ASSEREN
Journal of Educational Research and Development
VOLUME 10, SEPTEMBER 2023
ISSN: 2536-6899 E – ISSN: 2814 – 3248 Website: www.theasseren.org.ng

ASSESSMENT OF ABILITY ESTIMATE AND MODEL FIT OF STUDENTS IN 2020 NECO MATHEMATICS IN BENUE STATE, NIGERIA

Onuh Omale, Gloria Adaku Dike & C. A. Chibundum

Submitted: 5 November, 2022 Revised: 3 July, 2023 Accepted: 5 July, 2023

Abstract

The basic aim of test development is to construct a test of desired quality by choosing the appropriate items through item analysis and ensuring their reliability and validity. In developing quality test items to effectively measure students' achievement, it is pertinent that the best practices in test construction be employed by NECO. The study was guided by two research questions. The study adopts a non-experimental design. The study was carried out in the Benue State of Nigeria. The population for the study comprises 18,252 Senior Secondary School three (SSS3) student who registered and sat for the NECO Mathematics Examinations in 2020. The sample for this study consists of 1,825 students out of the 18,252 that registered and sat 'for that Mathematics examination. The sample size was arrived at by taking 10% of the population. Data collected was analyzed using Ability Estimate and model fit statistics build in jmetrik for research questions, at 0.05 level of significance. The findings revealed that students' ability ranges from 0.03 above while the data was fitted into 3prarmeter logistic model Based on the results of this study it was concluded that the test items are within the acceptable ranges of ability estimate and fitted to 3-PLM can be utilized in comparing students' latent abilities for sound educational decision in our schools. And it was recommended that examination bodies, researchers that wish to use IRT in solving measurement problems.

Key word: Assessment, Ability Estimate, Model Fit, NECO, Mathematics

Introduction

Education, in its broadest sense, is a process designed to inculcate knowledge, skills, and attitudes necessary to enable individuals to cope effectively with their environment. The primary purpose of education is to foster and promote the fullest individual self-realization for all people. Education is the best legacy a nation can give to its citizens, especially the youths. This is because education is very important in the development of any nation or community. The role of education can be seen to provide pupils with skills that will prepare them physically, mentally, and socially for the world of work. The overall objective of education is to maintain a high standard through a quality test.

Tests and examinations play important roles in learning for both students and their teachers. They are used to measure students' knowledge, intelligence, or other characteristics in a systematic way. There are many reasons why teachers give tests to students. Teachers give tests to discover the learning abilities of their students; to see how well students have learned a particular subject; some tests help the students to choose a vocation, and other tests help them to understand their personality (Azpsychology, 2010). Every student is a unique member of his/her class. Some students are good in certain aspects while others are good in other aspects. When students come into a class, the teacher also needs to know as much as possible, about what they know and how they differ. In this way, the teacher can match classroom teaching with the specific needs of each student through the test.

Test is an instrument for sampling a person's behavioural traits. Alonge in Aduloju, Obinne, and Omale (2017) see test as a technique of identifying or assessing certain human behaviour or traits which include attitude, performance, and interest. Also, a test consists of a set of uniform questions or tasks to which a student or testee is to respond independently, and the result of which can be treated in such a way as to provide a quantitative comparison of the performance in different students. (Nworgu, 2011). Test is one of the most important parameters with which society adjudges the product of her education system. Test has always been an important part of the school system that even the habitual absentees normally turn up to school and present themselves to be tested on examination days. The essence of test is to reveal the latent ability of the examinee and to make grounds for assessment across the country to be as uniform as possible which may be lacking in many measuring instruments (Olabode and Adeleke 2015). One of the primary purposes of tests in our educational system is to provide a means of measuring or evaluating a group of examinees' ability and skills that is as fair and objective as possible. Test has been fully accepted in most modern societies as the most objective method of decision-making in schools, industries, and government establishments.

Though test result is accepted to be used in most societies as one of the most objective methods of decision-making, nonetheless, the use of test has sparked off some concerns among the members of the public in recent years. These concerns have tended to erode people's faith in the power and efficacy of tests. (Anastasia and Urbina, 2006). This concern can be addressed by the use of standardized tests. They are different forms of test which include: objective, essay, among others.

Objective tests are a popular form of tests; they require candidates to choose or provide a response to a question that the correct answer is predetermined. An objective test item is one for which the scoring rules are so exhaustive and specific that they do not allow scorers to make subjective inferences or judgments, thereby any scorer that marks an item following the rules will assign the same test score (Murayama, 2009). Objective tests are known to have high reliability and predictive validity as means of evaluating learning outcomes. A major criticism of objective tests is that they expose test-takers to the correct answer among the available options. The argument is that candidates only have to recognize the correct answer and that the tests fail in engaging the kind of retrieval processes that support long-term retention (Chan, McDermott, & Roediger, 2014). It has however been demonstrated that objective tests could be designed to call upon these retrieval processes. The idea is that if the alternative answers are all plausible enough, then the test-takers would have to retrieve information about why correct alternatives are correct and also why incorrect alternatives are incorrect to distinguish between the two. Properly constructed objective tests can trigger productive retrieval processes, and even have potentially important advantages over tests in which only the question is presented (Little, Bjork, Bjork, & Angello, 2012). Objective tests help test takers to remember the information they are being tested on. Objective tests, which are of various types, can therefore be useful in ways that exercise the very retrieval processes they have been accused of by-passing. An objective test is the type of test that presents students with a highly structured task that limits their responses to supplying a word, brief phrase, number, and symbol or to select the answer from among a given number of alternatives.

Objectives test come in form of Multiple-Choice Questions (MCQ), Alternate Response Format (ARF) or True or False type, Completion Test Format (CTF), matching test format (MTF) but the focus of this study is on the multiple-choice test of item. Multiple-Choice Questions are test items that are given and the examinee is expected to pick the correct answer out of those options given. MCQ can be used to measure both simple and complex concepts. According to Case and Donahue (2008), "high-quality multiple-choice questions contain three components: the stem (a scenario or vignette setting up the question), the lead-in (the question), and the options (answer choices, typically labeled A, B, C,). The correct answer is referred to as the key and the remaining options are called distracters".

Connie (2010) outlines the rules for developing quality multiple-choice questions as; specification of the content of the test, outlining the purpose of the test, use of a table of specifications, use of simple sentence and precise wordings, placing most of the words in the question stem, making all distracters plausible, keep all answer choices in the same length, avoiding double negatives, mix up the order of the correct answers, keeping the number of options consistent and avoiding tricking test takers.

It is critical that the MCQs test is efficient and effective at measuring ability and the measurement scores are reliable and precise measures of examinee ability if carefully and well-constructed. Criteria used to establish test quality generally focus on the areas of test design, test analysis techniques, and test score Interpretation. Quality MCQ test design is impacted by many elements including format, length, administration procedures, construction, validity, and scoring schema (Kinsey in Ado, 2015). The nature and the quality of information gathered from the achievement test (MCQ) can control the educational development efforts and direct the instruction (Kimberlin & Winterstein, 2008).

Hence the quality of the NECO Mathematics test will be called to question if haphazardly constructed. Mathematics is a critical skill for all, particularly in the quest to meet up the increasing demands of technological change. The main points of view for the importance of Mathematics fall into three categories: Mathematics is a core skill for all adults in life generally; a mathematically well-educated population will contribute to the country's economic prosperity; and Mathematics is important for its own sake (Joubert, 2013). The demand for mathematical skills is increasing (Burghes, 2011; Vorderman, Porkess, Budd, Dunne, and Rahmanhart, 2011). Many reports explain why Mathematics matters; why it is important to produce young people who are good in Mathematics, and why it has become increasingly urgent that the problems with Mathematics education should be addressed.

The subject is so important that universities and other higher institutions in Nigeria require ordinary level (O level) credit pass in it for admission into most courses. Mathematics forms the bedrock of knowledge for developments in the fields of Science and Technology. Students that want to excel in Science and Technology studies should be grounded in Mathematics according to Ekanem, Ekanem, Ejue, and Amimi (2010). It is important to accurately evaluate students' understanding of the subject since assessment itself contributes to enhancing the study of Mathematics. Hence the issues of assessing data model fit of NECO examination using Item Response Theory is the focus of the study.

There are various assessment agencies such as the National Examination Council (NECO), West African Examinations Council (WAEC), National Business and Technical Examinations Board (NABTEB), Joint Admission and Matriculation Board (JAMB), which are saddled with the responsibility of implementing the objectives as stated in national policy on education (NPE 2013). Most of these examination bodies carry out psychometric analysis on their items using Classical Test Theory. Though it is assumed that JAMB has adopted the use of Item Response Theory, this is why the present study focuses on NECO. In NECO the analysis of psychometric qualities of their objective test is mostly done with Classical Test Theory (Adonu 2015). Thereafter the qualities of these items are kept as classified information and can be hardly assessed by the public, researchers, or other educational agencies. It is therefore pertinent that the psychometric properties of the examination body should be determined.

Under the Item Response Theory (IRT) framework, an important issue in the calibration of data is whether the Multiple-choice measure the same construct. It has been reported by Liu (2015) that some large-scale tests such as NECO are nearly unidimensional for the constructs that are measured, in either case, theoretical models and estimation programs are available for calibrating items. The crucial benefits of IRT models are realized to the degree that the data fit the different models, 1-, 2-, and 3 parameters. Model-data fit is a major concern when applying Item Response Theory (IRT) models to real test data.

Though there is an argument that the evaluation of fit in IRT modeling has been challenging, the use of IRT model checking and item fit statistics serve as crucial factors to effective IRT use in psychometrics for information on items and model selections. Obtaining evidence of model-data- fit when an IRT model is used to make inferences from a data set is recommended as the standards for educational and psychological testing by the American Association of Educational Research, American Psychological Association, and National Council on Measurement in Education (2014). Failure to meet this requirement invalidates the application of IRT in real data set evaluation. it is on this basis that the researcher seeks to assess the model data fit NECO 2019/2020 Mathematics since it uses the IRT model in the standardization of the items. Scholars such as Cyrinus, Idaka, and Metibemu (2017), indicated that model checking remains a major hurdle to the effective implementation of IRT in which, failure to assess item level and model-data- fit statistics in the applications of IRT models, before any inferences can be drawn from the fitted model, is capable of leading to any potentially misleading conclusions derived from poorly fitted models. (Liu and Maydeu-Olivares 2014). The need to effectively assess model-data fit is imperative for correctly choosing the right model that adequately fits the data. Studies have shown an extension beyond dichotomous IRT models to polytomous IRT models, including the generalized partial credit model and rating scale model on item fit statistics and model selection in recent times (Kang &Chen, 2011).

Wells, Wollack, and Serlin (2015) stressed that the fit of a model to the data must accurately portray the true relationship between ability and performance on the item. They held that model misfit has dire consequences leading to violation of invariance property. Thus, Kose (2014) emphasized that the property of invariance of item and ability parameters is the main crux of IRT that distinguishes it from CTT. The invariance property of item and ability is not dependent on the examinees' distribution and characteristics of the set of test items. Hence, Bolt in Cyrinus, Idaka, Metibemu. (2017) believed that test developers must establish that a particular model fits the data before operationalizing a valid item. Orlando and Thissen (2003) opined that the appropriate use of IRT models is predicated on the premise that some IRT assumptions are made about the nature of the data, to ensure that the model accurately represents the data. When these assumptions are not met, inferences regarding the nature of the items and tests can be erroneous, and the potential advantages of using IRT are not gained. Besides, Sinhary (2005) held that failure to ensure the appropriateness of model-data fit analysis carried the risk of drawing an incorrect conclusion. According to Hambleton and Swaminathan cited in McAlphine, (2012), the measure of model-data fit should be based on three types of evidence. Firstly, the validity of the assumption of the model for the data set such as unidimensionality, the test is not speeded, and guessing is minimal for 1 and 2PL, (d) also all items are of equal discrimination for 1PL. Secondly, that the expected properties are obtained to reflect; invariance of item and ability parameter estimates. In a study, Osarumwense (2019) assessed the model fit of 2016 and 2017 Biology multiple choice test items of the National Business and Technical Examination Board (NABTEB). Results from this study showed that 44, 29 and 35 items representing 88%, 58% and 70% items fitted IPL, 2PL and 3PL models for NABTEB Biology multiple choice items, while 6, 2, and 15 items representing 12%, 4% and 30% did not fit into the IRT models used for analysis, hence 1PLM was the model that fit the data in the 2016 May/June Biology test items. While Ayanwale, Adeleke and Mamadelo (2018) carried out an assessment of item statistics estimates of Basic Education Certificate Examination through Classical Test Theory (CTT) and Item Response Theory (IRT) measurement frameworks. The findings from the study showed that the items of 2017 mathematics basic education certificate paper I fitted 3 – parameter logistic model. Finding, also showed that 33 (55%) items were considered poor items which fell outside the set range of 0.20 to 0.80 for the difficult and discrimination parameter rpbis $\geq$ 0.30. Agah (2015) used three equating methods in a study to ascertain the relative efficiency of test score equating methods in comparing students' continuous assessment

measures. Findings from this study showed that, 6-items representing 15% did not fit the 3 PLM, whereas 34-items (85%) of total items fitted the 3 PLM in both states. Atsua,Uzoeshi and Oludi (2018) in a study on equating 2015 and 2016 Basic Education Certificate Examination (BECE), compared the scores of candidates of JSS III students who sat for BECE in the 2015 and 2016 sessions in Civic Education using classical test theory equating method in Ibadan North, Oyo state. The results also showed that when test items were modeled using1-PL, 2-PL and 3-PL, the smallest chi-square value was observed when the data set were modeled with 3-PL model. Findings from this study also indicated that, ability estimates of students using the two tests did not differ. Chikezie (2017) carried out a study assessing the unidimensionality of West African Senior School Certificate Examination in Chemistry with principal component analysis and Item Response Theory model. Findings from this study revealed that, the chi-square goodness of fit test showed that 94% of the items were statistically significant and do not fit the IRT 3-paramenter model.

Oku and Iweka (2018), developed and standardized a Chemistry achievement test using One-parameter logistic model (1-PLM) of IRT. The validity of the instrument was estimated using the item fit statistics, in which 74 items fitted the 1-PLM. The OKUKINS CAT on analysis yielded favourable statistics under the 1-PLM with regards to difficulty (*b*) parameter and ability estimates using the X- caliber 4.2 software. The *b* parameters ranged from - 2.417 to + 2.834 while the standard errors of measurement associated with the difficult (*b*) parameter ranged between 0.148 to 0.344. Also the ability estimates of the OKUKINS CAT ranged between - 2.276 to + 2.163 while the standard errors of measurement associated with ability estimates ranged between 0.2457 to 0.3128.

In developing quality test items to effectively measure students' achievement, it is pertinent that the best practices in test construction be employed by examination bodies. Researchers, over the years have pointed out that some best practices in an item and test analysis are too infrequently used by most examination NECO inclusive. It is expected that examinations such as NECO should be Valid, Reliable and all other psychometric properties that made up test are ensured. Since is presume that the examination body uses IRT model is calibrated and standardization of their instrument, one would expect that the item is fitted into the right IRT model that will reflect the true abilities of the test takers. This Item response theory (IRT) modeling involves fitting the responses obtained from questionnaire/test items intended to measure the educational achievement of students in NECO Examination to ascertain the discrepancy between the model and the data (i.e., the absolute goodness of fit of the model). if NECO Item is properly fitted into the right IRT model and all other assumptions of IRT are checked then it can be said that the test has met all standard practices of international and global test standardization practices, if these items are psychometric developed with IRT, it is expect that model fit of the data is checked or ascertained, also the assumption undelaying the use of IRT framework is used in the calibration of test item administered by NECO are psychometric test since it meant to review the latent ability of the student taking the examination.

Research Questions

The following research questions were raised to guide the study

1. What are the mean ability estimates of students based on mathematics test?
2. Which of the IRT model data fit do NECO 2020 Mathematics test items fit?

Methodology

The study adopts the non-experimental design of descriptive research type. The study was carried out in Benue State of Nigeria. The population for the study comprises 18,252 Senior Secondary School III (SSS3) Student who registered and sat for the NECO Mathematics Examinations in 2020. The sample for this study consists of 1,825 students out of the 18,252 that registered and sat 'for the NECO 2020 Mathematics examination. The sample size was arrived at by taking 10% of the population. This is following Borg and Gall

(cited in Emaikwu, 2015) who stated that, for a population that is up to 5,000 and above, 10% of the population is large enough to be considered a representation of the population. Data collected was analyzed using Ability Estimate and model fit statistics build in jmetrik for research questions, at 0.05 level of significance.

Results and Discussion

Research Question 1: What are the mean

Table 1: Examinees Ability Estimate Mathematics **Achievement Test**

Examinee's Number	MAT _Ability	T_score_MAT_
A1	0.22	52.16
A2	0.07	50.73
A3	0.11	51.09
A4	0.29	52.87
A5	0.08	50.76
A6	-0.14	48.58
A7	0.18	51.77
A8	0.09	50.90
A9	0.09	50.88
A10	-0.01	49.93
A11	0.11	51.08
A12	0.10	51.03
A13	-0.04	49.64
A14	-0.29	47.14
A15	-0.27	47.30
A16	-0.09	49.07
A17	0.06	50.64
A18	0.03	50.25
A19	-0.33	46.68
A20	-0.03	49.67
A21	-0.06	49.44
A22	-0.08	49.23
A23	-0.22	47.79
A24	0.04	50.42
A25	-0.25	47.52
+	+	+
+	+	+
+	+	+
+	+	+
A1821	**0.27**	**52.70**
A1822	-0.13	48.74
A1823	0.23	52.33
A1824	0.22	52.19
A1825	0.38	53.79
Mean =	**1.01**	
SD =	**0.86**	

Key:
++ Abridged ability estimate of students' scores on MATs

Table 1: reveals that the mean ability estimates of students in MAT is 1.01 with SD 0.86, while t-score represent the achievement Scores of the student.

1. Which of the IRT model data fit do NECO 2020 Mathematics test items fit? To answer this research question, the responses of examinees to the tests were subjected to test calibration using 3-PLM with jmetrik software and the results are presented in tables 2. All items fit and misfit are determined at 0.05 level of significance. An item is fit when p-value (calculated) is greater than 0.05 and not fit when p-value is less than 0.05.

Table 2: Calibration Analysis of MAT using 3 -PLM with Jmetrik Software.

Item	S-X^2	Df		Remarks	Item	S-X^2	df	P–Value	Remarks
1	63.370	33	0.001	**Misfit**	26	49.216	32	0.027	**Misfit**
2	30.347	32	0.550	Fit	27	25.994	33	0.802	Fit
3	135.755	32	0.000	**Misfit**	28	57.893	33	0.005	**Misfit**
4	44.651	33	0.060	Fit	29	26.246	33	0.607	Fit
5	41.535	32	0.121	Fit	30	84.868	33	0.000	**Misfit**
6	25.744	33	0.812	Fit	31	26.017	33	0.617	Fit
7	40.350	33	0.051	Fit	32	68.430	33	0.000	**Misfit**
8	39.514	33	0.202	Fit	33	97.070	33	0.000	**Misfit**
9	42.843	33	0.117	Fit	34	27.852	32	0.743	Fit
10	85.654	32	0.000	**Misfit**	35	44.807	32	0.066	Fit
11	108.945	33	0.000	**Misfit**	36	44.515	33	0.064	Fit
12	36.833	32	0.301	Fit	37	27.507	33	0.737	Fit
13	38.886	33	0.222	Fit	38	31.550	33	0.539	Fit
14	44.478	32	0.070	Fit	39	42.056	33	0.134	Fit
15	43.171	32	0.142	Fit	40	24.919	33	0.843	Fit
16	42.951	33	0.115	Fit	41	40.036	33	0.178	Fit
17	27.222	33	0.750	Fit	42	44.540	33	0.087	Fit
18	40.013	33	0.187	Fit	43	65.770	32	0.000	**Misfit**
19	35.256	33	0.362	Fit	44	33.718	32	0.384	Fit
20	41.328	32	0.125	Fit	45	54.221	33	0.011	**Misfit**
21	23.035	33	0.902	Fit	46	44.165	33	0.071	Fit
22	34.311	32	0.381	Fit	47	44.939	33	0.080	Fit
23	36.255	33	0.319	Fit	48	33.421	33	0.447	Fit
24	290.954	33	0.000	**Misfit**	49	32.363	33	0.449	Fit
25	37.873	33	0.236	Fit	50	43.552	33	0.104	Fit

Key: *s-x^2* = Chi–square statistic
df = degree of freedom

Table 2 shows that out of the 50 items of MAT, 38 items representing 76% fitted the 3-PLM. The table also revealed that the remaining 12 items representing 24% were statistically significant and did not fit the 3-PLM.

Discussion of Findings

The analysis of results from research question 1 revealed that the mean ability estimates of students' scores. This result is in consonance with that of Asiret and Sunbul (2016) who reported similar findings from their study. The result of the present study may be due to the administration of tests that were parallel to the students' groups that have similar ability distribution on the content measured by the test forms.

The analysis of result from research question 2 revealed that 76% of the test items were not statistically significant and thus fitted the 3-PLM. This finding is in line with those of Agah (2015) and Ayanwale *et al*. (2018), whose findings from their separate studies indicated 92%, 85% and 100% of items that fitted the 3-PLM respectively. Also the finding is in agreement with that of Eleje and Esomonu (2018), and Atsua *et al.* (2018) who used -2loglikelihood values for IRT to establish model fit. Their studies revealed 3-PLM with the lowest -2loglikelihood value which represented the model with a better fit for the test items. Also, the finding of this study is in line with that of Osarumwase (2019) whose study revealed that the NABTEB May/June 2017 Biology test items fitted the 3-PLM. The findings however, disagreed with that of Okwu and Iweka (2018), Bichi *et al* (2016), and Chikezie (2017) whose works revealed the items of their test instruments fitted the 1 and 2-PLM respectively. The result of the present study may be ascribed to the fact that most items in the two test are unidimensional thus fitting a 3-PLM. Also, that using large sample sizes in item parameter estimation might have led to significant item fit-statistics, hence the result of the present study.

Conclusion/Recommendation

Based on the results of this study it was concluded that the test items are within the acceptable ranges of ability estimate and fitted to 3-PLM can be utilized in comparing students' latent abilities for sound educational decision in our schools. And it was recommended that examination bodies, researchers that wish to use IRT in solving measurement problems especially those involving tests and scales should make efforts to conform to the IRT assumption.

References

Abonyi, S.O. (2009). Instrumentation in Behavioural Research; A Practical Approach. 2nd edition, Enugu. Fulladu Publishing Company. Pp. 33 – 65.

Adonu, I.I. (2014). Psychometric analysis of WAEC and NECO practical Physics tests using partial credit model. A Ph.D. thesis. Department of Science Education, University of Nigeria, Nsukka. Nigeria.

Aduloju, M.O., Obinne, A.D.E & Omale, O. (2017) Detection of Gender Bias in General Study Examination of University of Agriculture Makurdi Using Differential Item Functioning Techniques *African Journal of Theory and Practice of Educational Assessment* (AJTPEA) Vol 5 pp 87-100

Agah, J. (2015). Relative Efficiency of Test Scores Equating Methods in Comparison of Students' Continuous Assessment Measures. A Ph.D. Thesis. Department of Science Education, University of Nigeria, Nsukkka. Nigeria

American Educational Research Association, and National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing T.p. verso. Include index. ISBN 978-0-935302-35-6 (alk. paper)

Anastasia, A, & Urbina, S. (2006). Psychological testing. New Delhi:

Prentice Hall

Asiret, A.M. & Sunbul, A.S. (2016). Investigating Test Equating Methods in Small Samples Through Various Factors. *Educational Sciences: Theory and Practice*. 16, 647 – 668.

Atsua, T.G, Uzoeshi, V.I & Oludi, P. (2018). Equating 2015 and 2016 Basic Education Certificate Examination on Civic Education using Classical Test Theory and Item Response Theory in Oyo state, Nigeria. *Journal of Pristine, 14*(1) 2250-9593.

Ayanwale, M.A., Adeleke, J.O. & Mamadelo, T.I. (2018). An Assessment of Item Statistics Estimates of Basic Education Certificate Examination through Classical Test Theory and Item Response Theory Approach. *International Journal of Educational Research Review, 3(4)*, 55 – 67.

Azpsychology (2010). *Importance of testing in psychology and education*. Retrieved from http://www.a2zpsychology.com/articles/importance_of_testing_in_psychology.htm

Bichi, A.A., Hafiz, H. & Bello, S.A. (2016). Evaluation of North West University, Kano Post-UTME Test Items Using Item Response Theory. *International Journal of Evaluation and Research in Education (IJERE), 5(4)*, 261 – 270.

Burghes, D. (2011). International comparative study in Mathematics training:

Recommendations for initial teacher training in England. Education Trust. Retrieved https://www.nationalstemcentre.org.uk/res/documents/page/International%20comparative%20study%20in%20mathematics%20teacher%20training.pdf/

Chan, J. C., McDermott, K. B & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially non-tested material can benefit from prior testing of related material. *Journal of Experimental Psychology, 135*, 553–571

Chikezie, I. J. (2017). Assessment of Unidimensionality of West African Senior School

Certificate Examination in Chemistry with Principal Component Analysis and Item Response Theory Model. *African Journal of Theory and Practice of Educational Assessment.* vol.5, November 2017, 47-57

Cyrinus B. E., Idaka E. I., and Michael A. M.(2017) Item level diagnostics and model - data fit in item response theory (IRT) using BILOG - MG v3.0 and IRTPRO v3.0 programmes GLOBAL JOURNAL OF EDUCATIONAL RESEARCH VOL 16,: 87-94 DOI*: http://dx.doi.org/10.4314/gjedr.v16i2.2*

Ekanem, S. A., Ekanem, R. S., Ejue, J. B., & Amimi, P. B. (2010). Science and technology research for sustainable development in Africa: The imperative of education. *African Research Review, 4* (3b), 71-89.

Eleje, L.I. &Esomonu, N.P. (2018). Test of Achievement in Quantitative Economics for Secondary Schools: Construction and Validation Using Item Response Theory, *Asian Journal of Education and Training 4(1)*, 18 – 28.

Emaikwu, S. O. (2015). *Fundamentals of Research Methodology and Statistics*. Selfers Academic Press Ltd., Makurdi, Nigeria.

Federal Republic of Nigeria (2013). *National Policy on Education*. (Revised) Lagos. NERDC Press.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Netherlands: Springer.

Joubert, M. (2013). *Mathematics is important*. Retrieved from https://mathsreports.wordpress.com/overall-narrative/mathematics-is-important

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American journal of health-system pharmacy: AJHP: official journal of the American Society of Health-System Pharmacists*, *65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364.

Kose, I. A., 2014. Assessing model data fit of unidimensional item response theory in simulated data. Educational Research and Reviews, 9, (17): 642-649. Retrieved from: http/www.academicjournals.org/ERR.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23,* 1337–1344.

Liu, Y., and Maydeu-Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behav. Res.* 49, 354–371. doi: 10.1080/00273171.2014.910744

Murayama, K. (2009). *Objective Test Items.* Retrieved from https://www.education.com/reference/article/objective-test-items/.30/09/2020.

Nworgu B. G. (2011). *Differential item functioning: A critical issue in regional quality assurance.* Paper presented in NAERA conference.

Olabode, J.O. & Adeleke, J.O. (2015). Comparative analysis of item local independence of WAEC and NECO 2012 Mathematics (Objectives) test items. *ASSEREN Journal of Educational Research and Development (AJERD) 1&2,* 182-190.

Oku, K. & Iweka, F. (2018). Development, Standardization and Application of Chemistry Achievement Test using One-Parameter Logistic Model (1-PLM) of Item Response Theory (IRT). *American Journal of Educational Research, 6*(3), 1-58. Retrieved from. http://pubs.sciepublicon education/6/3/11/index.html#cor on 05/09/18.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2 : An item fit index for use with dichotomous item response theory models. Applied Psychological Measurement, 27, 289-298.

Osarumwense, H.J. (2019). Assessment of Model-fit for 2016 and 2017 Biology Multiple Choice Test Items of the National Business and Technical Examination Board. *International Journal for Innovation Education and Research, 7(4)*, 12–22.

Sinharay, S., 2005. Assessing fit of unidimensional item response theory models using a Bayesian approach. Journal Educational Measurement. 42, (4): 375-394.

Wells, C. S., Wollack, J. A and Serlin, R. C., 2015. An equivalency test for model fit. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada

Vorderman, C., Porkess, R., Budd, C., Dunne, R., & Rahman-hart, P. (2011). *A world-class mathematics education for all our young people*. London.