Comparative Analysis of Item Parameter Estimates of Mathematics Multiple-choice Test of the National Examination Council under Two Statistical Packages in Southwester Nigeria

Taiwo Oluwafemi AJEIGBE[1, *] and Ayobode Patricia ASOWO[2]
Department of Educational Foundations and Counselling
Faculty of Education, Obafemi Awolowo, University, Ile-Ife, Nigeria
[*]Correspondent author email: taiaje@oauife.edu.ng, ajetaiolu@gmail.com
[2]ayobodeasowo@gmail.com

Abstract

The study investigated dimension, parameters and compared the item parameters of the Mathematics Multiple-Choice test of the National Examination Council (NECO) under X-Calibre and mirt-R packages.Ex-post-facto research design was used adopted for the study. The population for the study was1,030,629 students who sat for the 2017 Mathematic Paper III of NECO in Nigeria. The Sample consisted of 76,146 (44,357 male (58.3%) and 31,788 female (41.7%) using multi-stage sampling procedure. The instrument for the study was 60 Mathematics multiple-choice questions. The students' responses dichotomously scored were extracted from the database and were calibrated for the analysis. The results showed that the questions was unidimensional under X-Calibre and mirt-R. The results also showed that the same reliability coefficient of 0.92 was recorded under the two packages. The results further showed a statistical difference between the estimated difficulties (t=-5.225; df=118; p<0.05) and discrimination (t=-3.767; df=118; p<0.05) parameters. In contrast, there was no statistical significant difference between the estimated guessing (t=-1.624; df=118; p>0.05) parameterunder X-Calibre and mirt-R. The study, therefore, concluded that the X-Calibre method is flexible but the mirt-R displayed more proficient in estimating item parameters for dichotomously scored items.

Keyword: Item, parameter estimates, Test, Multiple-choice, National Examination, Statistics

1. INTRODUCTION

Earlier studies have focused on what and how measurement experts should assess so as to provide evidence in the pattern of test scores across tests and measurement. That is why, it is important in testing, measurement and evaluation to put forward relevant test theories needed for practical purposes in test improvement and security. First of the theories was the Traditional Test Theory (TTT), which explains the overall performance of examinee based on the sum total score gained from exams. Second of the theories was Generalizability Theory (GT), that explains examinee performance by isolating the variation that may occur in the test due to the test takers, the items itself, the occasion that affected the test during administering, raters and other variable that may seem to generate error in observed scores of examinees. Lastly, the third of the theories was the Latent Trait Theory (LTT) that aims at predicting individual examinee ability levels to each items using different item parameter estimates. Although the TT and G theories had someweakness theoretically. For instance, the TTT had less assumptions and its item parameter estimates could easily be predicted than the LTT with unclear conclusion on how examinee true ability on a test is explained. Also, the G theory's assumptionon various indices of reliability is not sufficient enough for test fairness. However, the popularity of LTT kept risen from the 1950s since it fulfilled all the need on examinee ability identification, calculations and explanations made from individual behaviouron various tests as they move along the item response function curve.

Within the LTT framework, several choices of item parameter models could be applied in mathematical form of individual item response curves. The appropriateness of a particular parameter model to any set of test data may be established by conducting a suitable goodness -of-fit investigation through test dimensionality. Test dimensionality would reflect the relationship betweenexaminee observable and unobservable traitsrelated to the test described by a mathematical function. For this reason, test dimensionality provides information on whether a single dimension or more accounts for examinee item responses in a test through the appropriateness of the best model to use based on specific assumptions about the test data.

1.1 Latent Trait Theory (LTT)

Latent trait theory (LTT) is a powerful psychometric procedure used to provide evidence in the pattern of test development, delivering, analyzing and scoring examinee test scores used for practical purposes. This means that the LTT creates a scale for interpretation of score assessment with useful attributes by following a set of guiding rules for adequate measuring of examinee ability levels. LTT according to Baker (2001) primarily focused on all items on a test and whether individual examinee would be able to answer each item correctly or not.Under the LTT, persons and items are organised on the same scale and for an item to be efficient, it must be able to differentiate examinees who are placed on different points of the scale. If the power of items to differentiate between examinees fails or reduces, then the capacity of the test to discriminate between examinee ability denoted by (θ) at different point on the scale may be held constant.

In the perspective of the latent trait theory, calibration differs from one another in at least two significant ways. In a way, the significant difference between measurement models exist in the behaviour of items and parameters (difficulty or discrimination) that are structured in the test. The other significant difference between measurements models exist in terms of the response option format. The simplest LTT model is frequently called the Rasch model or one-parameter logistic model (1PL). In explaining research models, individual examinee

responses to dichotomous test that could either be right or wrong, true or false, agree or disagree is influenced by individual trait and the difficulty levels of the item. Basically in tests and measurement to apply any LTT models or software for calibration, an evident assumption of test dimensionality should be carried out to ensure test data is of adequate quality to fit any of the LTT models. There are two basic LTT models, these are:
Unidimensional LTT model, Yu, Popp, DiGangi, and Jannaasch-Pennel, (2007) enlightened that a test is unidimensional if only a dimension accounted for variations in examinees' test performance. There are one (Rasch), two, three, and four parameter models that could be used to illustrate the estimation of this type of model with other assumptions such as unidimensionality of the test, local independence and item representatives curve that underlie this model. The second model is the Multi-dimensional MTTS model which is used in modelling test data when two or more construct accounted for variations in explaining the behaviour of examinees on a test. Multidimensional data do not have score scale consistency but examinee responses could reflect ability skills such as comprehension, writing, sport, or maths. Under the multidimensional MTTST, the selection of a suitable model for calibration depend on which of the models fits the data, where existing parameters likethe M1PLM, M2PLM, M3PLM, and M4PLM may be applied in estimating examinees ability level and test item parameters.

1.1.1 Item Parameters

Item parameter estimates under the LT theory is divided into two, first is the estimation of examinee ability level while the latter is the estimation of test item parameters. The first estimation is in the form of a mathematical function that concerns a graphical representation of examinee probability of success or failure on items in the test to its ability through item response curve (IRC). The second estimation is the calibration of item parameters which is divided into difficulty, discrimination, pseudo-guessing,carelessness and distractor indices. Under the LT theory, the first technical property is the difficulty b parameter also known as item locationwhich identifies the item functions along the power scale. The second technical property is the discrimination a parameter also known as item slope,according to Baker and Kim (2004) depict how well the itemsdifferentiates between examinees of various ability levels with steeper slopes associated with greater discrimination on the scale.The third technical property is the pseudo-guessing c parameter which assesses the magnitude of probability to which low ability examinees got difficult item correctly by chance or testwiseness in a test. The fourth index is the reflection of how high ability examinee respond incorrectly to items on a test due to distraction and carelessness, this is denoted as d parameter. Finally, the fifth index is the distractor index Dthat provide information on how effective the distractors draw away examinees who do not have the required knowledge to answer an item correctly from the correct key.

Under the LTT framework, many approaches such as Xcalibre, Mplus, Raju Area Test, Multilog-Mg, Pascale, Testfact, EQsirt, mirt, ltm, Winstep/Bigsteps Irtpro, irtoys, Python, Conquest, Wingen, Winmira, Testinfo, T-Rasch, SScore, ScoreAll, RummFold, TestGraf, SAS, STATA, Lertap5, jMetrik, flexMIRT, Mplus and several other sophisticated packages have been developed to estimate various dichotomous and polytomous data sets. But all these LTT approaches rely on item parameters that have been advanced for use in order to sort, select and estimate proper items during test development procedure. This is done in establishing whether items on the test are well constructed for its intended purposes either before or after its administration on the set of examinees that requires it. For the purpose of this study, Xcalibre and mirt of R have been engaged as procedures in LTT to accurately estimate individual examinee actual responses to dichotomous test with theattempt ofestablishing whether the item parameter estimates of both methods are comparable or can be used interchangeably.

1.1.2Application of Xcalibre

According to Zhao and Hambleton (2009), Xcalibre is an advanced modern marginal maximum-likelihood estimation (MMLE) program that is sufficient in calibrating person and item parameters of the latent trait theory framework. This could perform calibration related to one, two and three-parameter logistic models of the latent trait theory (LTT) models. Xcalibre could also estimates the parameters from dichotomously and polytomously scored data using the expectation-maximization (EM) algorithm called loop to implement MMLE. Similarly, its graphical user interface is easy to use if the data is correctly formatted using the CSV formatbut not sufficient in handling larger sample size of data. According to its description by Zhao and Hambleton (2009), Xcalibre can handle up to750 items at a time, user friendly with no on-line help call for calibration.

1.1.3Application of mirt-R

Multidimensional Item Response Theory Description Analysis in R as described by Chalmers (2012) is both a unidimensional and multidimensional latent trait models under the Latent Trait Theory (LTT) paradigm. Its calibration allows the fittings of uni- and multivariate Rasch/1 to 4 parameter logistic models as well as confirmatory and exploratory item response models for polytomous nominal items, partial credit items and rating scales. The scholar's description of mirt applications also showed that the confirmatory bi-factor and two stages analyses of data is available for modeling item bundles. Likewise, multiple group analysis and mixed effects designs are available for detecting differential item functioning and modelling item and person covariates. It also allow partialcompensatory item response modelling in conjunction with other IRT models where factor scores and person parameters could be extracted with fscores. Various visualisations of the fitted models can be obtained with itemplot. Additionally, there are methods for many generics like coef, print, anova, fit, show,residuals,logLik, plot and summary for item classes returned from the model fitting functions.

Finally, mirt application of R also support the latent class models such as the DINA, DINO,and several other discrete latent variable models, including mixture and zero-inflated response models.It is also worth noting that the package strives for easy integration with the functions in plink.

2. Problem of the Statement

The efficacy of statistical package use in calibrating parameter estimates in Multiple-choice items cannot be overemphasised. To this end, many statistical packages abound, but the statistical power and strength of these packages needed to be empirical documented with reference to subjects specification. In view of this, the X-Calibre 4.2 and the mirt Rpackages are compared in calibrating parameter estimates of Mathematic Multiple-choice items for the Senior Secondary school Certificate Examination conducted by the National Examination Council, Nigeria. Hence, the study provided information on whetheritem parameter estimates under X-calibre 4.2 and mirt-R. approaches are comparable in terms of statistical strength.

3. Significant of the Study

The findings from this study will be useful for the psychometricians, statisticians, test developers, teachers, students, policy makers, examination bodies. The psychometricians and statisticians will be inform on the strength of each package in item calibration. The test developers will be able to get feedback on the parameters characteristics of the items developed and subsequently improve their item bank. Also, the issue of item fairness, validity, reliability and useability of test items across different groups of examinees would be addressed appropriately, especially on the part of the examination bodies.The policy makers would also be sure that decisions emanated from interpretations of students' scores are properly guided and be free of making decision in error(s). furthermore, the students' interest regardless of the group membership would have been taken care of in terms of providing and administering items that are consideredunbiased.

4. Research Objectives

(a) Investigate the dimension of the Mathematics Multiple-Choice Test under X-Calibre and mirt-R approaches;

(b)Estimates the parameters of the Mathematics Multiple-choice items underX-Calibre and mirt-R approaches; and

(c). Comparable the item parameter estimates of the Mathematics dichotomously scored items underX-Calibre and mirt-R approaches.

5. Research Questions

1. Does the Mathematics Multiple-Choice Test has the same dimension under X-Calibre and mirt-R approaches?

2.What are the parameter estimates of the MathematicsMultiple-choice items underX-Calibre and mirt-R approaches?

3. How comparable are theX-calibre and mirt R item parameter estimates of the Mathematics dichotomously scored items in Nigeria?

6. Research Hypotheses

1. There is no significant difference in the difficulty parameter estimate generated by the X-calibre 4.2 and mirt-R in the Mathematics Multiple-choice items.

2. There is no significant difference in the discrimination parameter estimate generated by the X-calibre 4.2 and mirt-R in the Mathematics Multiple-choice items.

3. There is no significant difference in the guessing parameter estimate generated by the X-calibre 4.2 and mirt-R in the Mathematics Multiple-choice items.

7. Methodology

The study adopted ex-post facto research design. The population for the study consisted of all 1,030,629 candidates that registered and sat for 2017 National Examinations Council Examination (NECO) in Nigeria. A sample of 76,146 cases were selected using simple random sampling technique comprising of 41.7% (31,788 female students) and 58.3% (44,357 male students). From six states in Southwestern Nigeria, three state were selected using random sampling technique. From each of the states, 10 schools were selected randomly to make a total of 30 schools used for the study. An intact class of the students, totalling 76,146 students in the selected school constituted the actual sample use in the data analysis for the study.The2017 NECO 60 multiple-choice Mathematics questions was used as the instrument for the study. The students' responses to the 60 multiple-choice Mathematics questions were extracted from the NECO database. The 60 multiple-choice Mathematics questions were score dichotomously (1 for correct option and 0 for incorrect option). Thereafter, Microsoft Excel, Notepad and SPSS were used as the preliminary packages before calibration using X-Calibre version 4.2.2.0 (Assessment Systems Corporation, 2014) and mirt-R for the data analysis. The items were calibrated with 65 loops but the actual calibration was 2500 cases for X-Calibre and 10 Ncyclesfor mirt-R

8. Results

ResearchQuestion One: How many dimensions underlies the Mathematics dichotomously scored items under X-Calibre and mirt-R frameworks? To answer this research question, the responses of the examinees that sat for the 2017 National Examinations Council (NECO) mathematics test in Nigeria was subjected to Stout's test of essential unidimensionality of LTT procedure. The result is presented in Table 1 as follow

Table 1: Dimensionality Assessment of the 60 Mathematics Dichotomously Scored Itemsunder X-Calibre and mirt-R frameworks

			Reliability Coefficient	
	Unweighted	Weighted	Mirt-R	X-Calibre 4.2
DETECT	-0.28	-0.28	0.92	0.92
ASSI	-0.31	-0.31		
RATIO	-0.28	-0.28		

Source: Own Calibration under LTT Framework

Table 1 shows the number of dimension that underlay the 2017 NECO Mathematics test among senior secondary school students in Nigeria. The criteria for adjudging essential dimensionality of a test according to Jang and Roussos (2007) and Zhang (2007) is based on the following basis: Essential Unidimensionality: .20 < DETECT < 1.00, ASSI < 0.25, Ratio < 0.36. The result showed that the 2017 Mathematics multiple-choice test was essentially unidimensional considering a maximum DETECT value = -0.28 (< .20), ASSI = -0.31 (< 0.25) and RATIO = -0.31 (< 0.36). This indicated that only one dominant dimension accounted for the variation observed in student's responses to the Mathematics test in Nigeria. To further ascertain the dimensionality of the test, thereliability coefficient of the 60 Mathematics dichotomously scored items under mirt-R and X-Calibre 4.2 approaches were 0.92 respectively. This result implies that the reliability of scores among examinees in the 60 Mathematics dichotomously scoreditemswas very high and consistent. This indicated that the 60 Mathematics dichotomously scored items under mirt-R and X-Calibre 4.2 measured what is it was intended to measure.

Research Question Two: What are the parameter estimates of the Mathematics dichotomously scored items based X-calibre 4.2 and mirt-R? To answer this research questions, the responses of the examinees that sat for the 2017 National Examinations Council (NECO) mathematics test in Nigeria was subjected to item calibrations in X-Calibre 4.2 and mirt-R procedures. The result is presented in Table 2 as follow

Table 2: Dichotomous Item Parameters of the 2017 NECO Mathematics Multiple-choice Test in Nigeria under X-Calibre4.2 and mirt-R

			X-Calibre 4.2				mirt-R	
Item ID	**N**	**a**	**b**	**c**		**a1**	**d**	**g**
1	76139	0.60	-1.53	0.03	IT1	1.00	1.69	0.03
2	76139	0.13	0.23	0.06	IT2	0.24	-0.19	0.12
3	76138	0.33	-0.30	0.60	IT3	0.32	1.16	0.21
4	76138	0.91	-0.97	0.06	IT4	1.43	1.76	0.01
5	76138	0.06	-0.79	0.30	IT5	0.06	0.56	0.06
6	76138	0.96	0.25	0.69	IT6	1.81	-0.69	0.71
7	76138	1.01	-0.69	0.09	IT7	1.54	1.44	0.04
8	76138	1.07	-0.39	0.52	IT8	1.74	0.82	0.51
9	76138	1.00	-0.49	0.07	IT9	1.46	1.09	0.01
10	76138	1.17	0.00	0.41	IT10	1.86	0.18	0.40
11	76137	1.14	-0.47	0.05	IT11	1.72	1.18	0.01
12	76136	0.93	-0.21	0.05	IT12	1.35	0.59	0.01
13	76136	3.50	2.68	0.08	IT13	-0.12	-3.14	0.03
14	76135	1.06	-0.26	0.09	IT14	1.52	0.74	0.04
15	76135	0.85	-0.16	0.07	IT15	1.21	0.51	0.01
16	76135	1.28	-0.06	0.46	IT16	1.99	0.35	0.45
17	76135	1.01	-0.76	0.06	IT17	1.60	1.59	0.01
18	76135	0.28	0.51	0.08	IT18	0.41	-0.02	0.01
19	76135	0.91	-0.53	0.70	IT19	1.62	0.85	0.69
ItEM20	76135	0.68	-1.02	0.12	IT20	1.05	1.44	0.02
itEM21	76135	0.37	0.90	0.06	IT21	0.52	-0.36	0.01
22	76134	0.92	-0.95	0.05	IT22	1.53	1.75	0.00
23	76134	0.20	0.39	0.12	IT23	0.28	0.09	0.02
24	76133	1.02	-0.31	0.34	IT24	1.66	0.68	0.33
25	76132	0.06	2.68	0.13	IT25	-0.03	-0.04	0.02
ItEM26	76130	0.70	-1.16	0.05	IT26	1.12	1.58	0.00
ItEM27	76128	0.62	-0.23	0.05	IT27	0.92	0.42	0.00
28	76127	0.40	1.50	0.10	IT28	0.49	-0.65	0.01
29	76125	0.40	1.29	0.09	IT29	0.51	-0.56	0.01
30	76121	0.92	-0.13	0.50	IT30	1.43	0.37	0.47
31	76119	1.08	-0.40	0.06	IT31	1.61	1.00	0.01
32	76118	1.23	-0.53	0.05	IT32	1.91	1.43	0.00
33	76117	1.05	-0.55	0.06	IT33	1.59	1.27	0.01

34	76115	1.09	-0.68	0.05	IT34	1.77	1.55	0.00
35	76112	1.10	-0.73	0.07	IT35	1.76	1.68	0.02
36	76108	0.99	-0.16	0.06	IT36	1.39	0.57	0.01
37	76106	1.29	0.13	0.13	IT37	1.68	0.08	0.08
38	76103	1.75	0.09	0.43	IT38	2.84	-0.04	0.42
39	76099	1.37	-0.03	0.43	IT39	2.18	0.31	0.42
40	76097	0.94	-0.21	0.05	IT40	1.37	0.58	0.00
41	76090	1.52	0.20	0.43	IT41	2.32	-0.27	0.41
42	76086	1.25	0.25	0.15	IT42	1.64	-0.14	0.11
43	76085	1.01	-0.54	0.12	IT43	1.56	1.25	0.04
44	76082	0.17	-0.18	0.13	IT44	0.27	0.28	0.02
45	76078	0.89	-0.55	0.21	IT45	1.49	1.09	0.16
46	76076	0.98	-1.01	0.07	IT46	1.66	2.02	0.01
47	76075	0.44	-0.29	0.05	IT47	0.70	0.38	0.00
48	76074	1.31	-0.04	0.38	IT48	2.11	0.29	0.36
49	76068	1.23	-0.16	0.51	IT49	1.99	0.53	0.49
50	76057	0.80	-0.20	0.28	IT50	1.21	0.46	0.23
51	75993	0.59	-1.74	0.10	IT51	1.01	1.99	0.01
52	75977	0.80	-1.51	0.08	IT52	1.40	2.38	0.01
53	75954	1.33	-0.01	0.56	IT53	2.21	0.23	0.53
54	75936	0.75	-1.11	0.08	IT54	1.31	1.73	0.01
55	75908	0.24	1.93	0.08	IT55	0.33	-0.53	0.01
56	75858	1.16	0.18	0.50	IT56	1.83	-0.09	0.46
57	75786	0.74	-0.99	0.09	IT57	1.23	1.51	0.01
58	75658	0.72	-0.99	0.11	IT58	1.16	1.46	0.02
59	75408	0.84	-0.46	0.13	IT59	1.23	0.94	0.02
60	72686	0.32	0.10	0.09	IT60	0.49	0.09	0.01

Source: Own Calibration in X-Calibre 4.2 and mirt-R

Table 2: Summary of Parameter Estimates under the Two Packages

Packages	Parameters	Low	Moderate	High
mirt-R	a1	9%(15 items)	16.67% (10 items)	68.33%(41 items)
	d	38.33%(23 items)	16.67% (10 items)	30% (18 items)
	g	73.33 % (44 items)	6.67% (4 items)	20% (12 items)
X-Calibre	a	16.67 % (10 items)	77.33% (44 items)	68.33%(41 items)
	b	10% (6 items)	10% (6 items)	80% (48 items)
	c	70% (42 items)	16.67% (10 items)	13.33% (8 items)

Table 3 showed the summary of item parameter estimates from Xcalibre (discrimination, difficulty and guessing indices of 2017 National Examinations Council Mathematics) dichotomously scored items. The results revealed that 80% (48 items) was considered difficult (indices <0.20), 10% (6 items) were considered ideal (.20 - .69), while 10% (6 items) were easy items (.70 - .90) IAR (2011). This implies that most of the items on the test is difficult for both low and high ability candidates despite their level of chance. Also, 3.33% (2 items) had high slope index (1.35 – 1.69), 6.67% (4 items) had a very high slope index (>1.70), 11.67% (7items) weregood items (0.35 – 0.64), 61.67% (37 items) were moderate items (0.65 - 1.34) while16.67% (10 items) were poor items (0.01 -0.34) IAR 2011).Also, the guessing parameter ranged from .03 to .70. This implies that low ability candidates had equal chance of getting most items on the test right. The results imply that the test could no longer differentiate between examinees with low and high abilitydue to the high level of chance.
Table 3 also showed the summary of item parameter estimates from mirt (discrimination, difficulty and guessing indices of 2017 National Examinations Council Mathematics) dichotomously scored items. The results revealed that 30% (18 items) was considered difficult (indices <0.20), 28.33% (17 items) were considered ideal (.20 - .69), 3.33% (2 items) were easy items (.70 - .90), while 38.33% (23 items) were very easy items (>.0.90) IAR (2011). This implies that most of the items on the test were easier for both low and high ability candidates. Also, 21.67% (13 items) had high slope index (1.35 – 1.69), 46.67% (28 items) had a very high slope index (>1.70), 8.33% (5 items) were good items (0.35 – 0.64), 8.33% (5 items) were moderate items (0.65 - 1.34) while 9% (15 items) were poor items (0.01 -0.34) IAR 2011). Also, the guessing parameter ranged from .04 to .71. This implies that low ability candidates had equal chance of getting most items on the test right considering that most items on the test were easy items. The results imply that the test could not differentiate between examinees that have mastered the subject concern well due to chance.

Hypotheses Testing

Hypothesis One: There is no significant difference in the difficulty parameter estimate generated by the X-calibre and mirt-R in the Mathematics Multiple-choice items. To test this hypothesis, the difficulty parameter estimates under the two X-Calibre and mirt-R procedures were subjected independent sample t-test to establish whether the parameters are significantly different from each other. The results is presented in Table 4.

Table 4: Independent Sample t-test Analysis of the Differences in the Difficulty Parameter Estimates of the Mathematics dichotomously scored Items under the two packages

Parameters	Packages	N	$\overline{X}$	SD	df	t	p	Evaluation
b	X-Calibre 4.2	60	-.186	.857	118	-5.225	.000	Significant
	Mirt-R	60	.654	.903				

Significant at p<0.05

Table 4 showed that there was a significant difference (t=-5.225; df=118; p<0.05) in difficulty parameter estiamtes of mathematics dichotomously scored items under X-calibre and mirt-R approaches (p=.000 < 0.05). This implied that the difficulty parameter estimates generated under X-Calibre 4.2 was not similar as the estimates generated under mirt-R.

Hypothesis Two: There is no significant difference in the discrimination parameter estimate generated by the X-calibre and mirt-R in the Mathematics Multiple-choice items. To test this hypothesis, the “a” and “a1” parameter estimates under the two packages (mirt-R and X-calibre 4.2) were subjected to independent sample t-test to establish whether the parameters are significantly different from each other. The results is presented in Table Five

Table 5: Independent Sample t-test Analysis of the Differences in Discrimination Parameter Estimates of the Mathematics dichotomously scored Items under the two Packages

Parameters	Packages	N	$\overline{X}$	SD	df	t	p	Evaluation
a	X-calibre 4.2	60	.891	.51254	118	-3.767	.000	Significant
	Mirt-R	60	1.292	.64430				

Significant at p<0.05

Table five showed that there was a significant difference (t=-3.767; df=118; p<0.05) in discrimination parameter estimates of the Mathematics Multiple-choice items scored dichotomously under X-Calibre 4.2and mirt-R approaches (p=.000 < 0.05). This implied that the discrimination parameter estimates generated under Xcalibre was not similar as the estimates generated under mirt-R.

Hypothesis Three: There is no significant difference in the guessing parameter estimate generated by the X-calibre and mirt-R in the Mathematics Multiple-choice items. To test this hypothesis, the “c” and “g” parameter estimates under the two packages (mirt-R and X-Calibre 4.2) were subjected to independent sample t-test to establish whether the parameters are significantly different from each other. The results is presented in Table 4.2.3

Table 6: Independent Sample t-test Analysis of the Differences in chance Parameter Estimate of the dichotomously scored Mathematics Items under the two Packages

Parameters	Packages	N	$\overline{X}$	SD	df	t	p	Evaluation
c	X-Calibre 4.2	60	-.194	.190	118	-1.624	.107	Not Significant
	Mirt-R	60	.136	.201				

Significant at p<0.05

Table six showed that there is no significant difference (t=-1.624; df=118; p>0.05) in the guessing parameter estimates’ of the Mathematics Multiple-choice items scored dichotomously under X-Calibre and mirt-R approaches (p=.107 > 0.05). This implied that the guessing parameter estimates under X-Calibre 4.2 is similar to the estimates generated under mirt-R.

9. Discussion

Findings from research question one showed that (the 2017 National Examinations Council Mathematics) dichotomously scored itemsis essentially one-dimensional considering a maximum DETECT value = -0.28 (< .20), ASSI = -0.31 and RATIO = -0.31.The results revealed that the 60 multiple-choice mathematics items fulfilled the assumption of unidimensionality as the maximum DETECT index were in line with the set of conditions used for assessing unidimensionality DETECT < .20, ASSI < .25RATIO < .36 according to Jang and Roussos (2007) and Zhang (2007).Likewise, the result showed that the reliability coefficient of examinee responses to the 60 Mathematics dichotomously scored items was high (0.92). The result imply that there was stability, constancy and credibility among examinee responses to the Mathematics dichotomously scored items; that reflected the quality of the test. Therefore, the assumption holds thatthe Mathematics dichotomously scored test measured what it was designed to measure in Nigeria.

Findings from research objective two showed that more difficult, easy and moderateitems were reflected under the Xcalibre method than how the items were under the mirt.In contrast, there were more ideal items and items with high/very high slope items under the mirt method. The study showed that the pattern of estimating item parameters under Xcalibre and mirt differs. For instance, item 60 under Xcalibre is a poor item and could discriminate well among low and high ability candidates because it was a difficult item with guessing parameter (-0.09). This indicated that low ability candidates do not have the chance of correctly answering the item right. In contrast, item 60 under mirt is a good item because it discriminates well among high and low ability candidates. This implies that low ability candidates have a slim probability of answering the item correctly due chance (0.01) while high ability candidates have the probability of answering the item correctly. From the findings under Xcalibre method based on how the items fitted the IAR (2011) criteria for making decisions. The study revealed that even at the slightest probability of guessing, low ability candidates may not have the chance of answering an item correctly. This findings specified that the mirt method was effectual in explaining individual examinee behaviour as they move from one item to the other along the scale than the Xcalibre method does.

Other findings from hypotheses one and two established that a significant difference exist between the difficulty and discrimination parameter estimates under Xcalibre 4.2 and mirt R approaches. The results imply that the item parameter estimates under the two approaches are not similar. In disparity, findings from hypothesis three revealed that the guessing parameter estimates under Xcalibre 4.2 and mirt R is similar. This shown that guessing parameter under both approaches can be used interchangeable. The findings from research question three further showed that the Xcalibre-based discrimination and difficulty parameters are not comparable with the mirt–baseddiscrimination and difficulty parameters. It showed there is a large distinction between the item discrimination and difficulty of Xcalibre and mirt approaches. This means that the Xcalibre and mirt software cannot be used to complement each other or interchangeable in interpreting students' performance in decision-making.Lastly, the findings showed that the Xcalibre- based guessing index is comparable with the mirt chance parameter. This means that the chance parameter under Xcalibre and mirt can be used to explain the probability of answering an item correctly on a scale.

10. Conclusion

Based on the findings, the study showed that the 2017 National Examinations Council dichotomously scored Mathematics items is essentially one-dimensional. The study also showed that the item parameter estimates of the 2017 National Examinations Council dichotomously scored Mathematics items under the mirt method were neither too difficult nor too easy for individual examinee than the items were under the X-Calibre 4.2 method. This indicated that low ability candidates had the chance of answering some difficult items right under mirt method than they would have answered under X-Calibre 4.2 method. Finally, the study also found that there is a statistical difference between the estimated difficulties and discrimination parameters associated to individual examinee capability at the level of .05. Especially in contrast to the others, there is no significant difference between the estimated parameters related to guessing under X-Calibre and mirt-R.The study, therefore, concluded that the Xcalibre method is flexible but the mirt of R language is more proficient in estimating item parameter estimates of dichotomously scored items. Nevertheless,both methods item parameter estimates cannot be used comparably or to complement each other.The study as a consequence recommended that examination agencies in Nigeriamust embrace the use of diverse LTT approaches in test development and item analysis to know the method that would produce the most reliable and valid result.

References

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, pp. 136-330.

Baker, F. B., and Kim, S. (2004). Item response theory: Parameter estimation technique, 2nd edn, New York, NY: Marcel Dekker.

Chalmers, R. P. (2012). A Multidimensional item response theory package for the environment. *Journal of Statistical Software, 48*(6): 1-29

Instructional Assessment Resources (2011), *'Item analysis*', Retrieved November 9, 2013 from University of Texas at Austin, Instructional Assessment Resources, IAR (Online) Available: http://www.utexas.edu/academic/ctl/assessment/iar/students/report/itemanalysis.php (October 12, 2021)

Jang, E. E., & Roussos, L. A. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based non-parametric approach. Journal of Educational Measurement, 44(1), 1-21.

R Development Core Team. (2010). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Yu, C. H., Popp, S. O., DiGangi, S. & Jannaasch-Pennel, A. (2007). Assessing unidimensionality: A comparison of Rasch modelling, parallel analysis and TETRAD. Practical Assessment, Research and Evaluation, (12)14.

Zhao, Y., and Hambleton, R. (2009) Software for IRT Analyses: Descriptions and Features. University of Massachusetts Amherst Centre for Educational Assessment Research Report No. 652.

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika*, *72*(1), 69-91. doi: 10.1007/s11336-004-1257-7