred to as reliability. It can range from 00.0 (if no variance is consistent) to 1.00 (if all variance is consistent) with all values between 00.0 and 1.00 also being possible. For example, if the Cronbach alpha for a set of scores turns out to be .90, you can interpret that as meaning that the test is 90% reliable and by extension that it is 10% unreliable (100% - 90% = 10%). Professionally developed high-stakes standardized tests should have internal consistency coefficients of at least .90. Lower-stakes standardized tests should have internal consistency of at least .85 or .80. For teacher made test, it is desirable to have a reliability coefficient of .70 (Willnat & Manheim, 2022).

Test item correlation used to support the validity, that the items are good contributor to what the test measures; it has sometimes been called an index of item validity (Borsboom, 2021). It is the correlation between an individual item and the total score without that item. For example, if you had a test that had 20 items, there would be 20-item total correlations. For item 1, it would be the correlation between item 1 and the sum of the other 19 items. Item-total correlation is performed to check if any item in the set of tests is inconsistent with the averaged behaviour of the others (Everitt, 2020).

Items' discrimination power of test measure how well an item is able to distinguish between examinees who are knowledgeable and those who are not or between masters and non-masters. The desired level of discrimination power for a criterion referenced test was based on item's ability to discriminate between high and low

achievers and is determined by the learning outcome of the test contents designed to measure.

Objectives of the Study

The objectives of the study were to determine:

1. alpha's reliability of Basic Science criterion referenced test for junior secondary school students in Maiduguri Metropolis.
2. items' score and total score correlation of Basic Science criterion referenced test for junior secondary school students in Maiduguri Metropolis.
3. items' discrimination power of Basic Science Criterion Referenced Test for junior secondary school students in Maiduguri Metropolis.

Research Questions

The following research questions were answered:

1. What is the alpha's reliability of the Basic Science criterion referenced test for junior secondary school students?
2. What is the relationship between item's score and total scores of the Basic Science criterion referenced test for junior secondary school students?
3. What is the items' discrimination power of the Basic Science Criterion referenced test for junior secondary school students?

Related Literature Review

Development and validation of criterion referenced test consist of a progressive activity, which depends on steps and principles. Careful planning and compulsive execution of the plan leads to test that validly measure examinee's ability in line with content domain of interest. Mackenzie and Padilla (2019) constructed and validated a test of graphing in science for junior secondary school in Nairobi. Descriptive survey research design was used. The population consist of 7,000 junior secondary school students and a sample of 380 students were sampled. Twenty 20-objectives and 5 essay tests items of graphing in Science were administered to 380 students. Data were analyzed using Cronbachs' alpha statistics. The reliability coefficient reported was .83. It was concluded that, the test item is highly reliable. Based on the finding, it was recommended that, test items should be used for graphing in Science.

Obioma (2021) developed, validated and normed a Mathematics criterion referenced test for Nigerian secondary school students in Anambra State. Conceptual framework from Achievement Testing was adopted. Instrumentation design was used. The curriculum and learning objectives of mathematics taught in junior secondary schools were used to develop 60 objectives test and validated in two stages. Pilot testing of the items was first done on 200 students and after that, 300 students were selected at random from a population of 3,200 in the Nsukka Education Zone of Anambra State. Sixty (60) objectives test were used for data collection and Pearson product Moment correlation statistics was employed to analyze the data. The finding showed that, the reliability coefficient was 0.77. It was concluded that, the internal consistency of the test was low and suggested that, some of the test items need to be rechecked through item's revision.

Audu (2022) determined the reliability of teacher-made tests in Agricultural Science for senior secondary schools in Borno State. The study was based on Latent Trait Model. Descriptive survey was utilized as research design. From a population of 5,300 senior secondary II students, a sample of 106 senior secondary II students' scores on teacher-made tests were used for the study. Data were analyzed using test-re-test methods of estimating reliability. The findings revealed a high reliability index of 0.83. It was concluded that, 0.83 index was good for teacher made test. It was recommended that, such index be used for teacher made test.

Nwankwo (2020) developed and validated Agricultural Science criterion referenced test for junior secondary school students. Concept Achievement Testing was utilized. Instrumentation research design was used. The study was carried out in the Education zones of old Imo State with a population of 5,800 students and a sample of 470 students. Data were collected with 50 objectives and 5 essays tests item of Agricultural Science criterion referenced test and analyzed using Cronbach alpha statistics. The reliability estimates of stability and internal consistency found was .90. It was concluded that the test items were internally consisted. Based on the finding, it was recommended that the test items were in line with standardized test.

Nduku (2019) developed and validated Integrated science criterion referenced test in Imo State. Conceptual framework based on Criterion Referenced Test (CRT) was utilized. Method research design was used for the study. A sample of 848 students were randomly selected from 42 secondary schools in five Education

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Zones of Imo State. The instrument for data collection was 60 items of test based on four level of cognitive domain. Data were analyzed using Pearson Products Moment Correlation coefficient. After the analysis .12~.59 correlation coefficient within the test items was reported. It was concluded that, all the test items were effective. From the findings, it was recommended that, items of criterion referenced test with relationship between items scores and total score should be considered.

In another study, Rissa (2019) developed and validated 50 objectives items of Physics criterion referenced test based on senior secondary school curriculum in Kano State. Test development concept was adopted as the conceptual framework of the study. Instrumentation and survey design were used as research design. The items were administered on a sample of 280 Senior Secondary Three (SS 3) students sampled from 3,100 students. Statistical analysis was conducted by Pearson Product Moment correlation statistics. Analyzed data indicated that, the item's score correlation among the content area of Physics were 19~.68. It was concluded that, all the items were positive and related to each other. Based on the finding, it was recommended that, standardized criterion referenced test should have at least index of .20 and above for relationship between items scores and total score.

Anene and Ndubisi (2019) analyze Agricultural Science criterion referenced test of junior secondary school one students in Delta State. The study developed and validated 50 objectives and 7 essays questions. Instrumentation research design was utilized. The study determined, the psychometric quality, construct related validity, items score correlation and reliability of the test items. The total population was 1,500 students and 300 junior school one students were selected through simple random sampling technique. Data were analyzed using descriptive statistics, principal components analysis, item analysis and Pearson Products Moment Correlation. The analyzed data revealed that, the relationship between item score and total score were .23 to .63 with reliability index of .90 and 4 extracted items. It was concluded that, the test items were valid and reliable. It was recommended that, test items that internally consistent and valid should be developed.

Inomiesa (2022) constructed and validated criterion achievement test for upper primary Science in Kenya. Item Response Theory served as the study's theoretical framework. The research design used in the study was descriptive survey. There

are 800 kids in upper primary 6, but 210 of them were randomly selected. After screening, the 200 items that were initially selected for the study, the items reduced to 102. Descriptive statistics and measure of discrimination formula were used to analyze the data. After the analysis, the finding indicated that test items have discriminating power of between 0.8 and 49%. It was concluded that, the test items have weak, moderate and strong discrimination power. It was also concluded that, about 4 items need to be revised. Based on the findings, it was recommended that, items 22, 23, 30 and 41 need to be modified.

Churchill (2018) investigated the discriminating power and condition of particular test items within a specific context of the business field (accountancy) at the University of Sao Paulo, Brazil. The study, which involved 66 and 62 students, used an experimental design (experiments E1 & E2). Data were gathered in the second semester when the same set of students participated in two distinct studies. The second experimental group was exposed to short-answer questions, regular-to-easy level and short items, the first group was exposed to a multiple-choice approach, regular-to-difficult level and lengthy examination. Measure of discrimination power statistics was utilized in this investigation to analyze the data. The result indicated experiment 1 (regular-to-difficult) had an average score of 60.06% of the total score and experiment 2 (regular-to-easy) had an average score of 70.01%.

The findings also indicated that, the degree of discrimination power in each experiment. Additionally, the standard deviation of the multiple-choice trial (experiment 1) is lower (0.1485) than the standard deviation of the short-answer questions (0.1952). Additionally, the outcomes demonstrated that experiments 1 and 2's Discriminating Power (DP) values were positive. The DP values ranged from 0.00 in experiment 1 to 0.67 (question 18) in experiment 2 and from 0.10 (question 1) to 0.39 (question 5) in experiment 2. It was concluded that, the discrimination power of the test items was positive and recommended that, the discrimination power of the test items should be based on learning outcome of the content.

Mahjabeen, Alam, Hassan, Zafar and Rizvi (2019) assessed the discrimination power of 65 multiple choice questions of 4th-year MBBS students at Medical and Dental College Islamabad. Survey was used to guide the study and from 1,200 population of students, 110 were selected and participated in the study. Data analysis was done using the percentage and discrimination formula statics. The

findings indicated that, 53 (81%) of the multiple-choice questions fell into the acceptable category, while just 1 % of the questions were too easy or too difficult. In terms of DI, a total of 34 (62%) MCQs indicated a great discriminating tendency to separate students who performed well from those who did not, while 15 (23%) and 5 (8%), as well as 11 (17%) MCT, showed good, acceptable and poor discriminating power. It was concluded that, the test items discriminating well and suggested, test items with positive discrimination index should be developed.

Omobolanle (2016) developed and standardized Biology criterion referenced test for Senior Secondary One (SS 1) students. The work was guided by concept achievement testing with an instrumentation research approach. The population of the study consisted of 1,258 Senior Secondary one (SS 1) Biology students in Etche Local Government Area, Rivers State, with a sample size of 636 Biology students. A Table of specifications was used to create 100 Form I B-CRAT multiple-choice questions. Experts in Biology and educational measurement and assessment face-validated the Form I B-CRAT. Percentage analysis was used to analyzed the B-CRAT. The finding results to discrimination power between 12% and 16%. It was concluded that, all the test items were effective. Based on the results, it was recommended that, discrimination power of the test items should be aligned with subject content and test items with positive discrimination power should be constructed.

Methodology

The study adopted descriptive research design. The choice of the design was in accordance with Schindler's (2018) who suggested that, survey be carried out using questionnaires, scales or any other pertinent tools to collect data reflecting the knowledge, attitudes, behaviour, views and beliefs of the population that cannot be immediately observed.

The population for this study consists of all students of the thirty-nine (39) junior secondary schools in Maiduguri Metropolis. Twenty (20) junior secondary schools were purposively selected and participated in the study with a total of six thousand, nine hundred and fifty (6,950) students. Out of the 6,950 population, three thousand, two hundred and fifty (3,250) were male and three thousand, seven hundred (3,700) were female JSS III students. To ensure the inclusion of the desired sample of the population of the student, stratified sampling technique was used to sample three hundred and sixty-two (362) male and female junior

secondary school three students using the Research Advisor (2015) Table for determining sample size as a guide and the sample size was at 95% confidence and 0.05 margin of error. Basic Science Criterion Referenced Test was the research instrument for data collection.

Results

Data were analyzed and presented as follows.

Research Question One: What is the alpha's reliability of the Basic Science criterion referenced test for junior secondary school students?

Table 1: Analysis of Data on the Internal Consistency Reliability of the Constructed Basic Science Criterion Referenced Test for Junior Secondary School (Objectives)

Total Items	N	Cronbach's alpha
60	362	.97

The result of data analysis in Table 1 on the Cronbachs' alpha reliability revealed that, the Cronbachs' alpha reliability of the objectives Basic Science criterion referenced test was .97. This shows that, the objective test items are highly consistent.

Table 2: Analysis of Data on the Internal Consistency Reliability of the Constructed Basic Science Criterion Referenced Test for Junior Secondary School (Essay)

Total Items	N	Cronbach's alpha
18	362	.90

The result of data analysis in Table 2 on the Cronbach alpha reliability of the essay constructed Basic Science criterion referenced test was .90 which indicates that, the essay test items are highly consistent.

Research Question Two: What is the relationship between item's score and total scores of the Basic Science criterion referenced test for junior secondary school students?

Table 3: Analysis of Data on the Relationship between Item's Score and Total Score of the Constructed Basic Science Criterion Referenced Test

Item	Scale Mean	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's alpha if Item Deleted
1	30.03	312	.14	.97
2	30.09	309	.33	.96
3	30.06	309	.33	.96
4	30.11	307	.40	.96
5	30.04	307	.23	.97
6	30.11	306	.47	.96
7	30.09	306	.48	.96
8	30.11	306	.49	.96
9	30.08	306	.48	.9
10	30.07	306	.49	.96
11	30.10	304	.60	.96
12	30.06	303	.42	.96
13	30.08	303	.62	.96
14	30.07	304	.40	.96
15	30.09	304	.60	.96
16	30.07	304	.61	.96
17	30.08	303	.67	.96
18	30.10	304	.63	.96
19	30.12	302	.69	.96
20	30.12	304	.63	.96
21	30.12	303	.64	.96
22	30.12	304	.61	.96
23	30.10	303	.66	.96
24	30.11	303	.65	.96
25	30.09	302	.69	.96
26	30.11	303	.63	.96
27	30.10	303	.66	.96
28	30.06	303	.47	.96
29	30.13	302	.71	.96

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Item	Scale Mean	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's alpha if Item Deleted
30	30.11	303	.68	.96
31	30.12	302	.71	.96
32	30.07	302	.49	.96
33	30.10	302	.70	.96
34	30.10	303	.68	.96
35	30.09	302	.72	.96
36	30.15	302	.69	.96
37	30.12	301	.72	.96
38	30.10	302	.66	.96
39	30.08	301	.73	.96
40	30.07	302	.67	.96
41	30.06	302	.70	.96
42	30.09	302	.71	.96
43	30.10	303	.66	.96
44	30.09	304	.62	.96
45	30.10	304	.61	.96
46	30.10	303	.66	.96
47	30.10	304	.56	.96
48	30.10	304	.59	.96
49	30.08	305	.53	.96
50	30.09	304	.54	.96
51	30.09	306	.46	.96
52	30.13	305	.51	.96
53	30.11	306	.50	.96
54	30.15	307	.45	.96
55	30.11	306	.47	.96
56	30.15	307	.42	.96
57	30.10	307	.42	.96
58	30.12	308	.37	.96
59	30.10	309	.30	.96
60	30.10	309	.30	.96

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

The results of data analysis in Table 3 on the relationship between item's score and total scores of the Basic Science criterion referenced test reveals a positive relationship between item's score and total scores of items 1 to 60. The results further show that, the reliability index of item 1 to 60 was between .96 and .97. Therefore, all the test items were positive and related.

Research Question Three: What is the item discrimination power of the Basic Science criterion referenced test for junior secondary school students?

Table 4: Analysis of Data on Item's Discrimination Power of the Basic Science Criterion Referenced Test for Junior Secondary School Students

Item	N of Students Who Got the Item Right at Upper	N of Students Who Got the Item Right at Lower	% of Discrimination Power
1	120	52	37
2	111	30	44
3	80	39	22
4	88	30	32
5	80	37	23
6	81	36	24
7	92	26	36
8	90	32	32
9	80	32	26
10	82	35	26
11	90	47	23
12	94	30	35
13	90	28	34
14	87	31	30
15	97	40	31
16	100	32	38
17	99	34	36
18	90	32	32
19	86	39	26
20	85	27	32
21	85	29	30
22	93	41	29
23	86	40	25
24	82	37	25
25	83	43	22
26	78	43	19
27	82	41	23

Item	N of Students Who Got the Item Right at Upper	N of Students Who Got the Item Right at Lower	% of Discrimination Power
28	80	47	18
29	85	29	35
30	87	44	24
31	90	48	23
32	90	38	29
33	80	45	19
34	100	53	25
35	95	41	30
36	99	44	30
37	116	66	28
38	90	28	34
39	102	47	30
40	112	33	44
41	104	36	38
42	117	38	44
43	107	41	36
44	128	62	36
45	99	70	16
46	87	29	32
47	119	40	44
48	140	62	43
49	124	71	29
50	116	30	48
51	140	40	55
52	151	61	50
53	136	49	48
54	144	43	56
55	129	38	50
56	162	71	50
57	128	48	44
58	109	39	39
59	141	49	51
60	167	79	49

The results of data analysis in Table 4 on item's discrimination power of the test revealed that, the discrimination power of item 1 to 60 was between 16 and 56%. The results indicated that, the discrimination power of all the Basic Science criterion referenced test were excellent.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Discussion

The finding's of the study with regard to reliability of objectives and essay test were .97 and .90 which indicated that, the reliability for internal consistency of the test is high. This further indicated that, the instrument is highly standard in term of reliability. The finding was similar to Meshkani and Abadie (2019) that reported a high reliability index of .90 at the University of Tehran for Basic Science and clinical students and that of Downing (2006) and Willnat and Manheim (2022) who recommended that a professionally developed and standardized tests should have internal consistency coefficient of .90. Lower-stakes standardized tests should have internal consistency of .85 and teacher made test, should have a reliability coefficient of .70.

The finding of the study on the relationship between item's score and total score of the constructed Basic Science criterion referenced test, revealed a positive correlation coefficient of .14 to .73 for the relationship between item's score and total and .96 to .97 reliability coefficient of internal consistency for item 1 to 60. From the findings, all the test items were internally related. The findings were in line with Rissa (2019) who developed, validated and found .27 to .70 for the relationship between item's score and total score of 50 objectives items of Physics criterion referenced test based on senior secondary school curriculum in Kano State. With regard to item's discrimination power of test, the finding revealed that, the discrimination power of item 1 to 60 was between 16 and 56%. The finding indicated that, the test has an excellent discrimination power. This finding was supported by that of Bable and Kuechler (2020) that analyzed the discrimination power of 50 objectives questions of WAEC and reported 18 to 59% discrimination power. The finding was also supported by that of Churchill (2018) who found 11 to 62%.

Conclusion

In view of the findings, it was concluded that, the developed Basic Science criterion referenced test was highly consistent, a positive relationship between item's score and total score of developed Basic Science criterion referenced test exists. It was also concluded that, the discrimination power of the developed Basic Science criterion referenced test were positive and excellent.

Recommendations

The following recommendations were made:

1. items of criterion referenced test with at least reliability index of 0.90 should be considered by test developers when developing reliable test to ensure the internal consistency for the test.
2. items of criterion referenced test with positive relationship between item score and total scores should be developed by test developers to support the validity of the test items.
3. when developing objectives or essay items of criterion referenced test by test developers, there is need to develop items with positive discrimination power that could identify poor, average and good performing test takers

References

Adams, I. K. (2015). Construction and validation of achievement test in integrated science for the Nigerian secondary school class one, Unpublished M. Ed. Thesis, University of Ibadan.

Anene, A. O. & Ndubisi, N. (2019). The influence of laboratory experiment on the performance of Nigerian secondary school students in the University of Nigeria, Nsukka, Unpublished M. Ed. Dissertation, University of Nigeria, Nsukka.

Audu, D. (2022). Assessment of procedures and quality of teacher-made tests in agricultural science of senior secondary school in Borno State, Nigeria, Unpublished Master Dissertation, University of Maiduguri.

Bable, L., M. G. & Kuechler, W. L. (2020). Using multiple-choice tests to evaluate students understanding of accounting, accounting education, *International Journal*, 15 (4), 20-25.

Borsboom, D. (2022). When does measurement invariance matter? *Medical Care, 44(11), 176-181*

Churchill, G. A. (2018). A paradigm for developing better measures of constructs, *Journal of Business Research,* 16 (1), 64-73.

Downing, S. M. (2006). Test item development: Validity evidence from quality assurance procedures, *Applied Measurement in Education,* 10 (1), 61-82.

Everitt, B. S. (2020). The Cambridge dictionary of statistics, (2nd Edition), Cambridge UK: *Cambridge University Press.*

Haladyna, T. M. & Downing, S. M. (2017). Construct-irrelevant variance in high-stakes testing, *Educational Measurement, Issues and Practice, 23 (1), 17- 27.*

Inomiesa, E. A. (2022). The development, validation and use of standardized instrument for the continuous assessment of pupil achievement in upper primary science, Unpublished Ph.D. Dissertation, University of Nigeria Nsukka.

Mackenzie, A. & Padilla, E. (2019). Implementing a class of structural change tests*: An Econometric, Computing Approach Computational Statistics and Data Analysis, 50(11), 2987-3008.*

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T. & Rizvi, M. (2019). Difficulty index, discrimination index and distractor efficiency in multiple choice questions, *Annals of PIMS, 18 (14), 310-315.*

Mehrens, W. A. & Lehmann, I. J. (2021). Measurement and evaluation in education and psychology, New York: Harcourt Brace.

Meshkani, Z. & Abadie, F. (2019). Multivariate analysis of factors influencing reliability of teacher-made tests, *Journal of Medical Education*, 6(2), 149-150.

Nduku, K. L. (2019). Development and validation of integrated science achievement test for junior secondary school students, Unpublished M. Ed. Thesis University of Nigeria Nsukka.

Nwankwo, R. (2020). Construction and validation of test. A seminar paper presented at 14th yearly conference, University of Port Harcourt.

Obioma, A. S. (2021). Empirical analysis of item difficulty and discrimination indices of senior school certificates multiple choice Biology tests in Nigeria, *Creativity Research Journal, 8 (3), 23-27.*

Omobolanle, I. M. (2016). Construction and measurement concept and applications, Owerri: Career Publishers.

Poham, W. J. (2015). Classroom assessment: What teachers need to Know (5th ed.), Boston: *Allyn and Bacon*.

Rissa, D. M. (2019). Further investigation of non-functioning options in multiple choice test items, Educational Measurement, 54(4), 861-872.

Schindler, M. (2018). Measurement: Reliability, construct validation and scale construction. In Reis, H. T. & Judd, C. M. (Eds.). Handbook of research methods in social and personality psychology, 339-370. Cambridge, UK: *Cambridge University*.

Willnat, D. & Manheim, G. (2022) Empirical political analysis (6th edition), Boston, MA: Longman Press.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]