International Journal of Educational Research Review

An Assessment of Item Statistics Estimates of Basic Education Certificate Examination through Classical Test Theory and Item Response Theory approach

Musa Adekunle Ayanwale[1], Joshua Oluwatoyin Adeleke[2], Titilayo Iyabode Mamadelo[3]

ARTICLE INFO

ABSTRACT

Article History:
Received 25.06.2018
Received in revised form 10.08.2018
Accepted
Available online 01.10.2018

In this study, the researchers assessed comparability of item statistics of 2017 basic education certificate mathematics examination of National Examinations Council (NECO) through Classical Test Theory (CTT) and Item Response Theory (IRT) measurement frameworks. The study adopted instrumentation design. A 60-item NECO basic education certificate education mathematics objective test paper I was administered to 978 basic nine examinees, randomly selected from Osogbo and Olorunda Local Government Area, Osun State, Nigeria. The responses of the examinees to the test data were analysed using Marginal Maximum Likelihood Estimation of JMETRIK software. The result showed that the test data obey the assumption of unidimensionality of 3-parameter logistic model and Classical Test Theory measurement framework deleted more items 33 (55%) compare to IRT measurement framework 12 (20%). Also, it was observed that item statistics from the two contrasting frameworks (CTT and IRT) were not comparable. Moreover, further analysis showed that there was low correlation among the item statistics index. The implication of this is that NECO should jettison the use of Classical Test Theory and embrace utilization of Item Response Theory framework during their test development and item analysis.

Keywords:
Classical Test Theory, Item Response Theory, Item Statistics, NECO and BECE mathematics multiple choice items

INTRODUCTION

National Examinations Council (NECO) was saddled with the responsibility of administering Basic Education Certificate Examination (BECE) at the end of the first three years of secondary school in order to measure the achievement level of examinees at that stage. The examination is a veritable tool to qualify examinees to proceed to the senior secondary category. It is also as an assessment that measures the extent to which basic competencies and skills have been acquired. Examinees must possess at least a credit pass in English language, Mathematics and any other three subjects before they are admitted into senior secondary school. Mathematics, the bedrock of technological development is a compulsory subject for all secondary school students in Nigeria (Ayanwale and Adeleke, 2016).

Developed countries, such as the USA, Canada, Ireland and Germany, have highly developed mathematics education programmes at the primary, secondary and post secondary education system that makes them to record significant success in their countries (Adamu, 2007). Thus, for any developing nation such as Nigeria to advance technologically and improve its social and economic status, mathematics education at the primary, secondary and post secondary levels should be well managed. Students need to develop more interest in mathematics and have a fair grasp of the basic concepts and fundamental principles because numeracy, reasoning, thinking and problem solving skills can be demonstrated through learning and application of mathematics (Adegoke, 2013). However, the current trends in the performance of students in mathematics at basic education certificate examination administered by National Examinations Council show that examinees' performance is consistently fluctuating over the years. More importantly, Ashikhia (2010) isolated various factors that could mar examinees' performances in Mathematics. Prominent among these factors are the nature of the test items and the examinees' characteristics. The performance of an examinee on a test item can be explained by the characteristics of the item.

1 kunleayanwale@gmail.com orcid.org/0000-0001-7640-9898,International Centre for Educational Evaluation, University of Ibadan
2,Institute of Education, University of Ibadan,3 National Examinations Council, Oke-mosan, Ogun State, Nigeria

The multiple-choice of test item has proved effective in measuring knowledge and skills and should continue to be used as a valid measurement of knowledge and skills (Ayanwale, 2017). Multiple choice tests have 'problems' (questions) which are called the stem and a list of alternative responses (the correct answer is the key while the incorrect ones are called distracters). The importance of assessing examinees with multiple choice items cannot be overemphasized because of its ability to cover representative samples of the universe of the content of interest without necessarily elongating testing time. It is used to complement constructed response test, because of its objectivity in scoring the responses of the examinees. As Nigeria is striving to remain competitive in the world of technology, there is the need for testing experts to make use of new techniques in constructing multiple choice test items. Thus, these items should meet the expected psychometric properties. For any achievement test to measure what it is purported to measure, it should be valid and reliable.

There are two main frameworks through which quality test items can be achieved in educational measurements. These are Classical Test Theory (CTT) and Item Response Theory (IRT). The Classical Test Theory comprises three concepts. These are: test (observed) score, true score and error score. According to Hambleton and Jones (1993), within these three concepts, several models have been formulated, of which the central model is the "Classical test model". This model links the observed test score (X) to the sum of the two unobserved (or often called latent) variables, true score (T) and error score (E). Mathematically, the Classical test model is represented by $X = T + E$

The equation has two unknowns (the true score (T) and error score (E)), thereby making it not solvable. However, according to Hambleton and Jones (1993), the use of the Classical test model represented by $X = T + E$ is made possible by three assumptions. These are:

a. True score (T) and error score (E) are uncorrelated
b. The average error score in the population of examinees is zero and
c. Error scores on parallel tests are uncorrelated.

Thus, under the Classical Test Theory, the examinee's test score would be the sum of the scores received on all the items in the test. This, according to Tomkowickz and Wright (2007), is referred to as number-correct scoring. This method of scoring produces maximum likelihood trait estimates based on raw scores (that is, total number of correctly answered items). In this method, examinees who answer correctly the same number of items, irrespective of the items' level of difficulties and discriminations, earn the same scale score. Thus, the nature of the items parameters (that is, difficulty and discrimination levels) are not considered in the scoring of examinees' performance.

The importance of item parameters estimation in test development cannot be ignored. In fact difficulty and discrimination indices are statistics that guide test development (Ayanwale, 2017). Test developers use these statistics to identify problematic items such as those that are too easy or too difficult for examinees and items that are unrelated to the overall test score. For instance in the construction of norm-referenced tests, items with difficulty level less than 0.2 and greater than 0.8 are deleted. Similarly, items with discrimination level less than 0.2 are considered as bad items (Croker and Algina, 1986).

Item difficulty (often denoted as p) is the proportion of correct responses to a particular item while item discrimination (usually denoted by D or r_{pbis}) is the index that helps to show the extent to which an item can differentiate between high ability examinees and low ability examinees. Under the classical test theory, there are several ways of assessing item discrimination. These include: (a) finding the difference in the proportion of high achieving and low achieving students' who score the item correctly. Mathematically, discrimination index $D = P_u - P_l$ Where, P_u is the proportion of examinees in the high achieving group of students (usually the upper 27%) who answered the item correctly and $P_l$is the proportion of examinees in the low achieving group (usually the lower 27%) who answered the item correctly (Kelly, 1939), and (b) point biserial correlation between a dichotomously scored item and the scores on the total score. Although, the $D = P_u - P_l$ is being used in the development tests in Nigeria, it has attracted criticism because of its inability to account for 46% of the examinees' test scores and information (Courville, 2005; Adegoke, 2013; Metibemu, 2016).

However, considering the framework of classical test theory and its ability to produce discrimination and difficulty indices for test items in the process of test development, its ability to detect poor item as regards bias, and produce evidence of test validation would suggest that a test developed under the framework of CTT would be fair enough to assess psychological traits such as mathematics achievement of examinees. Unfortunately, it is fraught with shortcomings. Among these are what Fan (1998) summarized as circular dependency in terms of: the observed score is sample dependent and the item statistics (item difficulty and item discrimination) are sample dependent. Moreover, the difficulty and discrimination indices are not taken into consideration in the estimation of observed score. These shortcomings of classical test theory made the measurement community to shift ground to development of another test theory known as item response theory (IRT).

Item response theory attempts to model the ability of a test taker and the probability of answering an item correctly based on the pattern of responses to all the items that constitute the test. Under IRT, the primary interest is in whether an examinee gets an item correctly or not, rather than in the raw test scores (Ayanwale, 2017). The item-pattern scoring method produces maximum likelihood trait estimate based on pattern of item responses (Lord 1980; Tomkowick and Wright, 2007). More importantly, in order to effectively estimate the ability of the examinee from his/her response to a particular test items, the items parameters of the test should be taken into consideration. The values of the item parameters and ability parameters depend on the type of parameter model used. In IRT, for test items that are dichotomously scored, there are four parameter models. These are: one, two, three and four parameter logistic models. These models provide mathematical equation for the relation of the probability of correct response to ability (Baker, 2001). Each model employs one or more parameters whose numerical values define a particular item characteristic curve (ICC). The one- parameter model is also called the Rasch model. This model assumes that all items discriminate equally among the testees. It is only interested in the difficulty level of the items.

Two-parameter model: This model considers the fact that items of a test cannot discriminate equally among all the testees. It estimates two parameters, ''a" (discrimination index) and ''b" (difficulty index) but assumes that an examinee cannot answer a question correctly by guessing. It can be expressed as: $P_1(\theta) = \frac{1}{1+ e^{-a_1(\theta-b_1)}}$
Three-parameter model: This model assumes that an examinee can answer an item correctly by guessing. Hence, in addition to estimating discrimination index 'a' and difficulty index ''b", it estimates guessing index ''c". By definition, the value of 'c' does not vary as a function of the ability level. Thus, the lowest and highest ability examinees have the same probability of getting the item correct by guessing. The parameter c has a theoretical range of 0 ≤ c ≤ 1.0, but in practice, values above 0.35 are not considered acceptable, hence the range ɵ ≤ c ≤ 0.35 is usually adopted when the 3- parameter logistic model is used. It can be expressed as : $P_1(\theta) = C_1 + (1 - C_1)\frac{1}{1+ e^{-a_1(\theta-b_1)}}$

Four-parameter model or Upper asymptote: This parameter is otherwise known as carelessness. The model assumes that there are some items so difficult that even with extreme level of trait, not every examinee will respond to the item correctly (Rupp, 2009; Ojerinde, 2013).

Nevertheless, despite the robustness of IRT models, the estimation of item parameters and ability parameters can only be obtained when the items in the test meet the assumptions underlying its framework. Typically, three assumptions are made in specifying IRT models. These are unidimensionality, local item independence, and item characteristic curves (Ojerinde, Popoola, Ojo and Onyeneho, 2012). The assumption of unidimensionality states that items of a test measure one's ability. Unidimensionality means that the items measure one and only one area of knowledge or ability (Ojerinde, 2013). Local item independence, as one of the assumptions of IRT, is referred to as the probability of an examinee getting a test item correct must not be dependent on the response given to other items in the test. The issue of local item independence does not imply that item will not correlate with each other, but that the items on the examinees performance on different items on the test should be independent (Ayanwale, 2017). Another assumption is that of item response function (IRF) also referred to as item characteristic curve (ICC). It is a graphical display of student proficiency (ability) level based on the student ability level (ϴ). The graph displayed takes the form of a normal *ogive* (normal distribution curve). After the probability of giving the correct answers across different levels of ϴ are combined, the relationship between the probabilities and ϴ are thus presented as an item characteristic curve.

Despite the fact that IRT differs considerably from CTT in theory and commands some crucial theoretical advantages over CTT, many examining bodies in Africa, such as West African Examinations Council (WAEC) and National Examinations Council (NECO), still use Classical Test Theory for their item development and item analysis. Based on this premise, many researchers have empirically assessed the comparability of IRT and CTT item parameters using different data sets. Some of them include: Fan (1998); Adedoyin, Nenty and Chilisa (2008); Nukhet (2002); Courville (2005) ;MacDonald and Paunonen (2002) and Metibemu (2016). They found that both CTT and IRT statistics were very comparable. There was a high correlation between the CTT and IRT difficulties. The discriminations were large and the spread of difficulty values was small. Thus, the item discriminations are very comparable and the item characteristics have shown signs of positive indications of a relationship between them. Despite all their submissions on comparability of item statistics using the two measurement frameworks, none of their studies considered establishment of item statistics of basic education certificate examination. Therefore, there is a need to carry out a research in this area.

Situation of the Problem

Item analysis of multiple choice tests using CTT and IRT have been in existence for decades in the developed countries. In Nigeria, despite the theoretical advantages of Item Response Theory over Classical Test Theory, many examining bodies such as National Examinations Council (NECO),West African Examinations Council (WAEC) and NABTEB still operates within the confines of CTT framework during their test development and establishment of item statistics such as difficulty and discrimination parameters. Past studies reported the invariance advantage of IRT of item statistics which makes it preferable to CTT. But, are these demands enough to jettison CTT for IRT? There is a need for more in-depth research in that area. Relevant literature had shown that there was dearth of empirical studies on the use of CTT and IRT frameworks in estimating item parameters of Basic Education Certificate Mathematics Examination. Therefore, the study assessed item statistics estimates of NECO 2017 mathematics BECE using the two contemporary measurement frameworks.

Aim of the study

This study was set out to assess item statistics of 2017 mathematics basic education certificate examination through classical test theory and item response theory conducted by national examinations council. However, within the context of this goal, questions advanced for this study were answers; do 2017 NECO BECE items obey unidimensionality assumption of IRT framework, what are the item statistics of 2017 BECE mathematics items and how comparable are the items statistics of 2017 BECE using the two contrasting frameworks (CTT and IRT)?

METHOD

The research adopted instrumentation design. The sample of the study comprised of basic nine students of private schools randomly drawn from Osogbo and Olorunda Local Government Area, Osun State, Nigeria. Table 1 below presents frequency distribution of examinees demographic variables that participated in the study.

Table 1: Presents Frequency Distribution of Demographic Variables

Variables		N	%	Mean	Standard dev.
Gender	Male	425	43.5	1.57	0.496
	Female	553	56.5		
Age	10-12	596	60.9	1.39	0.488
	13-above	382	39.1		
Ethnicity	Yoruba	788	80.6	1.31	0.704
	Hausa	110	11.2		
	Ibo	51	5.2		
	Others	29	3.0		

Material

The instrument used for this study was 2017 NECO Basic Education Certificate Examination Mathematics Paper 1. It was a multiple choice examination with 5-option response format, consisting of 60 items that measures a single trait of the examinees that is knowledge of mathematics. The test items covered junior school mathematics curriculum in Nigeria with 35% of number and numeration, 30% of algebraic process, 25% of geometry and mensuration and 10% of everyday statistics. The instrument was administered to basic nine students in the sampled school by twelve research assistants. The time allowed for the mathematics question by NECO was 1 hour 20 minutes. The Guttman's L2 was used to establish the internal consistency of the instrument. The reliability coefficient was 0.799.

Data analyses

Data collected was analysed using statistical package program such as SPSS version 20, NOHARM (Ogive Harmonic Analysis - Robust Method) and Jmetrik in order to establish item parameters of the instrument. Test items were scored dichotomously (1- right, 0- wrong). Also, paired sample t- test was carried out in order to investigate whether the mean score of the item parameters differed significantly from classical test theory and item response theory frameworks. Level of significance used for the study was 0.05.

Findings

Assessment of unidimensionality assumption of 2017 National Examinations Council BECE mathematics items was done using non-linear factor analysis implemented on Normal Ogive Harmonic Analysis - Robust Method (NOHARM statistical software). It produces residual matrix in order to aid model-data fit analysis. The residual matrix establishes the difference between the observed covariances and that of the items after the model has been fitted to the data. Thus, the best condition is where the differences are zero (0). From table 2 below, it can be observed that unidimensional solution's residuals are relatively small compared to the item covariances. More so, scrutiny of the residual matrix does not disclose any large residuals. Therefore, to review the residual matrix, NOHARM provides its root mean square (RMS). The RMS is the square root of the average squared difference between the observed and predicted covariances. Thus, root mean square with small values indicates good fit. McDonald (1997) suggested that the overall measure of model-data fit may be evaluated by comparing it to four times the reciprocal of the square root of the sample size which can be expressed mathematically as :RMS criterion

$$= 4 \quad \frac{1}{\sqrt{sample\ size}}$$

For this study, the sample size was 978 and this gave RMS criterion of 0.128. Thus, if the estimated value from root mean square (RMS) residual (0.022) was significantly small compare to that of RMS criterion (0.128), you conclude that the test data is measuring only a single construct. Another measure of number of dimension is Tanaka's (1993) goodness -of- fit index (GFI). According to McDonald (1999) suggested that a GFI of 0.90 indicates an acceptable level of fit, a value of 0.95 indicates ''good fit and GFI of 1.00 indicates perfect fit. Therefore, the estimated GFI (0.9009) indicates an acceptable level of fit. In conclusion, it can be observed, based on the aforementioned indices, that one-dimension model fits the data substantially. Similarly, reliability test analysis was used to corroborate the result gotten from model-data fit using NOHARM for establishing unidimensionality, Guttman's L2 gave reliability coefficient of 0.799, and the standard error of measurement (SEM) was 3.2134 with 95% confidence interval; the reliability coefficient was between 0.781 and 0.817. This indicated that the 2017 BECE mathematics was unidimensional.

Table: 2: Residual Matrix (lower off-diagonals)

47	-0.042								
48	-0.001	0.001							
49	-0.008	0.004	0.011						
50	7.2e-6	-0.001	-0.009	-0.004					
51	-0.037	-0.043	-0.002	-0.009	-0.029				
52	0.047	0.037	-0.008	0.011	0.035	0.028			
53	-0.001	-0.008	-0.002	0.007	0.002	-0.007	-0.037		
54	0.039	0.035	0.003	0.011	0.034	0.027	-0.015	-0.042	
55	0.020	0.015	-0.001	2.3e-4	0.022	0.005	-0.014	-0.061	-0.019
56	0.028	0.022	-0.002	0.004	0.026	0.013	-0.004	-0.042	-0.002
57	0.024	0.015	-0.006	-3.6e-4	0.030	0.020	0.021	-0.009	0.015
58	-0.003	-0.008	0.006	-0.015	-0.002	-0.011	0.013	0.008	0.014
59	-0.001	-0.003	0.009	-1.2e-4	-0.003	-0.006	0.004	-0.005	0.003

60	-0.013	-0.016	-0.008	0.003	-0.007	-0.021	0.006	-0.010	-0.004
	55	56	57	58	59				
56	-0.016								
57	0.007	0.011							
58	0.002	0.008	-0.001						
59	-0.004	-0.002	-0.007	-0.031					
60	-0.008	0.003	-0.039	0.037	0.032				

Sum of squares of residuals (lower off-diagonals) = 0.8199
Root mean square of residuals (lower off-diagonals) = 0.0215
Tanaka index of goodness of fit = 0.9009

More importantly, a model fit assessment was conducted to see items that fitted the model using -2 LogLikelihood (-2LL) and it was revealed that items fitted 3-parameter logistic model. The outputs were from item analysis and IRT item calibration segment of JMETRIK software representing the Classical Test Theory (CTT) and the Item Response Theory (IRT) respectively. Table 3 below presents these statistics as follows; For the CTT statistics, ***p*** represents the item difficulty indices while ***r_{pbs}*** represents the discrimination indices using point biserial correlation. For the IRT, '***b***' represents the difficulty parameter while ***'a'*** represents the discrimination parameter.

Table 3 : Item statistics of CTT and IRT models for NECO BECE 60- mathematics items

Item number	CTT		IRT		
	P	***r_{pbs}***	***a***	***b***	***c***
1	0.08	-0.15	0.82	2.03	0.08
2	0.26	0.38	2.86	1.67	0.01
3	0.51	0.50	2.85	0.70	0.31
4	0.35	0.31	1.14	1.04	0.09
5	0.36	0.14	0.20	5.98	0.17
6	0.14	-0.11	0.82	2.93	0.14
7	0.39	0.51	0.82	0.95	0.19
8	0.44	0.19	2.84	1.86	0.36
9	0.38	0.31	4.06	1.74	0.29
10	0.31	0.14	0.82	1.63	0.30
11	0.34	0.12	0.82	1.02	0.33
12	0.40	0.06	0.82	2.83	0.30
13	0.26	0.21	5.48	2.16	0.20
14	0.41	0.03	0.82	2.22	0.11
15	0.41	0.30	7.70	1.73	0.32

16	0.30	0.17	0.82	7.94	0.30
17	0.29	0.30	2.85	1.83	0.19
18	0.31	0.63	2.71	1.01	0.10
19	0.29	0.34	2.86	1.87	0.19
20	0.33	0.22	2.84	1.88	0.24
21	0.30	-0.18	0.82	1.19	0.30
22	0.06	-0.16	0.82	2.32	0.06
23	0.42	0.66	2.77	0.36	0.10
24	0.32	0.74	2.87	0.54	0.02
25	0.38	0.49	9.64	1.73	0.29
26	0.49	0.54	2.38	-0.03	0.04
27	0.38	0.08	0.82	2.19	0.30
28	0.42	0.58	2.87	0.88	0.23
29	0.33	0.12	0.13	1.99	0.21
30	0.36	0.54	2.83	1.08	0.18
31	0.42	-0.30	0.82	2.87	0.41
32	0.31	0.73	2.72	0.77	0.05
33	0.54	0.53	2.50	0.69	0.35
34	0.43	0.08	0.30	3.35	0.26
35	0.36	0.51	1.72	1.24	0.18
36	0.35	0.36	1.07	0.99	0.06
37	0.39	0.54	2.75	0.93	0.20
38	0.35	0.71	2.84	0.63	0.08
39	0.37	0.06	0.35	5.99	0.11
40	0.41	0.68	2.14	0.31	0.04
41	0.23	-0.11	0.82	3.62	0.23
42	0.35	0.07	0.82	1.79	0.34
43	0.31	0.14	0.82	1.04	0.31
44	0.28	0.07	0.82	2.49	0.28
45	0.49	0.55	2.56	0.37	0.21
46	0.44	0.30	5.58	1.72	0.35
47	0.48	0.61	2.57	0.30	0.17
48	0.33	0.71	2.53	0.63	0.03
49	0.25	-0.19	0.82	2.52	0.25
50	0.39	0.55	2.82	0.98	0.20
51	0.21	-0.06	1.28	5.93	0.21
52	0.29	0.08	1.02	5.82	0.28
53	0.45	0.06	0.15	5.96	0.22
54	0.16	-0.08	1.54	5.97	0.17
55	0.33	0.12	0.82	2.13	0.33
56	0.32	0.05	1.53	5.97	0.32

57	0.21	-0.07	1.53	5.97	0.21
58	0.33	0.02	1.23	5.91	0.33
59	0.41	0.10	0.82	1.77	0.35
60	0.32	0.17	0.30	8.50	0.26

It was observed from Table 3 that the CTT and IRT framework gave the estimates of all the item parameters of the 60 NECO BECE mathematics items subjected to item analysis process. The results suggested that within the CTT framework, all the items were analysed while under IRT framework all the items fitted the 3- parameter logistic model. However, the researchers further set criteria for CTT framework as (0.20 ≤ p ≤ 0.80 and r_{pbs} ≥ 0.30) which are categorized as being good items under remarks column with double asterisk (**) presented in Table 4

Table 4 : Reproduced item statistics of CTT and IRT models for NECO BECE 60-MAT items

Item number	CTT			IRT			
	P	r_{pbs}	***Remarks***	***a***	***b***	***c***	***Remarks***
1	0.08	-0.15	*	0.82	2.03	0.08	**
2	0.26	0.38	**	2.86	1.67	0.01	**
3	0.51	0.50	**	2.85	0.70	0.31	**
4	0.35	0.31	**	1.14	1.04	0.09	**
5	0.36	0.14	*	0.30	5.98	0.17	*
6	0.14	-0.11	*	0.82	2.93	0.14	**
7	0.39	0.51	**	0.82	0.95	0.19	**
8	0.44	0.19	*	2.84	1.86	0.36	**
9	0.38	0.31	**	4.06	1.74	0.29	**
10	0.31	0.14	*	0.82	1.63	0.30	**
11	0.34	0.12	*	0.82	1.02	0.33	**
12	0.40	0.06	*	0.82	2.83	0.39	**
13	0.26	0.21	*	5.48	2.16	0.20	**
14	0.41	0.03	*	0.82	2.22	0.41	**
15	0.41	0.30	**	7.70	1.73	0.32	**
16	0.30	0.17	*	0.82	7.94	0.30	*
17	0.29	0.30	**	2.85	1.83	0.19	**
18	0.31	0.63	**	2.71	1.01	0.10	**
19	0.29	0.34	**	2.86	1.87	0.19	**
20	0.33	0.22	*	2.84	1.88	0.24	**
21	0.30	-0.18	*	0.82	1.19	0.30	**
22	0.06	-0.16	*	0.82	2.32	0.06	**
23	0.42	0.66	**	2.77	0.36	0.10	**
24	0.32	0.74	**	2.87	0.54	0.02	**
25	0.38	0.49	**	9.64	1.73	0.29	**

26	0.49	0.54	**	2.38	-0.03	0.04	**
27	0.38	0.08	*	0.82	2.19	0.38	**
28	0.42	0.58	**	2.87	0.88	0.23	**
29	0.33	0.12	*	0.13	1.99	0.21	**
30	0.36	0.54	**	2.83	1.08	0.18	**
31	0.42	-0.30	*	0.82	2.87	0.41	**
32	0.31	0.73	**	2.72	0.77	0.05	**
33	0.54	0.53	**	2.50	0.69	0.35	**
34	0.43	0.08	*	0.30	3.35	0.26	*
35	0.36	0.51	**	1.72	1.24	0.18	**
36	0.35	0.36	**	1.07	0.99	0.06	**
37	0.39	0.54	**	2.75	0.93	0.20	**
38	0.35	0.71	**	2.84	0.63	0.08	**
39	0.37	0.06	*	0.35	5.99	0.11	*
40	0.41	0.68	**	2.14	0.31	0.04	**
41	0.23	-0.11	*	0.82	3.62	0.23	*
42	0.35	0.07	*	0.82	1.79	0.34	**
43	0.31	0.14	*	0.82	1.04	0.31	**
44	0.28	0.07	*	0.82	2.49	0.28	**
45	0.49	0.55	**	2.56	0.37	0.21	**
46	0.44	0.30	**	5.58	1.72	0.35	**
47	0.48	0.61	**	2.57	0.30	0.17	**
48	0.33	0.71	**	2.53	0.63	0.03	**
49	0.25	-0.19	*	0.82	2.52	0.25	**
50	0.39	0.55	**	2.82	0.98	0.20	**
51	0.21	-0.06	*	1.28	5.93	0.21	*
52	0.29	0.08	*	1.02	5.82	0.28	*
53	0.45	0.06	*	0.35	5.96	0.22	*
54	0.16	-0.08	*	1.54	5.97	0.17	*
55	0.33	0.12	*	0.82	2.13	0.33	**
56	0.32	0.05	*	1.53	5.97	0.32	*
57	0.21	-0.07	*	1.53	5.97	0.21	*
58	0.33	0.02	*	1.23	5.91	0.33	*
59	0.41	0.10	*	0.82	1.77	0.40	**
60	0.32	0.17	*	0.30	8.50	0.26	*

Note: (*) indicates poor item and (**) indicates good item.

Table 4 showed items whose difficulty index and discriminating index fall outside the range of 0.20 to 0.80 and less than 0.2 were considered poor and denoted by an asterisk (*). Also, based on the criteria set for classical item statistics, thirty-three (33) items were considered poor. However, unlike the CTT where statistics of discrimination and difficulty indices are used to determine good items, IRT models are not as straight forward. Rather, the contribution of each item to the assessment model is used.

Column 8 of Table 2.2 presented items based on the information each of them contributes to the overall information supplied by the whole test. This assessment requires looking at the test information function produced for the 60- NECO BECE mathematics items Paper I. Thus, using the test information function according to De Ayala, (2009), the solid line gives the total information, while the dotted line gives the standard error for a specific ability. The test information function shows that the maximum amount of information provided by the whole 60-NECO BECE mathematics items was 21.00 at an ability level of 0.83 (that is the point at which the curve peaks). From the test information function, items whose difficulty level fall between -1.0 and 3.0 (that is points where the dotted line cross the solid line and points where the dotted line and the solid line meet the vertical line on the right hand side) were selected as good items indicated by double asterisk (**) while items whose difficulty parameter fall outside the range were poor items indicated by an asterisk (*). Thus, IRT approach deleted twelve (12) items which are considered poor items.

Furthermore, assessment of comparability of item statistics under the two measurement frameworks showed that item difficulty mean value and standard deviation under CTT approach was 0.34 (SD = 0.09), under IRT approach was 2.41 (SD = 2.07) and the mean value difference was -2.07. Also, paired-sampled t-test statistics showed that this mean difference was statistically significant (t = -7.598, df = 59, p = 0.000). While item discrimination mean value and standard deviation under CTT method was 0.25 (SD = 0.28), under IRT method was 1.97 (SD = 1.76) and the mean value difference was -1.72. Also, Paired-samples t-test statistics showed that this mean difference was statistically significant (t = -8.090, df = 59, p = 0.000).

RESULT, DISCUSSION, AND SUGGESTIONS

The present study assessed item statistics estimates of 2017 mathematics basic education certificate examination through classical test theory and item response theory frameworks. Findings of the study revealed that model-data fit assessment of 2017 NECO BECE mathematics items Paper I fits 3-parameter logistic model and measured a single trait (knowledge of mathematics). Also, further analysis showed that moderately high reliability coefficient using Guttman's L2 method indicated unidimensionality. The study also found out that the Classical Test Theory measurement framework deleted more items than the IRT measurement framework. This finding suggests that IRT framework is superior to CTT in item analysis. This is because under CTT, good items were deleted on the basis difficulty index. This could be as a result of sample dependence of test items under CTT framework. Meanwhile, the results of comparability between CTT and IRT item statistics showed that they were not comparable. This submission contradicted findings from studies carried out by researchers (Fan (1998); Adedoyin, Nenty and Chilisa (2008); Nukhet (2002); Courville (2005) ;MacDonald and Paunonen (2002) and Metibemu (2016)) that item statistics from the two contrasting frameworks are quite comparable. Based on the findings, the study concluded that CTT and IRT item statistics were not comparable. Thus, suggested that National Examinations Council should embrace the use of IRT for test development and item analysis in order to have a very reliable and valid result.

Reference

Adamu, H. A. (2007). State of learning science and mathematics in Katsina state secondary schools. A Report Submitted to the Department of Research Statistics, Katsina State Ministry of Education, 11, 8-10.

Asikhia, O. A., (2010). Students and teachers' perception of the causes of poor academic performance in Ogun state secondary schools: Implications for counseling for National development. *European Journal of Social sciences,* 13(2), 28-36.

Adedoyin, O. O., Nenty, H. J.& Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Review, 3,* 83-93.

Adegoke, B.A. (2013). Comparison of item statistics of physics achievement test using classical Test theory and item response theory frameworks. *Journal of Education and Practice,* 4(22), 87 – 96.

Ayanwale, M.A. & Adeleke, J.O. (2016). Relational analysis of personal variables and marking skills of national examinations council's examiners. *African Journal of Pedagogy*, Kampala International University College, Tanzania, 8, 25-38.

Ayanwale, M.A. (2017). Efficacy of Item Response Theory in score ranking and concurrent validity of dichotomous and polytomous response mathematics achievement test in Osun State, Nigeria. Unpublished ph.D thesis. Institute of Education. University of Ibadan.

Baker, F.B. (2001). *The basic of item response theory*. Test calibration. ERIC Clearing House on Assessment and Evaluation. University of Maryland, College Park, MD, 136-330.

Courville, T.R. (2005). An empirical comparison of item response theory and classical test theory item/person statistics. Unpublished Doctoral Thesis, Texas A & M University.

Crocker, L. & Algina, J. (1986). *Introduction to classical test and modern test theory*. New York: Holt, Rinehart and Winston.

De Ayala, R.J. (2009). *The theory and practice of item response theory*. 1st ed. New York, NY: The Guilford Press. 144-200.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational andPsychological Measurement, 58,* 357-381.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Education Measurement: Issues and Practice, 12*(3) , 38-47.

Kelly, T.L. (1939). Selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology,* 30, 17-26.

Lord, F.M. (1980). *Application of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W.J. vander Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer

McDonald, R. P. (1999). *Test theory*: A unified treatment. Mahwah, NJ: LEA Publisher.

McDonald, R.P. & Paunonen, S. (2002). A Monte Carlo Comparison of Item and Person Statistics based on Item Response Theory versus Classical Test Theory. *Journal of Educational and PsychologicalMeasurement,*62, 921-943.

Metibemu, M.A. (2016). *Comparison of classical test theory and item response theory in the development and scoring of senior secondary school physics tests in Ondo State*. Unpublished ph.D thesis. Institute of Education. University of Ibadan.

Nukhet, C. (2002). A study of raven standard progressive matrices test's item measures under classic and item response models: An empirical comparison. *Ankara University, Journal of Faculty ofEducational Science, 35*(2), 71-79.

Ojerinde, D., Popoola, K., Ojo, F. & Onyeneho, O. P. (2012). *Introduction to item response theory:Parameter models, estimation and application*. Goshen Print media Ltd

Ojerinde, D. (2013). *Classical test theory (CTT) vs item response theory (IRT): An evaluation of comparability of item analysis results.* Lecture Presentation at the Institute of Education, University of Ibadan.

Rupp, A. A. 2009. Item response theory modeling with Bilog-MG and Multilog for windows. *International Journal of Testing,* 3(4), 365-384.

Tanaka, J.S. (1993). Multifaceted conceptions of fit in structural equation models. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models*. Newbury Park, CA: Sage.

Tomkowickz, J.T. & Wright, K.R. (2007*). Investigation of the effect of test equating and scoring methods on item parameter estimates and student ability scores*. A Paper presented at the annual conference of American Educational Research Association, Chicago, April 10.