EFFICACY OF ITEM RESPONSE THEORY IN THE VALIDATION AND SCORE RANKING OF DICHOTOMOUS RESPONSE MATHEMATICS ACHIEVEMENT TEST

[1]Musa A. AYANWALE, [2]Joshua O. ADELEKE

[1]Kampala International University, UGANDA

[2]University of Ibadan, NIGERIA

Abstract. The authors investigated the efficacy of Item Response Theory in the validation and score ranking of dichotomous response Mathematics Achievement Test. The study employed scale development research type of counterbalance design. The sample consisted of 1080 senior secondary schools three from 36 schools, who were drawn randomly from the Osun East senatorial district of Osun State. Two instruments with empirical reliability of 0.94 and 0.81 were used. Data obtained were subjected to Stout's Test of Essential Unidimensionality, Chen-Thissen LD, paired sample t-test and percentile rank. The results revealed that constructed multiple-choice items fulfilled the assumptions of IRT. Paired-samples t-test for difficulty and discrimination indices of the developed MAT and the NECO test items under CTT showed that their mean difference was statistically significant (t = 2.63, df = 59, P = 0.01) and (t = 12.19, df = 59, P = 0.00) while under IRT, the same trend was observed. Also, under CTT and IRT, the ability estimates showed a statistically significant mean difference between the two tests and the IRT percentile score ranking produced distinctive ranking for testees' who have a similar score ranking under CTT. It was concluded that the IRT method was more effective than CTT in test development and scoring. Examining bodies should calibrate their multiple-choice test using IRT.

Keywords: NECO-dichotomous items, Score Ranking-dichotomous items, Item Response Theory, developed-validation of dichotomous items

Introduction

Educational institutions and organizations appear to be facing difficulty in estimating the genuine abilities and ranking of individual examinees' to be chosen for admissions, scholarship awards and jobs. Examining bodies such as National Examinations Council (NECO) and West African Examinations Council (WAEC) are expected to pilot test the assessment instrument that is (multiple-choice test items) to enable them to establish the characteristics of their tests (item parameters and person statistic) using appropriate measurement framework. Unfortunately, what most examining bodies do is to pilot test their multiple-choice items using the Classical Test Theory (CTT) method to establish the item parameters, without exploring efficient framework like Item Response Theory (IRT). Since scoring and ranking of examinees was a product of many processes, part of which is exploring the assessment procedure and items parameters using appropriate methods. The neglect of these processes leads to faulty scoring and ranking which may blur judgment on the ability of examinees. In Nigeria, national examination tests are important to calibrate grades for certification and to give indications of the quality of education, specifically for admissions into higher institutions. Students must possess at least a credit pass in mathematics, English language and any other three subjects depending on the choice of course. Adegoke (2013) defined mathematics as a brilliant vehicle for the advancement and enhancement of an individual's intellectual competence in logical thinking, spatial perception, analytical and abstract idea. No doubt, an emphasis on functional mathematics education will ensure that Nigeria as a developing country has an inevitably focused workforce to address the difficulties of the 21st century.

However, the current trends in the performance of students in mathematics at secondary school certificate examination administered by public examining bodies showed that students' performance is consistently fluctuating over the years. The continually fluctuating performance in the subject by Nigerian secondary school students keeps attracting attention from major stakeholders in the education industry. Several researchers have tried to identify factors that might be responsible for the fluctuating performance of examinees in senior secondary school mathematics and proffered possible solutions, yet the interventions did not translate to improving the performance. Research in mathematics education has not focused on how test construction procedures and scoring frameworks contribute to examinees' performances. Assessment of examinees' learning is a crucial segment of the educational procedures. The characteristics of test items examinees' respond to and the inherent trait(s) being measured also can determine what the performance would be. Test items that are measuring other constructs different from what they are designed to measure will affect the performance of students adversely, just as scoring frameworks that cannot reflect the true performance of examinees in a test will ultimately result in abnormal test scores that do not reflect examinees' actual ability. This is the gap, the study sought to fill.

There are two contemporary approaches through which quality tests can be developed and examinee test scores obtained. These are classical test theory (CTT) and item response theory (IRT). Classical test theory (CTT), the foundation of measurement theory, has been the only measurement framework available to test developers and psychometricians for decades. The role of the classical test theory framework in test development cannot be overemphasized because of its ability to detect poor items through the estimation of item statistic values (difficulty index and discrimination index). In Nigeria, many public examining bodies still operate within the CTT framework to develop and determine the quality of their multiple-choice items despite its shortcomings of circular dependency of its item parameters and person statistics. However, to overcome these challenges associated with CTT, a better technique was developed known as Item Response Theory (IRT). This is a theory of testing based on the relationship between individuals' performances on a test item and the test takers' level of performance on an overall measure of the ability that item was designed to measure (Hambleton et al., 1991). This framework has assumptions (such as dimensionality, conditional independence and correct model specification) and models to her framework. When a single latent variable is accounted for the variance observed in examinees responses to the items, is called unidimensional if otherwise, is multidimensional. However, modelling examinees' responses depends on the type of dimensionality of the test. These are one-parameter logistic (1PL), two parameters logistic (2PL), three parameters logistic (3PL) and four parameters logistic (4PL) for the unidimensional model while multidimensional one-parameter logistic (M1PL), multidimensional two parameters logistic (M2PL), multidimensional three parameters logistic (M3PL) and multidimensional four parameters logistic (M4PL) are for multidimensional model.

As a result, IRT models produce item parameters that are independent of examinee samples and person statistics that are independent of the particular set of items administered (Hambleton & Swaminathan, 1985; Fan, 1998). The invariant property of item statistics of IRT models also, makes it theoretically feasible for the framework to provide solutions to vital measurement problems that are difficult to handle within the CTT framework. However, the significance of the invariant property of IRT model parameters cannot be overemphasized because, without it, the multifaceted nature of IRT models can scarcely be justified on either theoretical or practical grounds (Fan, 1998). Based on this premise, many researchers have compared the two frameworks of item parameters and person ability estimates using different data sets. Lawson (1991); Fan (1998); Macdonald & Paunonen (2001); Courville (2004); Progar et al. (2008)

Ojerinde & Ifewulu (2012); Ojerinde (2013); Guler et al. (2014); Bichi et al. (2015) found that IRT person parameters are more invariant across different item sets. Their findings further reveal that IRT item parameters and examinees' ability estimate are empirically superior to CTT parameters, though only if the appropriate IRT model is used for modelling the data. Despite their submissions on the comparability of item and person statistics using the two measurement frameworks, none of their studies developed and validated instruments for the establishment of item and person statistics of multiple-choice test items.

More importantly, ranking of individual examinee's scores to reflect their abilities had been used to guide major decisions like an admission of students into educational institutions, an award of scholarship and selection of candidates for jobs. The ability of an individual examinee is reflected as composite scores in multiple-choice tests and constructed-response tests. Under the classical procedure of ranking of scores for an individual examinee, every item is considered an independent draw for certain distributions which are a function of the ability of the individual. This assumption of the independent draw is questionable. There are several problems in classical test theory procedure of score ranking. For instance, in a standardized test for selections for jobs, scholarship award, admission process, and cowbellpedia competition conducted by NECO, examinees are ranked according to their total scores. The approach used in the ranking of examinees' ability is based on the CTT method. Item parameters are not considered during such scoring procedure. Two candidates may have the same raw scores that led to having the same ranking. This is because an item of the test attempted by the examinees differs in psychometrics properties.

However, IRT is an attempt to rectify the problem highlighted above. It is a model-based approach that provides additional tools for measuring traits and abilities by clearly separating test items, characterized by individual item parameters (threshold, slope and chance factor) from the characteristics of examinees and giving different weights to each item (question) according to their difficulty and discrimination level. This measures the true ability of examinees.

In the work of Zaman et al. (2008), compared students ranking of scores using CTT and IRT frameworks. They found that in CTT, a student attempting a difficult question and an easy question get equal credits which is not the case under IRT. Also, they submitted that in CTT two examinees with equal raw scores have the same ranking while in IRT they have a different ranking, making it easier for policy makers to take a decision. In Nigeria, students frequently return home at the end of the school term with the same ranking and parents/guardians were not bothered why their wards get the same ranking as another child in a class. This ought to be a worry for school authorities, policymakers, school teachers, parents and students since they cannot differentiate between the abilities of examinee A from examinee B. It is, therefore, worthwhile to carry out this study to show the efficacy of IRT in the validation and score ranking of dichotomous response mathematics achievement test.

Statement of the problem

One of the objectives of educational measurement is to precisely estimate an examinee's ability and use the outcome from this measurement to make pivotal decisions about the examinee. Such a decision may be for placement, promotion, scholarship award and certification. National tests at the senior secondary school level are administered in English, mathematics and other subjects by the public examining bodies. Mathematics is a compulsory subject for all the candidates that enroll for examination at that level. The central position given to mathematics might be because of the roles it plays in the carrier pursuits of the examinees. However, the fluctuating performance of examinees in the subject continues to give stakeholders a great concern. Researchers had isolated factors that might responsible for this problem and proffered solutions but less effort had been directed at item qualities and scoring system which may affect examinees performance adversely. Researching on the efficacy of IRT in the validation and score ranking of dichotomous response mathematics achievement test would be another empirical dimension for documentation of significant roles of IRT.

Research questions

Three questions were advanced for this study. These include: (1) does Draft Multiple-choice Mathematics Achievement Test (DFT-MAT) satisfies IRT assumptions; (2) is there a significant mean difference in the item and person statistics estimate of the developed multiple-choice item (DEV$_{mc}$-MAT) and 2015 NECO-Paper III using the two contrasting frameworks; (3) how comparable are the examinees' ranking of scores in the DEV$_{mc}$-MAT using the two contrasting frameworks.

Methodology

The study employed scale development research type of counterbalance design. The population comprised of mathematics testees in Senior Secondary School III (SSS3) in all schools that had presented testees for NECO examination in the last five years in Osun State, Nigeria. Selection of samples was conducted using a simple random sampling technique to select Six (6) Local Government Areas (LGAs) from the Osun East senatorial district of Osun State, Nigeria. Moreover, six (6) co-educational public schools were drawn in each of the selected LGAs, making a total of 36 schools, from which an intact science class was used. Thus, 1080 SSS3 examinees participated in the study. Their ages ranged between 16 and 20 years with 655 (60.6%) boys and 425 (39.4%) girls respectively. Two instruments were used for data collection: Self-developed mathematics multiple-choice (DEV$_{mc}$-MAT) with empirical reliability of 0.91 and NECO MAT for the year 2015 multiple-choice with empirical reliability of 0.83. Data collected was analyzed using Stout's Test of Essential Unidimensionality, Chen-Thissen LD and paired sample t-test at 0.05 significant level.

Findings

Research Question 1: Does the Draft Multiple-choice Mathematics Achievement Test (DFT-MAT) satisfies IRT assumptions?

To answer this question, Stout's Test of essential dimensionality was conducted using DIMTEST package version 1.0. An exploratory partitioning approach of DIMTEST was used for the analysis. The null and alternative hypotheses tested by DIMTEST are given by (Stout et al., 1996). They are: H_o: $AT \cup PT$ satisfies essential unidimensionality ($d = 1$), H_1: $AT \cup PT$ fails to satisfy $d = 1$. The null hypothesis posits that the assessment subtest (AT) and partitioning subtest (PT) partitions assess the same dominant underlying dimension, while the alternative hypothesis implies that the items in the AT partition are best represented by a dimension that is distinct from that driving response to the PT items. Table 1 presents Stout's Test of Essential Unidimensionality statistics of DFT_{mc}-MAT.

Table 1. Stout's Test of Essential Dimensionality Statistic DFT_{mc}-MAT

TL	TGbar	T	P-value
7.5916	10.1538	-2.5495	0.9946

Table 1 revealed that the draft multiple-choice mathematics achievement test (DFT_{mc} -MAT) fulfilled the assumption of unidimensionality. Since the p-value is greater than 0.05, we do not reject H_o. This showed that there was only one dimension that accounted for the variation observed in examinees' responses to the draft multiple-choice mathematics achievement test. The DIMTEST analysis result was in agreement with the set condition for assessing unidimensionality by (Stout et al., 1996). That null hypothesis is rejected if the test statistic is larger than the predetermined critical value from the normal distribution. This leads to the conclusion that the AT is dimensionally distinct from PT. Otherwise; the test is viewed as essentially unidimensional.

Another assumption which tests data must fulfill is conditional independence. Chen-Thissen LD analysis was conducted using IRT-PRO Version 3.0. Chen & Thissen (1997) proposed the LD χ^2 statistic, computed by comparing the observed and expected frequencies in each of the two-way cross-tabulations between examinees' responses to each item and each of the other items. These diagnostic statistics are standardized χ^2 values (that is, they are approximately z-scores) that become large if a pair of items indicates local dependence. That is if data for that pair of items indicates a violation of the local independence assumption. Table 2 presents the Marginal fit (χ^2) and Standardized LD χ^2 statistics among the items contained in the DFT$_{mc}$-MAT.

Table 2. Marginal fit (χ^2) and Standardized LD χ^2 statistics

Item	Label	Marginal X^2	1	2	3	4	5	6	7	8	9	10
1	Q1	0.0										
2	Q2	0.0	6.0									
3	Q3	0.0	3.6	9.8								
4	Q4	0.0	2.9	9.1	8.7							
5	Q5	0.0	3.2	5.7	7.0	9.2						
+	+	+	+									
Item	Label	Marginal X^2	141	142	143	144	145	146	147	148	149	
141	Q141	0.0										
142	Q142	0.0	5.6									
143	Q143	0.0	0.8	7.8								
144	Q144	0.0	1.7	6.1	5.4							
145	Q145	0.0	5.0	3.7	2.0	2.0						
146	Q146	0.0	3.7	5.1	4.6	6.8	8.6					
147	Q147	0.0	7.1	6.9	4.9	1.9	1.9	8.6				
148	Q148	0.0	9.5	0.2	6.2	5.4	5.6	1.7	6.6			
149	Q149	0.0	2.8	0.1	9.6	4.5	2.0	9.7	8.3	4.4		
150	Q150	0.0	8.2	0.5	1.2	0.9	5.0	8.6	7.4	4.9	5.0	

Table 2 depicts Chen-Thissen LD statistics with χ^2 values larger than 2 or 3 or 4 are not considered to be large. Rather, values larger than 10 were considered large, which indicates local dependence. It can be observed from Table 2 that the number of item pairs whose LD χ^2 value was over 10 was forty-five (45). Many of the values are relatively small, indicating no evidence of local dependence, and suggesting that test data satisfies the assumption of conditional independence.

Research Question 2: Is there a significant mean difference in the item and person statistics estimate of the developed multiple-choice item (DEV_{mc}-MAT) and 2015-NECO_{mc}-MAT using the two contrasting frameworks?

To answer this research question, item parameters in the developed multiple-choice mathematics achievement test (DEV_{mc}-MAT) and 2015-NECO_{mc}-MAT were compared under CTT and IRT frameworks. Comparison of person scores was also made in the two tests. Tables 3 and 4 presented the item parameters (that is discrimination and difficulty index) and descriptive statistics (item parameters) of the 60-items DEV_{mc}-MAT and the 2015-NECO_{mc}-MAT under CTT and IRT frameworks respectively.

Table 3 showed the item parameters (difficulty and discrimination indices) of the developed multiple-choice MAT items and the NECO_{mc}-MAT test items under CTT and IRT measurements. The data were obtained from BILOG-MG and Multivariate EQSIRT output. Table 4 further depicts the mean and standard deviation values for difficulty (p) of the developed MAT and NECO test under CTT were (M = 0.51; SD = 0.14) and (M = 0.43; SD = 0.20) respectively. The mean and standard deviation for discrimination (r_{pbs}) of the developed MAT and the NECO test were (M = 0.47; SD = 0.17) and (M = 0.07; SD = 0.23) respectively. Moreover, under IRT framework, the mean and standard deviations of difficulty parameters (b) of developed MAT and NECO test were (M = 1.89; SD = 2.09) and (M = 0.26; SD = 1.44) respectively, the means and standard deviations of discrimination parameters (*a*) of developed MAT and NECO test were (M = 1.49; SD =1.01) and (M =0.38; SD = 0.24) respectively, while the means and standard deviations of pseudo-guessing parameter (*c*) of developed MAT and NECO test were (M = 0.23; SD =0.24) and (M =0.24; SD = 0.13) respectively.

Table 3. Item parameters of 60 DEV$_{mc}$-MAT and 2015-NECO$_{mc}$-MAT

Item Number	CTT				IRT					
	DEV$_{mc}$-MAT		NECO$_{mc}$-MAT		DEV$_{mc}$-MAT			NECO$_{mc}$-MAT		
	p	r_{pbs}	*p*	r_{pbs}	*b*	*a*	c	*b*	*a*	c
1	0.56	0.48	0.58	-0.35	-0.54	2.01	0.01	-	-	-
2	0.53	0.50	0.48	-0.30	-0.33	1.72	0.03	-	-	-
3	0.52	0.36	0.17	-0.11	4.89	0.12	0.40	-	-	-
4	0.34	0.67	0.25	0.02	1.03	1.71	0.06	2.16	0.32	0.20
5	0.67	0.25	0.48	0.24	3.29	3.50	0.60	0.07	0.51	0.19
6	0.60	0.43	0.53	0.19	-1.23	3.10	0.01	- 0.24	0.38	0.31
7	0.52	0.47	0.42	0.05	-0.77	2.73	0.01	0.56	0.38	0.48
+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+	+	+	+
54	0.48	0.55	0.33	-0.14	-0.12	3.49	0.01	-	-	-
55	0.37	0.60	0.15	0-.05	0.67	1.01	0.01	3.78	0.28	0.31
56	0.45	0.53	0.23	0-.24	2.33	1.29	0.29	-	-	-
57	0.54	0.44	0.57	0.09	4.08	3.79	0.42	- 0.44	0.42	0.20
58	0.41	0.58	0.65	0.19	3.98	2.06	0.26	- 0.89	0.48	0.40
59	0.42	0.54	0.53	0.17	0.23	0.45	0.01	- 0.20	0.49	0.28
60	0.45	0.62	0.65	0.53	0.02	0.63	0.01	- 0.60	0.93	0.19

These statistics explained that on the average, under the CTT framework, items of the NECO test were a little bit more difficult than the multiple-choice items of the developed MAT. Nevertheless, multiple-choice items of the developed MAT items were better at discriminating among the examinees with high and low ability. It was observed that the same pattern of statistics obtained in terms of item difficulty and discrimination of the developed multiple-choice MAT and NECO items were also observed under the IRT framework. More importantly, paired sample t-test analysis was further carried out on the developed MAT and NECO test item parameters to determine whether the two test forms can be used concurrently. Table 5 presented paired sample t-test statistics of developed MAT and NECO test items.

Table 4. Descriptive statistics of Item parameters of 60 DEV$_{mc}$-MAT and 2015- NECO$_{mc}$-MAT

Statistics	CTT				IRT					
	DEV$_{mc}$-MAT		NECO$_{mc}$-MAT		DEV$_{mc}$-MAT			NECO$_{mc}$-MAT		
	p	r_{pbs}	p	r_{pbs}	a	b	c	a	b	c
Min.	0.16	-0.13	0.03	-0.42	0.12	-1.23	0.01	0.00	-3.21	0.00
Max.	0.77	0.72	0.87	0.65	3.86	5.80	0.70	2.37	3.78	0.52
Mean	0.51	0.47	0.43	0.07	1.49	1.89	0.23	0.38	0.26	0.24
SD	0.14	0.17	0.20	0.23	1.01	2.09	0.24	0.24	1.44	0.13

Table 5 showed the mean difference value for difficulty (p) of the developed multiple-choice MAT and NECO test under CTT to be 0.08. Paired-samples t-test statistics showed that the mean difference was statistically significant (t = 2.628, df = 59, P = 0.011). The mean difference value for discrimination (r_{pbs}) of the developed MAT and NECO test was 0.39, paired-samples t-test statistics showed that the mean difference was statistically significant (t = 12.189, df = 59, P = 0.000). Furthermore, under the IRT framework, the mean difference value for difficulty parameter (b) of developed MAT and NECO test was 1.64, paired-samples t-test statistics showed that the mean difference was statistically significant (t = 4.609, df = 59, P= 0.000), the mean difference value for discrimination parameters (a) of developed MAT and NECO test was 1.10. The paired-samples t-test statistics showed that the mean difference was statistically significant (t = 8.277, df = 59, P = 0.000) while the mean difference for the pseudo-guessing parameter (*c*) of developed MAT and NECO test was 0.00 while the paired-samples t-test statistics showed that the mean difference was not statistically significant (t = 0.011, df = 59, P = 0.992) respectively. The results suggested that there was a significant difference between the two test form item parameters, that is, the instruments cannot be used concurrently. Also, figures 2,3, 4, 5 and 6 respectively presented the bivariate picture between the item parameters in the DEV$_{mc}$-MAT and 2015-NECO$_{mc}$-MAT under the two measurement frameworks.

Table 5. Paired sample test of Item parameters of DEV$_{mc}$-MAT and 2015-NECO$_{mc}$-MAT

	Paired Difference					**t**	**df**	**Sig. (2-tailed)**
	Mean	Std. Deviation	Std.Error Mean	95% confidence interval of the difference				
				Lower	Upper			
DEV$_{p}$_ctt-NECO$_{p}$_ctt	0.08	0.24	0.03	0.02	0.14	2.63	59	0.01
DEV$_{rpbs}$_ctt-NECO$_{rpbs}$_ctt	0.39	0.25	0.03	0.33	0.46	12.19	59	0.00
DEV$_{b}$_irt-NECO$_{b}$_irt	1.64	2.76	0.36	0.93	2.36	4.61	59	0.00
DEV$_{a}$_irt-NECO$_{a}$_irt	1.10	1.03	0.13	0.84	1.37	8.28	59	0.00
DEV$_{c}$_irt-NECO$_{c}$_irt	0.00	0.29	0.04	-0.08	0.08	0.01	59	0.99

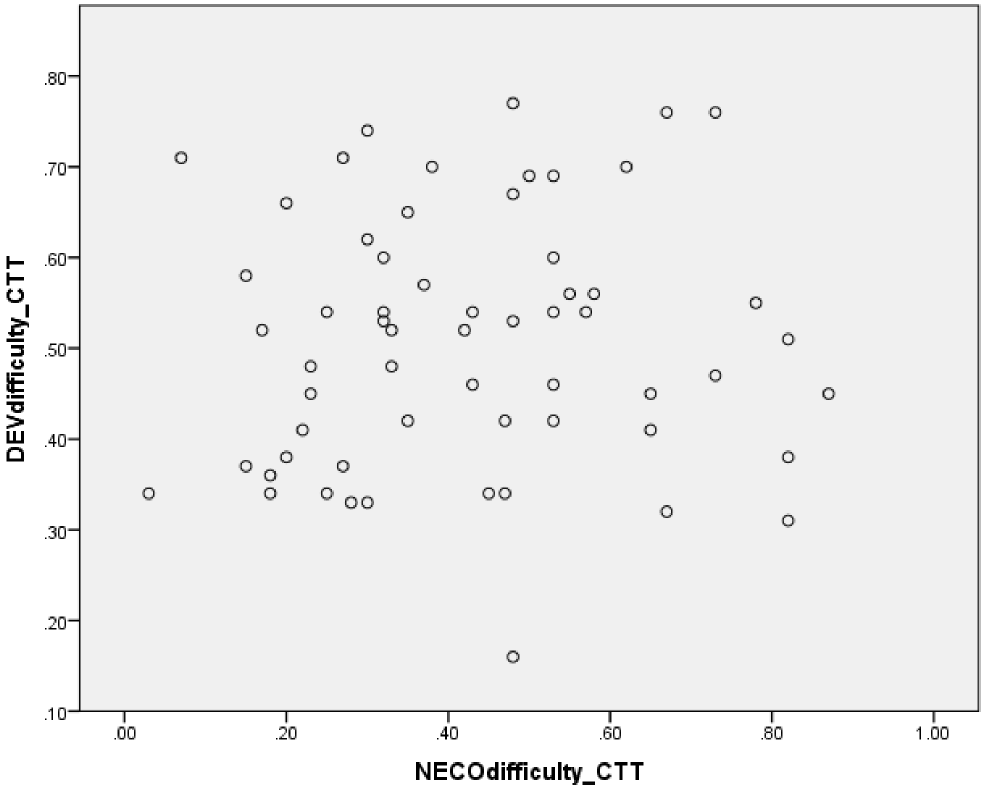

Figure 1. Scatter plot of item difficulty in the DEV$_{mc}$-MAT and 2015-NECO$_{mc}$-MAT under CTT

Fig. 1 showed that there was no significant linear relationship between the item difficulties of the two mathematics achievement instruments. The Pearson moment correlation coefficient was very low though positive but not significant (r = 0.06, p = 0.68). This implied that since the correlation between them was low, concurrent validity was not evident. This submission was in tandem with the earlier position that the item statistics were not related.

Fig. 2 showed that there was no significant linear relationship between the item discrimination in the DEV$_{mc}$-MAT and 2015 NECO paper III under CTT. The Pearson moment correlation coefficient was low and not significant (r = 0.23, p = 0.08). Thus, with the low relationship, the concurrent validity of the two instruments was not tenable.

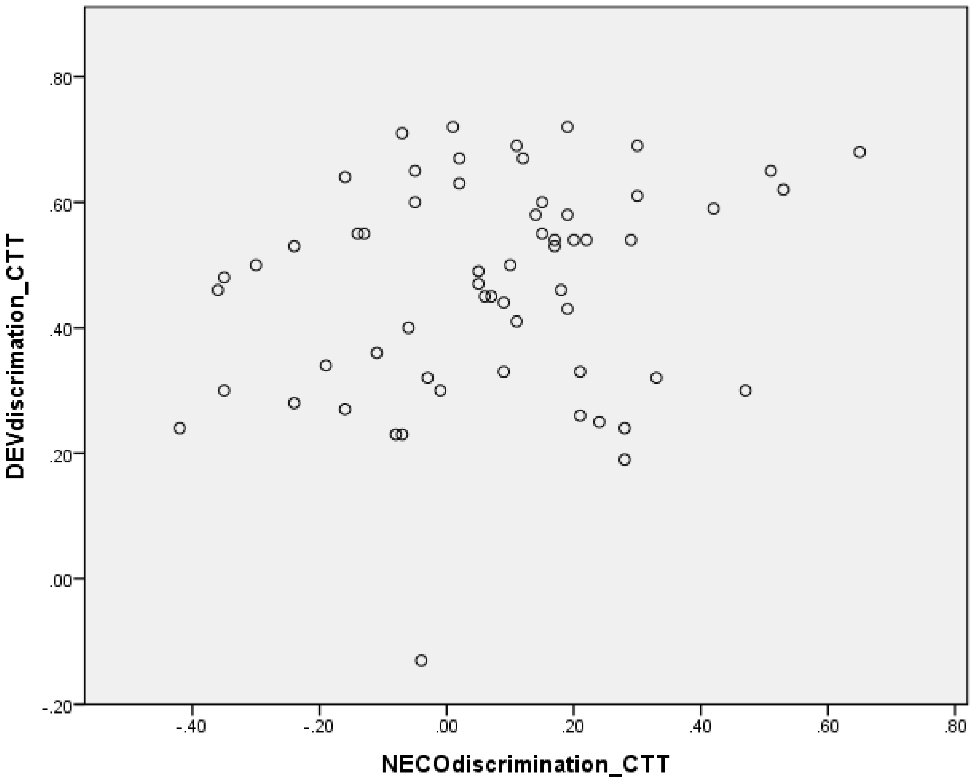

Figure 2. Scatter plot of item discrimination in the DEVmc-MAT and 2015 NECO$_{mc}$-MAT under CTT

Fig. 3 shows that there was no significant linear relationship between the difficulty parameters of the developed and 2015 NECO test items. The Pearson moment correlation coefficient was negatively low as well as not significant (r = -0.19, p = 0.14). This also corroborated earlier findings that developed-MAT and NECO test items cannot be used concurrently.

Fig. 4 shows that there was no significant linear relationship between the discrimination parameters of the two mathematics achievement instruments.

The Pearson moment correlation coefficient was low and positive but not significant (r = 0.11, p = 0.41). This indicated that since the correlation between them was low, concurrent validity was not evident.

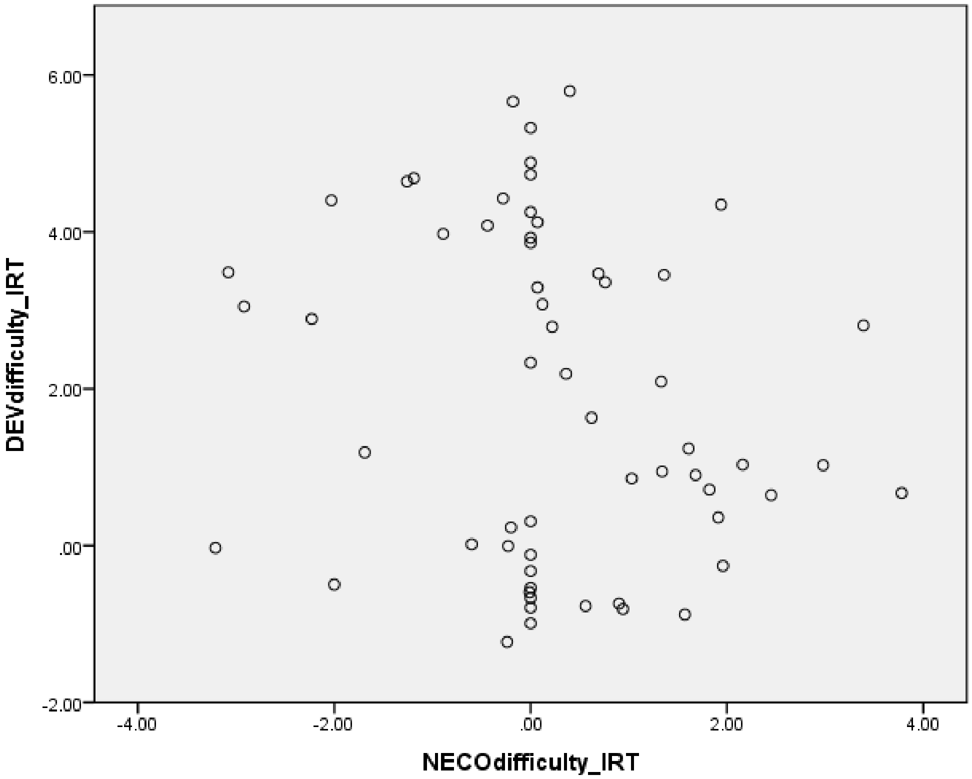

Figure 3. Scatter plot of item difficulty in the DEVmc-MAT and 2015-NECO$_{mc}$-MAT under IRT

Fig, 5 shows that there was no significant linear relationship between the chance factor parameter of the DEV$_{mc}$-MAT and 2015-NECO$_{mc}$-MAT. The Pearson moment correlation coefficient was low and negative and as well as not significant (r = -0.12, p = 0.35). In all, it was evident that correlations between all the parameters under the two frameworks were low and cannot be said to be equivalent.

More so, the mean difference of the test scores and ability scores for the two tests under CTT and IRT frameworks were examined. Tables 5 and 6 presented the descriptive and paired sample t-test statistics of the test scores and ability scores of developed MAT and NECO test.

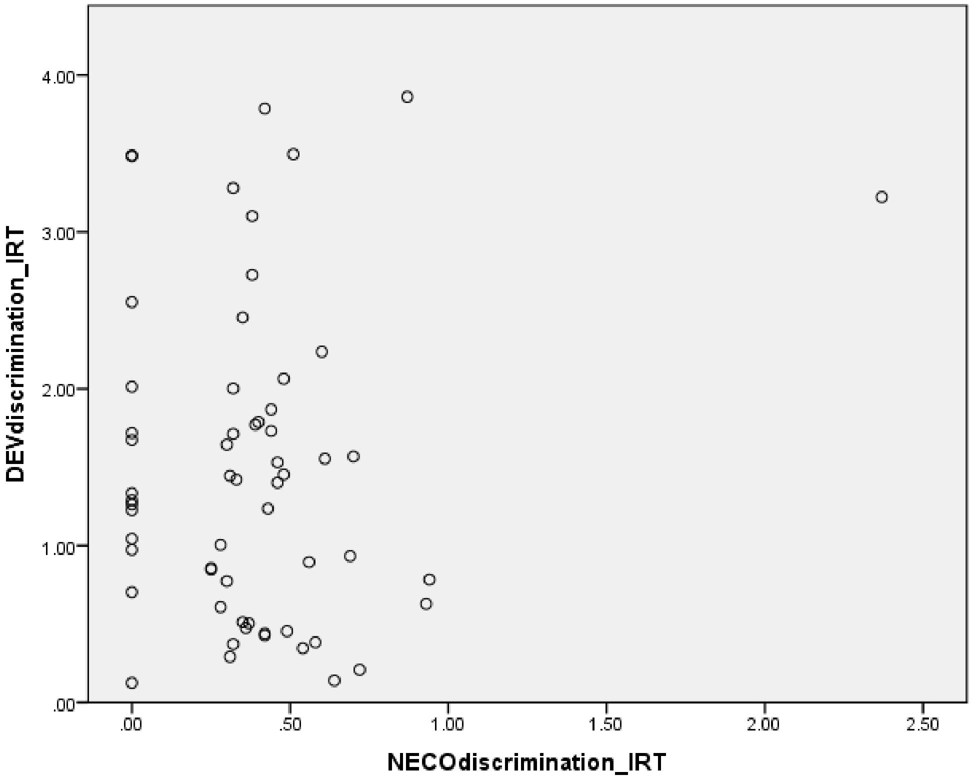

Figure 4. Scatter plot of item discrimination in the DEV$_{mc}$-MAT and 2015-NECO$_{mc}$-MAT under IRT

Table 5 shows the distribution and descriptive statistics of the test scores and ability scores of examinees' in DEV$_{mc}$-MAT and NECO$_{mc}$-MAT tests under CTT and IRT frameworks. The ability scores were obtained from the Multivariate EQSIRT program. It was revealed that under the CTT, NECO$_{mc}$-MAT test was more difficult (M = 24.62; SD = 4.13) than DEV$_{mc}$-MAT (M = 30.42; SD = 14.30). Similarly, under IRT, NECO$_{mc}$-MAT test was more difficult (M = -0.035; SD = 0.876) than the DEV$_{mc}$-MAT (M = 0.001; SD = 0.99).

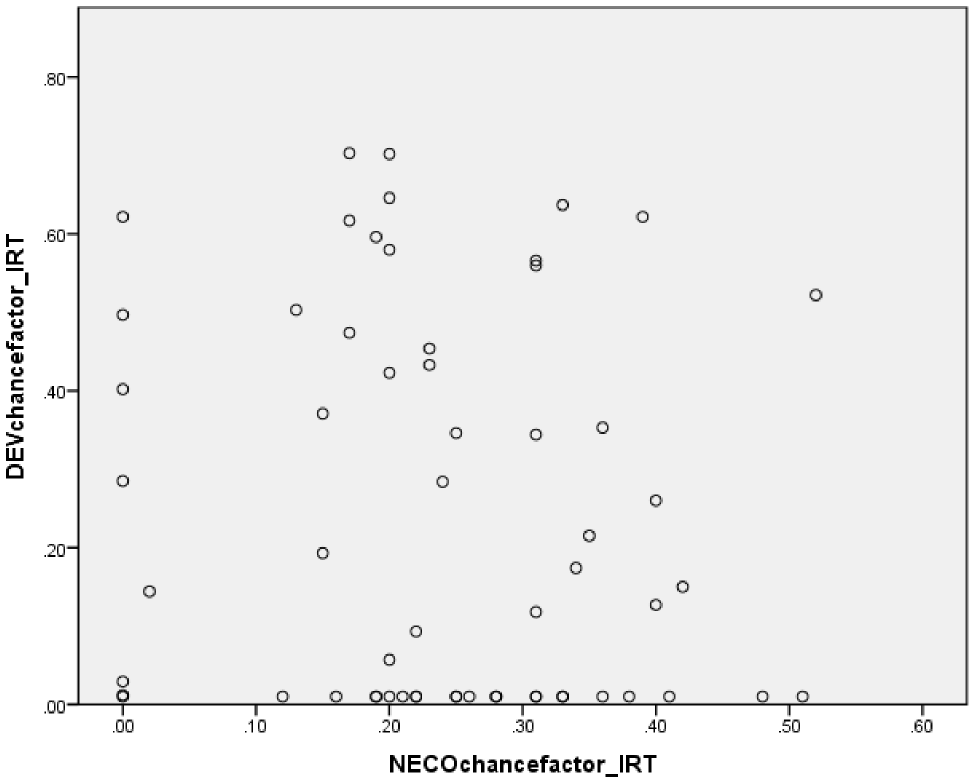

Figure 5. Scatter plot of a chance factor in the DEVmc-MAT and 2015-NECO$_{mc}$-MAT under IRT

Table 5. Descriptive statistics of test scores and ability scores of DEV$_{mc}$-MAT and 2015-NECO$_{mc}$-MAT

Statistics	N	CTT		IRT	
		DEV$_{mc}$-MAT	NECO$_{mc}$-MAT	DEV$_{mc}$-MAT	NECO$_{mc}$-MAT
Min.	1080	16.00	12.00	-0.60	-2.96
Max.	1080	59.00	37.00	2.03	1.78
Mean	1080	30.42	24.62	0.001	-0.04
SD	1080	14.30	4.13	0.99	0.88

Table 6 shows that the mean difference value for test scores of developed multiple-choice MAT and NECO test under the CTT framework was 5.81, while the paired-samples t-test statistics showed that the mean difference was statistically significant (t = 12.75, df = 1079, P = 0.00). Also, the mean difference value for ability scores of the developed multiple-choice MAT and NECO test under the IRT framework was 0.04, while the paired-samples t-test statistics showed that the mean difference was not statistically significant (t = 0.89, df = 1079, P = 0.37). It implies that examinees' ability from the two test forms was not related.

Table 6. Paired sample t-test statistics of test scores and ability scores of DEV$_{mc}$-MAT and 2015-NECO$_{mc}$-MAT

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
DEV$_{testscore}$_ctt-NECO$_{testscore}$_ctt	5.81	14.96	0.46	4.912	6.699	12.750	1079	0.000
DEV$_{abilityscore}$_irt - NECO$_{abilityscore}$_irt	0.04	1.33	0.04	-0.043	0.116	0.893	1079	0.372

Research Question 3: How comparable are the examinees' ranking of scores in the DEV$_{mc}$-MAT using the two contrasting frameworks?

To answer this research question, examinees' test scores and abilities from the two contrasting frameworks in the DEV$_{mc}$-MAT were converted to the same metric scale and ranked in descending order. Table 7 presented examinees' abilities and their ranking in the DEV$_{mc}$-MAT using the two measurement frameworks.

Table 7 depicted that scores obtained under the two contrasting frameworks for the examinees were converted into the same metric scale and ranking of scores were established. It was observed from the ranking that there was a considerable shift of examinees' ranking when it was done within the confine of IRT. For instance, in CTT the highest examinees got 70 scores while in IRT ranking, this relative standing changed and 69.210 was on top of the ranking and so on. This was evident that IRT ranking is superior because examinee with a score of 69.210 might select wrong options for easy items and therefore got less penalty while examinees with a score of 70 in CTT also could not answer the difficult items and got more penalty subsequently lost his or her relative standing. Furthermore, percentile and percentile rank was used to establish the comparability of examinees score ranking from the two measurement frameworks. Table 8 and Fig. 6 presented the percentile scores and graph of relative cumulative frequency curve of examinees in the DEV$_{mc}$-MAT using the two contrasting frameworks.

Table 7. Examinees ranking of scores in the DEV$_{mc}$-MAT under CTT and IRT

S/N	T_{score_CTT}	T_{score_IRT}	$CTT_{score\_descending}$	CTT Rank	$IRT_{score\_descending}$	IRT Rank
1	43	43.450	70	1	69.210	1
2	43	52.350	70	1	69.190	2
3	47	56.730	70	1	69.180	3
4	70	69.130	70	1	69.170	4
5	46	58.640	70	1	69.160	5
+	+	+	+	+	+	+
+	+	+	+	+	+	+
+	+	+	+	+	+	+
1071	47	51.600	41	1013	35.340	1071
1072	45	51.770	41	1013	35.240	1072
1073	43	51.700	41	1013	35.220	1073
1074	48	52.490	41	1013	35.130	1074
1075	51	51.730	40	1075	35.120	1075
1076	46	51.600	40	1075	32.140	1076
1077	48	51.650	40	1075	31.680	1077
1078	43	51.790	40	1075	31.370	1078
1079	52	52.670	40	1075	31.070	1079
1080	52	51.980	40	1075	30.640	1080

Table 8. Percentiles scores of examinees score in the DEV$_{mc}$-MAT using CTT and IRT

Statistics	**Percentiles**						
	5	***10***	***25***	***50***	***75***	***90***	***95***
CTT_score	41.00	43.00	44.00	46.00	49.00	70.00	70.00
IRT_score	35.80	38.74	51.58	51.83	52.31	68.98	69.06

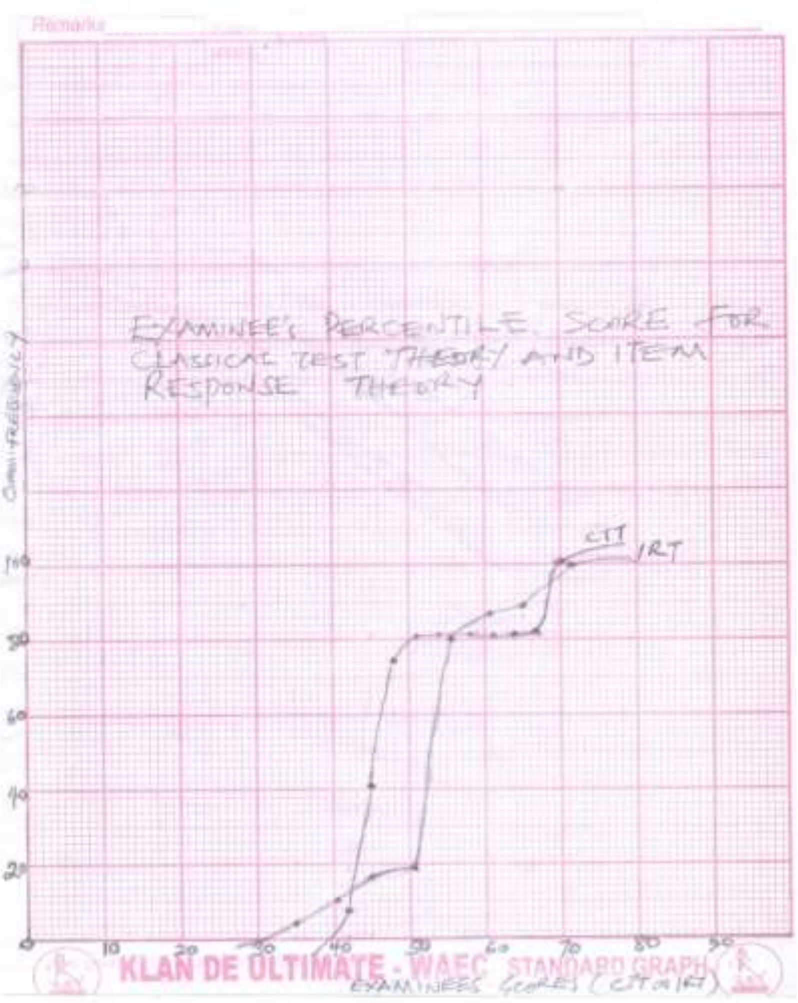

Figure 6. Examinees percentile score for CTT and IRT

Table 8 and Fig. 6 provide evidence on how examinees score relates to a larger group. The statistics depicted that percentiles scores from the two measurement frameworks differ. For instance, under CTT, a raw score of 46.00 points corresponds to a percentile of 50, then 50% of the scores in the distribution were equal to or less than 46 points. While under IRT, a distinctive raw score of 51.83 points falls under the same percentile with 50% of the scores in the distribution were equal to or less than 51.83 points and a similar trend was observed for other percentiles. This implies that there was a thin line drawn between the observed values within the distribution of the classical test theory and item response theory.

Discussions

Dimensionality assumption of IRT was conducted on the responses of the examinees to the items. The result showed that Stout's test of essential unidimensionality hypothesis was not rejected. This implies that there is no significant difference between the assessment subtest (AT) and partitioning subtest (PT). Thus, it was concluded that the test satisfied dimensionality assumption of IRT. The result of this study laid credence to the work of Nandakumar (1993); Finch & Habing (2007) who affirmed that test data are essentially unidimensional when items of AT are of the same dominant dimension as the rest of the items; therefore, DIMTEST does not reject the null hypothesis. When the test data is not unidimensional, the items of AT are dimensionally different from the rest of the items, and DIMTEST rejects the null hypothesis of essential unidimensionality. However, the submission from this study contradicted the work Metibemu (2016) that dimensionality of the 2014 National Examinations Council Mathematics objective test underlies with more than one trait in explaining the performance of examinees in the test data, which implies that the NECO Mathematics test violates unidimensionality assumption. Also, it was shown that the test satisfies the conditional independence assumption of IRT. This submission is in agreement with Chen-Thissen LD.

The results also showed that there was a statistically significant difference between the developed multiple-choice MAT and test of 2015-NECO$_{mc}$-MAT item parameters (difficulty and discriminating indices). Findings lay credence to the conclusion that there was a significant difference in the item parameters of economics multiple-choice items conducted by public examinations bodies. However, the study contradicted the submission of Bandele & Adewale (2013), Metibemu (2016) and Ogbebor (2017) that items parameters produced by developed multiple-choices Physics Achievement Test and 2014 WAEC were equivalent.

Furthermore, Findings also showed that there was a statistically significant difference between the developed multiple-choice MAT and test of 2015-NECO$_{mc}$-MAT person parameters. The results agreed with the submission of Fakayode (2018) that there was a significant difference in the examinees' ability from the two forms of WAEC mathematics achievement test for June and Nov 2015 diets. Similarly, the results corroborated earlier findings that the two test forms displayed very significant differences which make them not usable concurrently. Therefore, the developed mathematics achievement test instrument produced better estimates. The study disagreed with the submission of Metibemu (2016) that the ability scores of candidates in D-PAT is linearly related to the ability scores of candidates in the WAEC test.

Also, IRT scoring ranking produced different relative standing for examinees' who have the same scores under the CTT framework. This submission is in accordance with the findings of Zaman et al. (2008) that under CTT an examinee attempting a difficult question and an easy question get the same relative standing which is not the case in Item Response Theory (IRT).

Conclusion

The researchers concluded that the IRT method was more effective than CCT in test development and scoring. Also, examining bodies should calibrate their multiple-choice response Mathematics Achievement Test using Item Response theory**.**

Acknowledgement. We express our profound gratitude to the director of schools under the Ministry of Education, Osun State, Nigeria for approved all the schools and students used in this study. Also, the National Examinations Council (NECO) which was used as a baseline in this study is well appreciated. This paper was presented at the 2nd Annual Forum of Association of Behavioural Research Analysts and Psychometricians (AB-ReAP) held between 10th – 14th February 2020 at Onyema Ugochukwu Auditorium, Town Campus, University of Uyo, Akwa Ibom State, Nigeria.

REFERENCES

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

Bandele, S.O. & Adewale, G.A. (2013). Comparative analysis of the reliability and validity coefficients of WAEC, NECO and NABTEB constructed mathematics examination. *J. Educ. & Soc. Res., 3(*2), 397-402.

Bichi, A.C., Embong, R., Mamat, M. & Maiwada, D.A. (2015). Comparison of classical test theory and item response theory: a review of empirical studies. *Australian J. Basic & Appl. Sci., 9,* 549-556.

Chen, W.-H. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *J. Educ. & Behavioral Statistics, 22,* 265-289.

Courville, T.G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics: PhD thesis.* College Station: Texas A & M University.

Fakayode, O. (2018). *Comparing CTT and IRT measurement frameworks in the estimation of item parameters, scoring and test equating of West African examinations council mathematics objective test for June and November, 2015: PhD thesis.* Ibadan: University of Ibadan.

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

Finch, H. & Habing, B. (2007). Performance of DIMTEST-and NOHARM-based statistics for testing unidimensionality. *Appl. Psych. Measurement, 31,* 292-307.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: principles and applications.* Newbury Park: Sage.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). Fundamentals of item response theory. *J. Educ. Measurement, 30,* 84-87.

Lawson, S. (1991). One parameter latent trait measurement: do the results justify the effort (pp. 159-168). In: Thompson, B. (Ed.). *Advances in educational research: substantive findings, methodological developments, vol. 1.* Greenwich: JAI Press.

Macdonald, R.P. & Paunonen, S.V. (2001). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educ. & Psych. Measurement, 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: PhS thesis*. Ibadan: University of Ibadan.

Nandakumar, R. (1993). Assessing essential dimensionality of real data. *Appl. Psych. Measurement, 17.* 29-38.

Ogbebor, U.C. (2017). *Construct of mock economics test for senior secondary school students in Delta State, Nigeria using classical test and item response theories: PhD thesis.* Ibadan: University of Ibadan.

Ojerinde, D. & Ifewulu, B.C. (2012). Item unidimensionality using 2010 unified tertiary matriculation examination mathematics pre-test. *International Conference of IAEA (Kazakhstan).*

Ojerinde, D. (2013). *Classical test theory (CTT) vs. item response theory (IRT): an evaluation of the comparability of item analysis results*. Ibadan: University of Ibadan.

Progar, S. & Sočan, G. (2008). An empirical comparison of item response theory and classical test theory. *Horizons Psych.*, 17(3), 5-24.

Stout, W., Habing, B., Douglas, J., Kim, H.R., Roussos, L. & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Appl. Psych. Measurement, 20*, 331-354.

Zaman, A., Kashmiri, A-Ur-R., Mubarak, M. & Ali, A. (2008). Students ranking, based on their abilities on objective type test: comparison of CTT and IRT. *Proc. EDU-COM 2008 International Conference*, pp. 591-599.

✉Dr. Musa A. Ayanwale
Education Foundations
Kampala International University
E-Mail: adekunle.ayanwale@kiu.ac.ug

✉ Dr. Joshua O. Adeleke
Institute of Education
University Ibadan
E-Mail: joadeleke@yahoo.com