



Development and Item Response Theory Calibration of Economics Achievement Test for senior secondary school students Plateau State, Nigeria.

Dr Joseph John Mawak (Ph.D)¹; Dr Queen Efomo Igabari (Ph.D)²; Prof. Abbas Yusuf Mustapha (Ph.D)³

¹ Department of Educational Foundations, Faculty of Education, University of Jos.

Email: mawakjoseph74@gmail.com; Phone: +2348067773253

² Department of guidance and counseling, Faculty of education, Delta state university Abraka.

Email: qe-igabari@delsu.edu.ng, Phone: +2348039466534

³ Department of Educational Foundations, Faculty of Education, University of Jos.

Email: abbsmusty@yahoo.com. Phone: +2348031857059

ARTICLE INFO

Article No.: 021724023

Type: Research

Full Text: [PDF](#), [PHP](#), [HTML](#), [EPUB](#), [MP3](#)

Accepted: 21/02/2024

Published: 14/03/2024

*Corresponding Author

Dr Joseph John Mawak

Email: mawakjoseph74@gmail.com

Phone: +2348067773253

Keywords: item response theory, achievement, calibration, test development.

ABSTRACT

The study focused on development and item response theory calibration of economics achievement test for secondary school students in Plateau State, Nigeria. What motivated the study was the persistent poor performance of students in external examination in the subject which could be attributed to the quality of teacher-made test that is used in assessing students' achievement in the subject. Apart from the quality of teacher-made test that is used in assessing students' achievement, some teacher-made test used by teachers are developed using classical test theory with its attendant shortcoming of being sample dependant and test dependant. Again there is no standardized achievement test that can be used by teachers for continuous assessment of students' achievement in the subject in the study area and hence the need for this study. The designs of the study were instrumentation and cross-sectional survey research design. The population of the study consisted of all the 23712 SS 2 and a sample of 1454 from the population economics students was used for the study. Multistage sampling technique was used for the study. Multistage sampling technique was used to ensure that adequate number of students was selected from each zone, local government, schools and students. A sample of 134 schools, made up of 74 private schools and 60 public schools, 68 from rural and 66 from urban, 720 males and 734 were selected. The instrument for data collection was Economics Achievement Test (EAT) developed and calibrated by the researcher. Six research questions were raised to guide the study. The research questions were answered using difficulty indices, discrimination indices and guessing parameter indices. The validity of the instrument was established using test blue print and the judgement of experts from Economics Education and Research, Measurement and Evaluation. The reliability of the instrument was established using omega reliability procedure and it was found to be 0.83. The results of the findings indicate that the test was valid, reliable and a multidimensional test. That had moderate difficult items. The test was recommended to be used by teachers in conducting continue assessment of economics students in Plateau State Nigeria.

INTRODUCTION

Economics is a social science subject that is concerned with the study of human behaviour as a relationship between ends and scarce resources which have alternative uses. It is the study of how the society manages scarce resources, such as food, clothing and housing among others, and how human relates with these scarce resources. It is also concerned with choice because humans are faced with the problem of scarcity of resources with which to satisfy their wants and hence they are forced to choose which want to satisfy first and which want to satisfy last (Anyawuchi, 2008). The study of economics enables an individual to understand how humans behave when the price of a commodity is high and when it is low. The study of this subject is critical if Nigeria is to find itself among the top 20 world economics in the year 2030 Federal Republic of Nigeria (2008). Economics also helps in the process of production distribution and consumption of scarce resources.

The economics curriculum at the secondary school level is assessed using teacher-made test and standardized examination such as West African Examination Council (WAEC) and National Examination Council (NECO). Teacher-made test are constructed by the teachers and they are used in conducting continuous assessment in schools while the standardized examination are constructed by the examination bodies and are used in conducting certificate examination for students who graduate from the secondary schools. There is criticism that teacher-made test are poorly constructed and that results obtained from such poorly constructed tests are not valid and reliable (Wakjissa, 2010). There is therefore the need to develop a quality teacher-made achievement test for used in conducting continuous assessment of students in the subject so as to improve students' performance. Some reasons that could be held responsible for students' poor performance in the subject include teachers' poor knowledge of test construction skills which results in poor quality teacher-made economics achievement test EAT that are used in assessing students' achievement in the classroom, teachers' attitude, students' attitude, commitment and teacher's qualification among others.

Researchers Osabede (2013), Osabede (2014), Adedoyin and Adegoke (2014) have shown that there is paucity of valid and reliable economics achievement test for used in conducting continuous assessment in secondary schools with emphasis on feedback. Achievement test serves two main purposes in schools; formative and summative (Ugodulunwa, 2020). Formative assessment is used during the course of teaching for monitoring students' learning progress and providing feedback to teachers and students which is referred to as assessment for learning. While summative assessment is the type of assessment that is generally carried out at the end of a course or unit of instruction and is regard as assessment for learning. Achievement test helps teachers to identify students' areas of strength and

weakness in specific content areas. It also helps teachers to grade students.

According to Dishe (2018) and Ugodulunwa (2020) a good teacher-made test should be valid, reliable and objectively constructed. This will enable the teacher to obtain a valid and reliable result. Imo (2011) states that teacher-made test are normally prepared and administered for the purpose of finding out classroom achievement of students but unfortunately, most of the test are poorly constructed and a such results from such test are not usually valid and reliable and hence the need for teachers to develop quality tests for used in conducting continuous assessment of students.

Test development refers to the procedure involved in designing of test items by teachers or measurement experts for used in assessment (Amanda, 2014). It involves assembling test items for use to determine achievement of students in a course of study. Test development implies that some course contents area have been taught to students and the teacher is interested in finding out the achievement of students (Hamman-Tukur, Musa & Atsua, 2013). Construction of a valid and reliable test is enhanced when appropriate procedures re followed. This implies that there are a number of stages involved in developing a valid and reliable test before item calibration. These stages include; the determination of the purpose of the test, outlining the content area, development of a table of specification among others. Others are selection of appropriate item format writing the test items in line with the table of specification, field testing of items, item development of test norms and development of test manual.

According to Wakjissa (2010) Obilor & Akpan (2020) must test that are developed by teachers focused on lower levels of cognitive objectives such as knowledge and comprehension and therefore students usually do well in such examination but when confronted with items that are developed based on standard, such students usually perform poorly. Hence, the need for the teachers to consider the level of behavioural objectives in test development before the calibration of such items for use in conducting continuous assessment of students.

Achievement test are designed based on three theoretical approaches: classical test theory (CTT), generalizability theory (G-theory) and item-response theory (IRT). CTT has been the leading framework for developing and analyzing test and examination in Nigeria with its attendant advantages and disadvantages of being sample and test dependent (Hambleton & Jones, 2013). It is a theory that is concerned with the relationship between observed score, true score and error score of testees in a test.

Generalizability theory (G-theory) is a statistical theory that is used in evaluating the reliability or dependability of behavioural measure. It is a theory that is concern with persons, test items and testing situation. G-theory has it that the person administering the test, the test items and the testing situations has the potentials to affect the dependability of a test results. G-theory is an improvement on the CTT and

hence the need for teachers to adapt a modern measurement theory such as item-response theory in test development process so as to develop quality test for used in conducting continuous assessment of students.

Item-response theory is a modern measurement that is also used in developing achievement test just like the CTT and G-theory. The theory is based on the assumptions of unidimensionality, local independence and item characteristic curve. It is a theory that model, the interaction between students ability and the item difficulty, discrimination and pseudo guessing (Chalmers (2013). The theory holds that during testing, there is an encounter between testees and a test items and that the testees must have a trait level sufficient enough to be able to answer the item correctly. The focus of IRT is on the pattern of responses rather than the total score of the students.

The parameters of interest according to Amuche and Fan (2014) are difficulty, discrimination guessing and dimensionality of the test. Item difficulty or location parameter (b) is the amount of latent trait a testee must possess to be able to answer an item in a test correctly. It is the probability that a testee will be able to answer an item correctly if the student has the ability and unfortunately most teachers develop most tests without putting into consideration the difficulty parameters of the test items that are used in conducting continuous assessment of students. Hence, there is the need to consider the difficulty of test items before using the test for continuous assessment.

Item discrimination parameter (a) in test development using item-response theory indicates how well test items differentiate between individuals of different latent trait level and also indicates the differences between high and low achievers in a test. Baker (2003) opined that most achievement test by teachers do not take into consideration the need to have items that discriminates between high and low achievers and the need to discard items that do not discriminate well. Hence, there is the need for teachers to consider item discrimination in the test items that are developed for the purpose of continuous assessment.

The pseudo guessing parameter (c) is another parameter of interest which suggest that the respondent to a test item with a low trait level, my still have a small chance of answering the item correctly. It assumes that in multiple-choice test items and examinee who does not know the correct alternative may succeed in responding correctly by random guessing. As observed by Wambleton and Jones (2013) most test items that are developed by teachers are prone to guessing. This is because appropriate procedures are not usually followed in developing the test items. Hence, the need for teachers to check the guessing parameters of their items before using the items for assessment of students. Unidimensionality of a test is based on the belief that a test is supposed to measure one trait or dimension at a time but most test that are developed by teachers are usually multidimensional, this is because the items in most test usually measure more than one trait Li & Jiao (2012). A multidimensional test is a test that measure two or

more construct, ability, attribute or dimension or skills. Hence when a test violate the assumption of unidimensionality it means that the test is measuring more than one trait hence multidimensional procedure needs to be applied in calibrating the items

Item calibration is the process by which the parameters of test items are estimated (Baker, 2008). It is the scaling of test items based on their difficulty, discrimination and guessing parameter. It also involves the grouping of test items based on the ability of testees. The goal of item calibration is to develop a pool or bank of items which have the same scale and are comparable with known standard. Ojerimde Popoola, Ojo & Onyeneho (2012) found that most test items that are used for continuous assessment are not usually calibrated before they are used for the purpose of continuous assessment and hence the need for teachers to calibrate test items before using the items for continuous assessment of students.

There are several empirical studies on development and calibration of achievement tests. These include Ogbekor and Onuka (2014), Danjuma (2014) Madu (2014), Akwa (2014) and Adeleke & Joshua (2015). While Ogbekor and Onuka (2014) study was on development and validation of economics achievement attitude scale using exploratory factor analysis, Madu (2014) designed a study on the development and validation of workshop-based process skills test. The two studies found that the instruments developed were reliable. Similarly Madu (2014) conducted a study on development and validation of survey achievement test in Agricultural science which is also different from the present study that seeks to develop and calibrate and economics achievement test for secondary school students. Again, Akwa (2014) carried out a study on development and standardization of achievement test in senior secondary school mathematics and the study is different from the present study that seeks to develop and calibrate an economics achievement test for SS II economics students in Plateau State. Furthermore, Joshua and Adeleke (2015) carried out a study on the development and validation of scientific literacy achievement test to assess senior secondary school students' literacy in physics which is also different from the present work.

After reviewing literature, two limitations could be summarized from these past studies; the studies conducted are not in SS II economics and most of them used classical test theory with it attendant short coming of sample and test dependent. This therefore calls for a study on the development and IRT calibration of economics achievement test in Plateau State, Nigeria. Therefore, the broad question for this study is: what is the item parameter of the economics achievement test developed and calibrated by the researcher in Plateau State.

Research Questions

1. What is the validity of the Economics Achievement Test developed and calibrated in Plateau State?

2. What is the reliability of the Economics Achievement Test developed and calibrated in Plateau State Nigeria?
3. What is the estimate of the difficulty parameter (b) of the Economics Achievement Test items in Plateau State Nigeria?
4. What is the estimate of the discrimination parameter (a) of the Economics Achievement Test items in Plateau State Nigeria?
5. What is the estimate of pseudo guessing parameter (c) of the Economics Achievement Test items in Plateau State Nigeria?
6. What is the unidimensionality of the dichotomously scored Economics Achievement Test items in Plateau State Nigeria?

METHODOLOGY

The study used instrumentation and cross-sectional survey research designs. Instrumentation research design refers to the tool or means by which an investigator attempts to measure the variables or items of interest in a data collection process. It was used in developing and certifying the validity and reliability of the economics achievement test while cross-sectional survey was required in collecting data from the SS II students for the purpose of generalizing the finding on the entire population of SSII. The population consisted of all The 23712 SS II students in Plateau State made up of male and female while a sample of 1454 SSII students made up of 720 males and 734 females was used. Multistage sampling and proportional stratified sampling were used for the study. Multistage and proportional stratified sampling are sampling methods in which different strata in a population are identified and in which the number of elements are drawn from each strata proportionate to the relative number of elements in each stratum (Nwana, 2007). The instrument used for data collection was the economics achievement test developed and

calibrated by the researcher. Content validity of the instrument was established using table of content and by subjecting the instrument to expert judgement from Economics Education and Research, Measurement and Evaluation. The reliability of the instrument was established using Omega Reliability Procedures. The difficulty indexes, discrimination indexes guessing parameter indexes and stout's test of essentiality of item response theory were used in answering the research questions that were raised to guide the study.

Research Question One

What is the validity of the Economics Achievement Test?

The content validity of the Economics Achievement test was established using the test blue print based on the topics in the Nigerian senior secondary school economics curriculum and the objectives of each topic as contained in the curriculum. It was also subjected to experts' scrutiny. The test was adjudged to be valid based on the responses of the validators. Furthermore Kendell coefficient of concordance was computed and it was found to be 87%

Research Question Two

What is the reliability of the Economics Achievement Test?

The reliability of the Economics Achievement Test was established using Omega reliability procedure and a reliability coefficient of the dichotomous items was 0.83. This show that the EAT was reliable.

Research Question Three

What is the estimate of the difficulty parameter (b) of the Economics Achievement Test items?

Table 1: Results of Difficulty parameter Indexes of the Dichotomously Scored EAT Items

Item	Difficult	Item	Difficult	Item	Difficult	Item	Difficult
IT1	1.81	IT21	2.51	IT41	1.74	IT61	1.31
IT2	1.65	IT22	2.35	IT42	2.00	IT61	0.80
IT3	1.73	IT23	1.54	IT43	1.40	IT62	1.54
IT4	1.58	IT24	1.61	IT44	2.00	IT63	1.64
IT5	1.27	IT25	1.58	IT45	1.43	IT64	2.00
IT6	2.00	IT26	1.48	IT46	1.51	IT65	1.76
IT7	1.48	IT27	1.46	IT47	1.14		
IT8	1.44	IT28	1.36	IT48	1.42		
IT9	1.51	IT29	2.00	IT49	1.65		
IT10	2.00	IT30	2.00	IT50	1.31		
IT11	2.00	IT31	1.52	IT51	1.56		
IT12	2.00	IT32	1.53	IT52	2.00		
IT13	1.50	IT33	1.73	IT53	1.54		
IT14	1.33	IT34	1.52	IT54	1.35		
IT15	1.57	IT35	1.51	IT55	1.39		
IT16	1.40	IT36	2.00	IT56	2.00		
IT17	2.00	IT37	1.52	IT57	1.66		
IT18	1.67	IT38	1.59	IT58	1.54		
IT19	1.44	IT39	2.00	IT59	2.00		
IT20	1.75	IT40	-1.55	IT60	1.58		

The results in Table 1 show the difficulty indices of the dichotomously scored EAT items. From the results, 65 items that is (100%) had difficulty indexes that range between 0.80 to + 3. This shows that the items are good items because the difficulty indices are within -2 to $+2$ which is the benchmark for judging an item to have good difficulty index using item-

response theory analysis. The implication of this is that all the items are good items and were retained.

Research Question Four

What is the estimate of the discrimination parameters of the dichotomously scored EAT items?

Table 2: Results of Discrimination Parameter 'a' of the EAT Dichotomous Test Items

Item	ID	Item	ID	Item	ID	Item	ID
IT1	1.21	IT21	1.11	IT41	1.87	IT61	1.31
IT2	0.91	IT22	1.21	IT42	1.88	IT62	1.22
IT3	0.93	IT23	1.17	IT43	0.83	IT63	1.45
IT4	0.95	IT24	1.05	IT44	1.47	IT64	1.56
IT5	1.11	IT25	0.73	IT45	1.52	IT65	1.45
IT6	1.21	IT26	1.67	IT46	1.50		
IT7	1.34	IT27	1.14	IT47	1.52		
IT8	2.00	IT28	1.33	IT48	1.36		
IT9	1.34	IT29	1.34	IT49	1.344		
IT10	1.44	IT30	1.50	IT50	1.53		
IT11	1.45	IT31	1.36	IT51	0.93		
IT12	1.37	IT32	1.42	IT52	1.50		
IT13	1.35	IT33	1.88	IT53	1.52		
IT14	1.05	IT34	1.43	IT54	1.47		
IT15	1.22	IT35	1.87	IT55	1.14		
IT16	1.25	IT36	1.75	IT56	1.28		
IT17	1.33	IT37	1.36	IT57	1.29		
IT18	1.46	IT38	0.85	IT58	1.45		
IT19	1.34	IT39	0.42	IT59	1.51		
IT20	1.51	IT40	1.66	IT60	1.43		

The results in Table 2 presents the discrimination indexes of the dichotomously scored EAT items. From the analysis, it is evident that all the items discriminated well among the high and low examinees that sat for the test since none of the items has indexes below the range of 0.3 to 0.7 that is

considered to be below the thresholds of the discrimination indexes of 0.8 to 2.0.

Research Question Five

What is the guessing parameter of the dichotomously scored EAT items?

Table 3: Results of Guessing Parameter c of the Dichotomously Scored EAT Items

Item	c-parameter	Item	c-parameter	Item	c-parameter	Item	c-parameter
IT1	0.18	IT21	0.21	IT41	0.20	IT61	0.05
IT2	0.21	IT22	0.18	IT42	0.15	IT62	0.15
IT3	0.18	IT23	0.17	IT43	0.16	IT63	0.25
IT4	0.16	IT24	0.17	IT44	0.12	IT64	0.17
IT5	0.18	IT25	0.15	IT45	0.19	IT65	0.21
IT6	0.09	IT26	0.14	IT46	0.13		
IT7	0.22	IT27	0.25	IT47	0.13		
IT8	0.22	IT28	0.12	IT48	0.27		
IT9	0.23	IT29	0.22	IT49	0.11		
IT10	0.12	IT30	0.19	IT50	0.21		
IT11	0.14	IT31	0.21	IT51	0.15		
IT12	0.14	IT32	0.20	IT52	0.25		
IT13	0.17	IT33	0.20	IT53	0.21		
IT14	0.09	IT34	0.15	IT54	0.71		
IT15	0.12	IT35	0.12	IT55	0.24		
IT16	0.15	IT36	0.12	IT56	0.13		
IT17	0.17	IT37	0.20	IT57	0.17		
IT18	0.19	IT38	0.12	IT58	0.09		
IT19	0.24	IT39	0.19	IT59	0.16		
IT20	0.24	IT40	0.12	IT60	0.15		

The results in Table 3 shows the guessing parameter of the EAT items. From the results, the guessing parameter ranges from 0.00 – 0.25. This therefore shows that all the items were not prone to guessing. This is because all the items had guessing parameter that were between 0.00-0.25, this is good for four-multiple-choice items because the threshold is

between 0.00 to 0.25 hence all the items were retained in the final test

Research Question Six

What is the unidimensionality of the dichotomously scored EAT item?

Table 4: Results of Stout's Test of Essentiality of Unidimensionality of EAT

Test Level	Tenable level	t-crit	p-value
2.9556	0.8546	2.0906	0.0183

Table 4 shows the results of the analysis using IRTPRO cluster procedure of DETECT statistic in DIMTEST and 30% of the responses of the testees were tested to see if it was dimensionally distinct from the remaining items in the test and the results show that unidimensionality was not tenable, that is, the items that were found to form the secondary dimension

were dimensionally distinct from the remaining items of the test ($t = 2.0906$) ($p\text{-value} < 0.05$, one-tailed); therefore, the assumption of unidimensionality was rejected. This shows that the pooled 65-item of the Economics Achievement Test violated the assumptions of unidimensionality. This means that the

test was a multidimensional test because the test measured more than one trait or dimension.

DISCUSSION OF FINDINGS

The need to have a quality economic achievement test for use by teachers in conducting continuous assessment of students is an important component of teaching and learning. The results of the analysis shows that the validity of EAT was sought through experts judgement and test blue print based on the objectives in the economic curriculum and based on the results, the test was said to be valid. This is in agreement with the views of Awotunde and Ugodulunwa (2003) and Farooq (2013) that a valid measurement instrument leads to accurate measurement and evaluation while instruments that are not valid leads to wrong decision in measurement and evaluation. Furthermore, the reliability of the test items was established using omega reliability and it was found to be 0.83. This is also in agreement with the finding by Akwa (2014) that a good instrument should have a reliability of 0.70 and above.

The result of the analysis further showed that the difficulty indexes of the EAT fall within the acceptance region of between -2 – $+2$. This is in agreement with the findings by Ojerinde, Popoola, Ojo and Onyeneho (2012) and Ojerinde (2013) that difficulty index of achievement test items should be between -2 to $+2$. This is adequate because the items will be of moderate difficulty. While items that are of low difficulty need to be discarded because even low ability testees have the chance to answer those items correctly. Furthermore, the results of the funding reveals that all the items discriminated well between high and low achievers since none of the items has less than 0.05 thresholds. This is in tandem with the finding by Isaac (2011) and Ugwu (2012) that item with low-a-value discriminate poorly over a wide range of abilities and items with discrimination value below 0.80 are not good items. The higher the a-value, the more sharply the item discriminates between examinees at the point of inflection. The results of the analysis also reveals that most of the items were not prone to guessing because the items had guessing parameters indexes that fall within the range of 0.12 – 0.25. The findings is consistent with the finding by Ojerinde (2013) and Ojerinde, Popoola, Ojo & Ariyo (2014) that proper construction of achievement test reduces the chance of low ability students responding to an item correctly, while poor construction of multiple-choice items increases the chance of low ability students responding correctly to an item even when the testees do not have the ability to answer the item correctly.

The result of the findings further showed that the assumption of unidimensionality was rejected because the pooled 65-items of the EAT violate the assumption of unidimensionality. This also confirmed the views of Doran and Kingston (2018) and Li, Jiao (2012) that when a data violate the assumption of unidimensionality, then multidimensionality should be applied in determining the psychometric properties of the instrument before the calibration of items.

CONCLUSION

Based on the findings of the study, it was concluded that teachers need to use quality teacher-made achievement test like the EAT in conducting continuous assessment of students; so as to improve the performance of students in internal and external examinations in Plateau State, Nigeria.

RECOMMENDATION

Based on the findings of the study, the following recommendations are made:

1. The economics achievement test is hereby recommended for use in conducting continuous assessment of students in economics in the study area.
2. Teachers need to establish the psychometric properties of achievement test before using the test for assessment.
3. Teachers need to use item response theory achievement test for the purpose of assessment.
4. Teachers in the study area need training in test construction procedure using IRT so as to improve the test development skills of teachers in the study area.

REFERENCES

- Adedoyin, O. O., & Adegoke, J. A. (2013). Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test items parameters. *Herald Journal of Education and General Studies*, 2(3), 107 – 114.
- Akwa, A. M. (2014). *Development and standardization of achievement test in senior secondary school mathematics using item response theory*. Retrieved 29 September, 2017 from: https://acbrary.lated.org/...
- Amanda, C. (2014). *Achievement test-definition – objectives functions and characteristics*. Retrieved 17 October 2017 from: www.nsgmed.com
- Amuche, C. I., & Fan, A. F. (2016). An assessment of item bias using differential item functioning technique in NECO biology conducted examination in Taraba State Nigeria. *American International Journal of Research in Humanities, Arts and Social Science*, ISSN (print), 2328 – 3734.
- Anyawuchi, R. A. J. (2008). *Fundamentals of economics for senior secondary schools*. Onisha, African First Publishers Limited.
- Baker, J. O. (2003). *Testing in modern classrooms*. London: Gense Allen and Urwin Ltd.

- Chalmers, R. P. (2013). Mirt: A multidimensional item response theory package for the environment. *Journal of Statistical Software*, 48(6), 1 – 29.
- Danjuma, A. O. (2014). Development and validation of workshop based process skills test in metal grinding for assessing students in technical colleges for work. *Journal of Education and Practice*, 5(25), 768 – 920. Retrieved 14 June, 2017 from: www.iiste.org.
- Dishe, M. (2018) teacher-made test, meaning, features and uses/statistics retrieved on the 3rd July 2018 from www.yourarticlelibrary.com/statistics-2/teacher-made-test.htm
- Doran, N. J., Kingston, N. M., (2018). The effect of violation of unidimensionality on estimation of items and ability parameters on item response theory equating of GRE verbal scale, *Journal of Education* 22(4)22-45.
- Farooq, U. (2013). *Educational measurement, definition and concepts*. Retrieved 13 June, 2017 from: www.studylecturenotes.com.
- Federal Republic of Nigeria (FRN) (2008). *National Policy on Education* (Revised 5th ed.). Lagos: NERDC Press.
- Hambleton, R. K., & Jones, R. W. (2013). *Comparison of classical test theory and their applications to test development*. Retrieved 13 March, 2017 from: <http://ncme.org/linksevid/6696808013>.
- Hamman-Tukur, A., Musa, U., & Atsua, T. (2013). Development and validation of secondary school ranking parameters for quality assurance. *Nigerian Journal of Educational Research and Evaluation*, 12(special edition), 224 – 233.
- Hamafyelto, .R. S., Hmman-tukur, A & Hamafyelto, S. S. (2015). *Assessing teachers' competence in test construction and content validity of teacher made examination questions in commerce in Borno State Nigeria*. Retrieved 17 March, 2017 from: www.article.sapub.org
- Imo, G. C. (2011). *Effect of training in test construction on quality of teacher-made tests and study physics achievement in secondary schools in Plateau State*. Unpublished doctoral dissertation, University of Jos.
- Isaac, O. (2011). *Development and validation of psycho-productive skills multiple choice test items in agricultural science for students in secondary schools*. Retrieved 16 March 2017 from: www.unn.edu.ng.
- Joshua, E. O., & Adeleke, A. A. (2015). Development and validation of scientific literacy achievement test to assess senior secondary school students' literacy acquisition in physics. *Journal of Education and Practice*, 6(7), 81 – 96.
- Li, Y., & Jiao, W. S. & (2012). Applying multidimensional item response theory models in validating test dimensionality. *Journal of Applied Testing*, 13(2) 22 – 37.
- Madu, A. O. (2014). Development and validation of a survey achievement test in agricultural science for senior secondary schools in Plateau State. *Journal of Sustainable Agriculture and Environment*, 15 (1), 95 – 102.
- Nwana, O. C. (2007). *Introduction to educational research*. Ibadan: HEBN Publishers Plc.
- Ojerinde, D., Popoola, K., Ojo, F., & Onyeneho, P. (2012). *Introduction to item response theory, parameter models estimation and application*. Abuja: Marvelous Mike Press Ltd.
- Obilor, E. I & Akpan, U. T. (2020) Construction and Validation of Basic science achievement test for junior secondary three (JS3) students in public secondary schools in Akwa Ibom State.
- Ojerinde, D. (2013). *Classical test theory (CTT) vs. item response theory (IRT): An evaluation of the comparability of item analysis results*. Retrieved 18 March, 2017 from: UI.ed.ng.
- Ogbebor, U., & Onuka, A. (2013). Differential item functioning method as an item bias indicator. *International Journal Research*, 4(4), 367 – 373.
- Osadebe, P. U. (2014). Construction of economics achievement test for assessment of students. *World Journal of Education*, 4(2), 56-63.
- Ugodulunwa, C. A. (2020). *Fundamentals of educational measurement*. Jos: Fab. Educational Books, Nigeria.
- Ugwu, S. N. (2012). *Development and validation of criteria referenced achievement test in biology*. Retrieved 16 September, 2017 from: www.xandonline.com.
- Wakjissa, S. G. (2010). Appraisal of senior secondary 2 geography teachers' competency in assessing students Blooms levels of cognitive objectives in plateau state Nigeria. *Journal of Assessment in African*, 5(1), 177 – 188.