Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.11, No.7, 2020

Using Test Theories Models to Assess Senior Secondary Students Ability in Constructed-Response Mathematics Tests

Babatunde K Oladele* Dr. Benson A. Adegoke
Institute of Education, University of Ibadan, Nigeria

Abstract
Testing is essential in education and other behavioural science fields because many decisions and policies are made according to the results of testing. It is therefore, imperative that besides ensuring that the test items are valid and reliable, the scoring of the items must be reliably and validly conducted. It has been established by chief examiners of examining bodies that students most times obtained low scores in the constructed-response items aspect of mathematics and this could be as a result of the assessment procedures adopted by the examination bodies. Also, many research works have been carried out to confirm the similarity between West African Examinations Council (WAEC) and National Examinations Council (NECO) in Nigeria using multiple choice items. Against this backdrop, this study assessed the ability of senior secondary students in constructed-response mathematics tests of WAEC and NECO with test theories models and as well established the similarity between the two examinations. Non-experimental design of *ex post facto* type was adopted. The target population consists of all senior secondary school students in Ibadan Metropolis of Oyo State in Nigeria. Simple random sampling was used to randomly sample 24 schools and 1151 students. The compulsory section A of Paper II of three years past constructed-response items of WAEC and NECO were used as instruments for data collection. Data collected were analysed using mean, standard deviation, Person Product Correlation Movement; Classical Test Theory, generalized partial credit and graded response models of Item Response Theory. Results of the finding shows that students mean score in the examinations were below 50% using CTT and above 50% using IRT models respectively. It was concluded that the two examinations mathematics constructed-response items are equal and IRT models are more efficient and reliable in determining students' ability compare to CTT.
Keywords: WAEC Mathematics Constructed-Response Test, NECO Mathematics Constructed-Response Test, Classical Test Theories Model, Item Response Theory Models,
DOI: 10.7176/JEP/11-7-05
Publication date: March 31st 2020

1. Introduction

One of the best instruments to measure or assess the level of ability of students in educational system is test. Testing is essential in education and other behavioural science fields because many decisions and policies are made according to the results of testing. It is therefore, imperative that besides ensuring that the test items are valid and reliable, the scoring of the items must be reliably and validly conducted. Examinees performance in test could be affected positively or negatively as a result of assessment procedures adopted by examinations bodies.

In Nigeria, the public examination bodies are faced with mass failure especially in Mathematics and English language which are the two most compulsory subjects in secondary school education level. However, the senior secondary school certificate tests are set, conducted, scored and graded by bodies external to the schools. It is possible that some of the factors that account for poor performance could be related to the procedures of assessment of external examination bodies. Other important issue that needs to be examined is the appropriateness of the scoring models that are being adopted by public examining bodies (Adegoke 2016). Issues such as these are within the scope of psychometrics. Examinees' test data needs to be assessed and evaluated using psychometric measures in order to understand, monitor, control and improve the quality of assessments (Onuka and Ogbebor, 2016). The assessment practices being adopted by public examining bodies may be one of the reasons why students are performing poorly in mathematics. Assessment is a process used to determine the area of strength and weakness of students. Assessment of what students' have learnt in school subjects is done by administering achievement tests to students. An achievement test is used to measure the extent to which an individual has learnt from a given course of instruction (Metibemu and Omole 2016).

Achievement tests are of various types, namely; matching test, fill in the gap, short answer, selected-response (multiple-choice) and constructed-response (essay). The most common achievement tests in measurement are the selected-response (multiple -choice) and constructed-response (essay). In the assessment of students' performance in mathematics, two types of tests are usually used. These include multiple choice and essay tests. For multiple choice test, specifically WAEC uses 50 items with each item placed under four response format A, B, C and D, while NECO uses 60 items with each item placed under five response mode, A, B, C, D and E. For the essay tests also known as constructed-response mode and called Paper II, WAEC uses 13 items while NECO uses 12 items. Bandele and Adewale (2013) concluded that WAEC, NECO and NABTEB mathematics achievement examinations are highly reliable and valid and are as well comparable and equivalent. In the aforementioned study

multiple choice items were used but this present study will further confirm if WAEC and NECO constructed-response mathematics items are similar or comparable.

Test theories provide a general framework linking observable variables, such as test scores and item scores, to unobservable variables, such as true scores and ability scores. Thus, a test theory that introduces concepts such as true scores, test scores, and error scores cannot be judged as useful or useless until it is fully specified in the form of a particular model. On the other hand, particular test models are formulated within the framework of a test theory and do specify in considerable detail the relationships among a set of test theoretic concepts along with a set of assumptions about the concepts and their relationships. The appropriateness of such models can be evaluated with respect to a particular set of test data.

In measurement processes in education, there may be unobservable, latent variables that we are particularly interested in, such as achievement in test, reading ability, Mathematics ability, intelligence, and aptitude. Such variables cannot be measured directly since they are constructs rather than physical quantities. Two test theories have been developed by psychometrics experts and professionals in the field of testing enabling us to measure these latent traits. The two popular test theories developed are Classical Test Theory (CCT) and Item Response Theories (IRT).

Classical Test Theory (CTT) has been the foundation for measurement theory for decades. The conceptual foundations, assumptions and extensions of the basic premises of CTT have allowed for the development of some excellent psychometrically sound scales in the assessment practices of educational bodies in Africa. This is owing to the simplicity of interpretation which can usefully be applied to examinees achievement and aptitude test performance (Hambleton, 1989). In the past 30 years or more, the field of educational measurement all over the world has undergone changes to meet increasing demand for valid interpretation of individual score from educational tests or examinations (Adedoyin, 2010).

Classical Test Theory has been defined by experts to mean a simple linear model which states that the observed score on a test is the sum of true score and measurement error. It is a simple linear model which comprises three components, namely: the observe score, the true score and the error score. It is this central idea of the relationship among true score, observed score and error of measurement that enables classical test theory to describe factors which influence the test scores. Classical test theory, also known as ***true-score theory,*** assumes that each person has a true score, ***T*** that would be obtained if there were no errors in measurement (Cappelleri, Lundry & Hays, 2014). A person's*s* **true score** is defined as the expected score over an infinite number of independent administrations of the scale. Scale users never observe a person's true score, only an observed score, ***X***. It is assumed that observed score (***X***) = true score (***T***) + some error (***E***). Classical test theory and related models have been researched and applied continuously and successfully for well over 60 years, and many testing programs today remain firmly rooted in classical measurement models and methods.

Despite the popularity of classical item statistics as an integral part of standardised test and measurement technology, it is identified with so many limitations (Hambleton & Jones, 1993; Ojerinde, 2013). According to Hambleton and Jones (1993) the major limitations of CTT are: (a) the person statistics (that is, observed score) is (item) sample dependent, and (b) item statistics (i.e., item difficulty and item discrimination) are examinee sample dependent. Therefore, the estimates of CTT are not generalizable across populations.

Item response theory (IRT) is a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test. According to Schumacker (2010), "Item response theory (IRT) is based on latent trait theory, incorporates measurement assumptions about examinee, item and test performance, how performance relates to knowledge as measured by the item on a test. The outcome of measurement under item response theory is a scale to which examinees as well as items are placed. In that sense, it is necessary to have a scale of measurement. Since we do not have the exact image of the latent variable, scaling is a difficult task. To overcome this problem, it is generally assumed that the ability scale has a midpoint zero, a unit of measurement of one, and a range from negative infinity to positive infinity (Baker, 2001). Typically, two assumptions are made in specifying IRT models. One relates to the dimensional structure of the test data, and the other relates to the mathematical form of the item characteristic function or curve (denoted ICC).

Under item response theory, the primary interest is in whether an examinee got each individual item correct or not, rather than in the raw test scores. This is because the basic concepts of item response theory rest upon the individual items of a test rather than upon some aggregate of the item responses such as a test score.

However, the rules of measurement under IRT framework afford greater robustness, flexibility, efficiency and reliability in trait measurement than the classical test theory framework. The underlying principle used in IRT models for testing is that person and item parameters can be fully separated and this is brought to bear on measuring examinee traits and test characteristics with greater precision and flexibility.

Item responses can be discrete or continuous and can be dichotomously or polychotomously scored; item score categories can be ordered or unordered. There can be ability or many abilities underlying test performance; and there are many ways (that is models) in which the relationship between item responses and the underlying ability or abilities can be specified (Ostini & Nering, 2006). Within the general IRT framework, many models

have been formulated. Famous names associated with these various scoring models are dichotomous, binomial, poison, rating scale, facet, multinominal logit, or polytomous. These scoring models handle item responses that are discrete or continuous and dichotomous and polytomous scored.

Polytomous items are categorical items in the same way as dichotomous items. They simply have more than two possible response categories. Categorical data can be described effectively in terms of the number of categories into which the data can be placed. Ordered categories are defined by boundaries or thresholds that separate the categories (Adegoke, 2013). Logically, there is always one less boundary than there are categories. For example, a dichotomous item requires only one category boundary to separate the two possible response categories. In the same manner, a 4-point Likert-type (with response Strongly Agree; Agree; Disagree; Strongly Disagree) item requires three boundaries to separate the four possible response categories. In an essay test item that is scored over five, possible categories include 0, 1, 2, 3, 4, and 5 (Adegoke 2013). In this case there are six categories. However, using polytomous model, there are five category boundaries that is the six categories minus one. There are different types of dichotomous and polytomous models that have been developed for dichotomous and polythomous items. The most commonly used models for dichotomous items are the parameter logistic models. These include the one-parametre logistic model, two-parametre logistic model and three-logistic parameter model.

One of the psychometric issues that make polytomous items more attractive than dichotomous items is that polytomous items measure across a wider range of the trait continuum than dichotomous items (Adegoke 2016). This occurs simply by virtue of the fact that polytomous items contain more response categories than dichotomous items. Samejima (1969) introduced the first polytomous model (Graded Response Model). Although Bock and Samejima (1972) presented a different polytomous model (Nominal Categories Model), Interest in polytomous IRT began in 1980's. The polythomous models that have been developed for polythomous items include the Partial Credit Model, Generalised Partial Credit Model, Graded Response Model, and Nominal Response Model. It is only the Generalised Partial Credit Model and Graded Response Model that can effectively determine the difficulty and discrimination parameters of an item. This paper therefore assessed the ability of students in mathematics using test theories models so as to establish the relativeness among models.

The following figures display the statistical data of students' performance in mathematics for both WAEC and NECO.

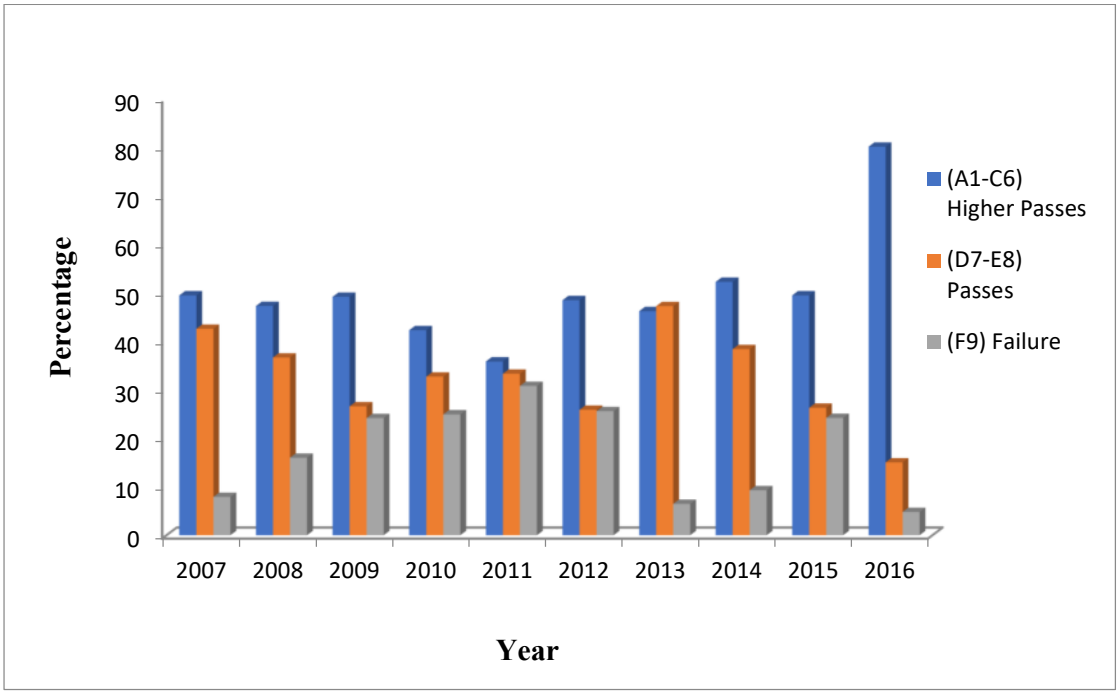

Figure 1.1: Multiple Bar chart showing Analysis of performance in WAEC Mathematics (May/June 2007-2016)

Figure 1.1 reveals that students' performance in WAEC Mathematics has been unimpressive for decades. This is evidenced, as the highest percentage of candidates who attained credit pass was 52.27% in year 2008. However, a cursory look at Table 1.1 and Figure 1.1 shows that there was improvement in candidate's performance from 2007 to 2008, and regression in students' performance from 2009 to 2011 and from 2012 to 2014 respectively. However, there was improvement in 2013 to 2016 but the ultimate if to have 100% performance.

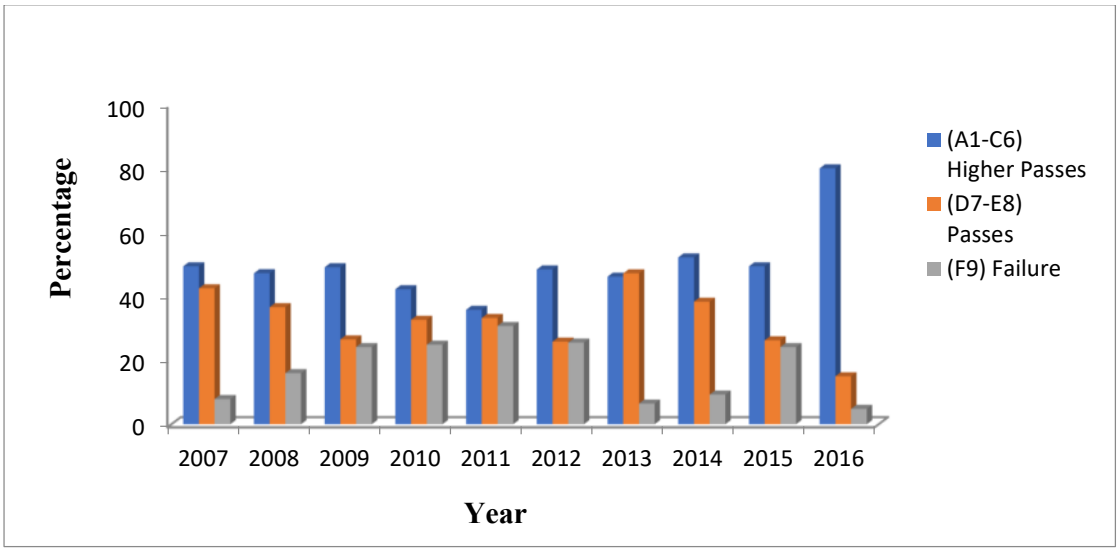

Figure 1.2: Multiple Bar chart showing Analysis of candidate's performance in NECO mathematics (May/June 2007-2016)

Figure 1.2 reveals that students' performance in NECO Mathematics has been unimpressive for decades. This is evidenced, as the highest percentage of candidates who attained credit pass was 52.27% in year 2008. However, a cursory look at Table 1.1 and Figure 1.1 shows that there was improvement in candidate's performance from 2007 to 2008, and regression in students' performance from 2009 to2011 and from 2012 to 2014 respectively.

Examiners' reports (WAEC Chief Examiners' Report, Mathematics, 2013, 2014, and 2015) revealed that candidates usually failed in theory test items that require solving numerical problems. Some candidates showed lack of skills in the application of formulae to solve problems. Researchers and stakeholders in education industry have in the recent past identified several factors as the causes of poor performance of students in Mathematics. Among such factors identified are poor location of the school, incessant changes in government policies, closure of schools, which is contingent upon teachers' strike action, home-school distance, high student teacher ratio, lack of supervision, monitoring and evaluation machinery, lack of good textbooks, poor content and context of instruction, poor and non -conductive environment among others (Adepoju, 2002). Yet, others blamed the Government for failure to provide human and material resources to facilitate good teaching and learning, some blamed the teachers for failure to inculcate the necessary knowledge, skills and behavior to students and also the students themselves for refusal to learn

Another issue that is associated with students' performance in mathematics is gender. There is a general held view that boys are better than girls in mathematics. From the available literature, gender issues have been linked with performance of students in mathematics in several studies but without any definite conclusion. Adeleke (2007) affirmed that male students performed better than the females in Mathematics. This was in consonance with the WAEC chief examiners' report for more than one and a half decade (1996 – 2011) which confirmed that boys performed better than girls in mathematics. Thus, the question of gender differences in mathematics achievement remains an issue that is not completely resolved at present. It is therefore considered necessary to really confirm the difference between male and female students' performance in mathematics constructed-response tests,

1.1 Research Questions

1. What are the ability estimate of students in NECO and WAEC mathematics constructed-response tests along:
   (a) Classical Test Theory (CTT) framework?
   (b) Item Response Theory (IRT) framework vis a vis:
      i. Graded Response Model (GRM)?
      ii. Generalised Partial Credit Model (GPCM)?
2. How related is the ability estimate of students in NECO and WAEC mathematics constructed-response tests along?
      i. Classical Test Theory (CTT) Model?
      ii. Graded Response Model (GRM)?
      iii. Generalised Partial Credit Model (GPCM)?
3. Is there any difference in male and female students ability in NECO and WAEC mathematics

constructed-response tests along:

i. Classical Test Theory (CTT) Model?
ii. Graded Response Model (GRM)?
iii. Generalised Partial Credit Model (GPCM)?

2. Method

The study adopted non-experimental design of *ex post facto* type. This was employed since the aim of the study was to assess the ability of students using test theories models without any form data manipulation. To test the research questions raised, two test frameworks were used. These are Classical Test Theory (CTT) and Item Response Theory (IRT). Simple random sampling technique was used to select 24 senior secondary schools in Ibadan metropolis of Oyo State, Nigeria. 1151 students were randomly sampled.

2.1 Material

The two prominent test theories were used in estimating and comparing the ability of students in mathematics in order to determine the relativeness of the models. The similarities between WAEC and NECO tests were also considered using the test theories models. SPSS 23 and IRT Pro 3 software were used to analyse the collected data.

2.2 Data analysis

The descriptive analyses for the data collected are as follows:

Table1: Distribution of Students by Gender

Gender	Number	Percentage
Male	565	49.1
Female	586	50.9
Total	**1151**	**100.0**

Table 1 presents the gender distribution of students showing that 49.1 % are male while 50.9 % are female. This indicates that on the average there are more female students than male students in the senior secondary school in Ibadan, Oyo State, Nigeria.

Table 2: Distribution of Students by Age

Students Age Range	Number	Percentage
12-16	537	46.7
17-19	603	52.4
20-22	11	1.0
Total	1151	100.0

Table 2 presents the age distribution of students showing that 46.7% are between 12 and 16 years old, 52.4% are between 17 and 19 years old while 1.0% are between 20 and 22 years old respectively. This indicates that on the average the age of students in senior secondary schools are in consonance with the age range stipulated in the 6 3 3 4 education system documents of the state.

Table 3: Distribution of Students Gender arranged by Age

Students Gender	Students Age			Total
	12-16	17-19	20-22	
Male	215	339	11	565
Female	322	264	0	586
Total	**537**	**603**	**11**	**1151**

Table 3 present the distribution of students' gender arranged by age showing that 339 students who are between age 17 and 19 are male while 264 are female. 215 male students are between age 12 and 16 respectively. Also, 11 male students are between age 20 and 22 while there is no female student that are between age 20 and 22. This indicates that male students are older than female students in senior secondary schools in Ibadan, Oyo state in Nigeria.

3. Findings

Research Question One: What are the ability estimates of students in NECO and WAEC mathematics constructed-response tests along:

(a) Classical Test Theory (CTT) framework?
(b) Item Response Theory (IRT) framework vis a vis:
i. Graded Response Model (GRM)?
ii. Generalised Partial Credit Model (GPCM)?

To answer the research questions raised, the following data were analysed:

Table 4: Summary of Descriptive Statistics of Estimation of Student Ability arranged by Models

Models	Number	Minimum Score	Maximum Score	Mean Score	Std. Deviation
NECO_CTT	1151	8	78	33.49	12.394
WAEC_CTT	1151	10	75	35.88	10.018
NECO_GPCM	1151	23	84	50.15	8.977
WAEC_GPCM	1151	29	77	50.07	8.858
NECO_GRM	1151	22	86	50.01	8.994
WAEC_GRM	1151	28	77	50.02	8.841

From table 4 presents the least minimum score of students across the models as 8 while the highest maximum score is 86 which is one of the estimated scores from the Item Response Theory models. Specifically, it was the Graded Response Model (GRM). Also, the least mean score is 33.49 while the highest mean score is 50.15 which is one of the estimated scores from IRT models. Specifically, it was the Generalised Partial Credit Model (GPCM). This indicates that the IRT models produced better estimates of students' ability in mathematics constructed-response tests.

Research Question Two: How related is the ability estimate of students in NECO and WAEC mathematics constructed-response tests along?

i. Classical Test Theory (CTT) Model?
ii. Graded Response Model (GRM)?
iii. Generalised Partial Credit Model (GPCM)?

Table 10: Correlation Analysis of the relationship between Students Ability Estimate in WAEC and NECO Using Classical Test Theory Model

Achievement Tests	Number	Mean	Standard Dev	*R*	P-Value	Remark
NECO	1151	33.49	12.394			
				0.96	0.01	Significant
WAEC	1151	35.88	10.018			

Significant at $p<0.05$

Table 10 presents the result of Pearson Product Movement correlational coefficient showing the relationship between students' ability in WAEC and NECO mathematics tests. The result shows that there is high and positive relationship between student ability in the two examinations achievement tests (0.96) which is significant at $p<0.05$ using Classical Test Theory model. This indicates that an increase in students' ability in WAEC test will cause a corresponding increase NECO test.

Table 11: Correlation Analysis of the relationship between Students Ability Estimate in WAEC and NECO Using Generalised Partial Credit Model

Achievement Tests	Number	Mean	Standard Dev	*R*	P-Value	Remark
NECO	1151	50.15	8.98			
				0.72	0.02	Significant
WAEC	1151	50.07	8.86			

Significant at $p<0.05$

Table 11 presents the result of Pearson Product Movement correlational coefficient showing the relationship between students' ability in WAEC and NECO mathematics tests. The result shows that there is high and positive relationship between student ability in the two examination body achievement tests (0.72) which is significant at $p<0.05$ using Generalised Partial Credit Model. This indicates that an increase in students' ability in WAEC test will cause a corresponding increase NECO test.

Table 12: Correlation Analysis of the relationship between Students Ability Estimate in WAEC and NECO Using Graded Response Model

Achievement Tests	Number	Mean	Standard Dev	*R*	P-Value	Remark
NECO	1151	50.01	8.99			
				0.87	0.00	Significant
WAEC	1151	50.02	8.84			

Significant at $p<0.05$

Table 12 presents the result of Pearson Product Movement correlational coefficient showing the relationship between students' ability in WAEC and NECO mathematics tests. The result shows that there is high and positive relationship between student ability in the two examination body achievement tests (0.87) which is significant at

p<0.05 using Generalised Partial Credit Model. This indicates that an increase in students' ability in WAEC test will cause a corresponding increase NECO test.

Research Question Three: Is there any difference in male and female students' ability in NECO and WAEC mathematics constructed-response tests along:

i. Classical Test Theory (CTT) Model?
ii. Graded Response Model (GRM)?
iii. Generalised Partial Credit Model (GPCM)?

Table 13: Descriptive Statistics of Students Ability Estimates arranged by Gender

Test	Students Gender	Number	Mean	Std. Deviation	Std. Error Mean
NECO	Male	565	34.17	12.36	0.52
	Female	586	32.83	12.40	0.51
WAEC	Male	565	36.51	9.75	0.41
	Female	586	35.27	10.24	0.42

Table 13 presents students' ability estimates showing the mean score of male students in NECO test as 34.17 while female students mean score is 32.83. Also, mean score of male students in WAEC is 36.51 while female students mean score is 35.27 respectively. This indicates that male student relatively performs better in both tests compared to their female counterpart.

Table 14: Analysis of Independent Sample T-test showing the Difference between Male and Female Students Ability in Mathematics Constructed-Response (CTT Model Estimates)

Test		Mean Diff.	Std. Dev	t	df	p-value	95% Confidence Interval	
							Lower bound	Upper bound
NECO	Equal variances assumed	1.34	0.73	1.84	1149	0.63	-0.09	2.77
	Equal variances not assumed	1.34	0.73	1.84	1148		-0.09	2.77
WAEC	Equal variances assumed	1.24	0.59	2.10	1149	0.56	0.08	2.40
	Equal variances not assumed	1.24	0.59	2.10	1149		0.08	2.39

Not Significant at p>0.05

Table 14 presents the difference between male and female students' ability in WAEC and NECO mathematics constructed-response tests using CTT estimated scores. It is revealed that there is no significant average difference between the ability estimate of male and female students in NECO test which is not significant at (t_{1149} =1.84, p>0.05) and WAEC test not significant at (t_{1149}=2.10, p>0.05) respectively.

Table 15: Descriptive Statistics of Students Ability Estimates arranged by Gender (GPCM Estimates)

Test	Student Gender	Number	Mean	Std. Deviation	Std. Error Mean
NECO	Male	565	50.42	8.88	0.37
	Female	586	49.88	9.07	0.37
WAEC	Male	565	50.24	9.36	0.39
	Female	586	49.90	8.35	0.35

Table 15 presents students' ability estimates showing the mean score of male students in NECO test as 50.42 while female students mean score is 49.88. Also, mean score of male students in WAEC is 50.24 while female students mean score is 49.90 respectively. This indicates that on the average male students relatively perform better in both tests that their female counterparts.

Table 16: Analysis of Independent Sample T-test showing the Difference between Male and Female Ability in Mathematics Constructed-Response (GPCM Estimates)

							95% Confidence Interval	
							Lower bound	Upper bound
Test		**Mean Diff.**	**Std. Dev**	**t**	**Df**	**p-value**		
NECO	Equal variances assumed	0.53	0.53	1.01	1149	0.63	-0.50	1.57
	Equal variances not assumed	0.53	0.53	1.01	1149		-0.50	1.57
WAEC	Equal variances assumed	0.34	0.52	0.65	1149	0.01*	-0.69	1.36
	Equal variances not assumed	0.34	0.52	0.65	1149		-0.69	1.37

*Significant at p<0.05

Table 16 presents the difference between male and female students' ability in WAEC and NECO mathematics constructed-response test using GPCM estimates. It is revealed that there is no significant average difference between the ability estimate of male and female students in NECO test (t_{1149} =0.53, p>0.05) while also no significant average difference exists between male and female student's ability estimate in WAEC test (t_{1149}=0.34, p<0.05) respectively.

Table 17: Descriptive Statistics of Students Ability Estimates arranged by Gender using Graded Response Model

Test	**Students Gender**	**Number**	**Mean**	**Std. Deviation**	**Std. Error Mean**
NECO	Male	565	50.22	8.93	0.38
	Female	586	49.81	9.06	0.37
WAEC	Male	565	50.27	9.35	0.39
	Female	586	49.78	8.32	0.34

Table 17 presents students' ability estimates showing the mean score of male students in NECO test as 50.22 while female students mean score is 49.81. Also, mean score of male students in WAEC is 50.27 while female students mean score is 49.78 respectively. This indicates that male students relatively perform better in both tests that their female counterparts.

Table 17: Analysis of Independent Sample T-Test on the Difference between Male and Female Ability in Mathematics Constructed-Response (GRM Estimates)

							95% Confidence Interval	
							Lower bound	Upper bound
Test		**Mean Diff.**	**Std. Dev**	**t**	**Df**	**p-value**		
NECO	Equal variances assumed	0.42	0.53	0.78	1149	0.70	-0.63	1.46
	Equal variances not assumed	0.42	0.53	0.78	1148		-0.63	1.46
WAEC	Equal variances assumed	0.49	0.52	0.94	1149	0.00*	-0.53	1.51
	Equal variances not assumed	0.49	0.52	0.94	1148		-0.54	1.51

*Significant at p<0.05

Table 17 present the difference between male and female students' ability in WAEC and NECO mathematics constructed-response test using CTT. It is revealed that there is no significant average difference between the ability estimate of male and female students in NECO test (t_{1149} =0.14, p>0.05) while a significant average difference exists between male and female student's ability estimate in WAEC test (t_{1149}=8.27, p<0.05)

4. Results, Discussions and Suggestions

The results of the findings show that students' ability in mathematics constructed-response item of both WAEC and NECO using the two Item Response Theory frameworks produced better estimates compared to the Classical Test Theory framework. The findings corroborate the findings of (Adegoke 2016) who found out that GPCM gave better estimates than RPCM in scoring physics essay test. Hence this suggests that the examination bodies should

switch from the present assessment procedures to a more robust and flexible one like the IRT so that examinees can benefit from the inherent good services of the models. Similarly, the study confirms that student ability in WAEC test correlates with their ability in NECO test. This means that the two tests are similar and parallel. It corroborates the findings of Bandele and Adewale (2013) who concluded that WAEC, NECO and NABTEB tests are similar and equivalent.

However, the study also established the differences between male and female students' ability in mathematics constructed-response item of both examinations. It was discovered that male examinees performed better than the female examinees. This supported the findings of Ogbebor & Onuka (2013), and Adeleke (2007) who affirmed that male students performed better than the female students in Mathematics. This may be connected to the facts that the study revealed that male students are older and had experience than their female counterparts. It is therefore suggested that the female students should be encouraged to step up their performance in mathematics in order to measure up with the male students' ability.

This study recommended that examination bodies should adopt the good test framework that will enhance the performance of students not only in mathematics but in all subjects. It was also recommended that students can write either WAEC or NECO tests because they are relatively equivalent.

Reference

Adedoyin, C. (2010). Investigating the Invariance of Persons Parameter Estimates based on Classical Test and Item Response Theories. An *international journal on education science* 2(2);107-113

Adegoke, B.A, (2013). Effects of Item-Pattern Scoring Method on Senior Secondary School Students Ability Scores in Physics Achievement Test. *West African Journal of Education* 24: 181-190.

Adegoke, B.A, (2016). Examination of Examinees's Ability in Physics under two IRT polytomous Models. Public examining in Sub-Sahara Africa:Issues, Challenges and Prospects. Vol 1. 229-240. *A book in honour of Professor Dibu Ojerinde published by JAMB, Nigeria.*

Adeleke, J.O. (2007). Identification and Effect of Cognitive Entry Characteristics of Students Learning Outcomes in Bearing Mathematics. An Unpublished Ph.D Thesis, University of Ibadan

Adepoju, T.L (2002). Location Factors as Correlates of Private cost and Academic Performance of Secondary School Students in Oyo State, Nigeria. Unpublished Ph.D. Thesis, 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.

Bandele, S.O. & Adewale A. E. (2013). Comparative Analysis of the Reliability and Validity Coeffient of WAEC, NECO and NATEB Contructed Examination. *Journal of Education and Social Research* 3(2): 397-402

Bock, R. D. (1972). Estimating Item Parameters and Latent Ability when Responses are Scored in two or more Nominal Categories. Psychometrika. 37, 29-51.

Cappelleri, J. C. Lundy, J. J. & Hays, R. D. (2014). Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures. Elsevier HS Journals, Inc. Published by Elsevier Inc.

Ojerinde, A. (2013). Classical Test Theory (CTT) VS Item Response Theory (IRT): An evaluation of the comparability of item analysis results. A guest lecture presented at the Institute of Education, University of Ibadan on 23rd May.

Hambleton, R. K. & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston, MA: Kluwer-Nijhoff.

Hambleton, R.K., Swaminathan, H. & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Press.

Metibemu M.A. & O.T. Omole (2016). Achievement Test in 21st Century Public examining in Sub-Sahara Africa:Issues, Challenges and Prospects. Vol 1. 215-227. A *book in honour of Professor Dibu Ojerinde published by JAMB, Nigeria.*

Ogbebor, U.C & Onuka A.O.U. (2013). Differencial Item Functioning as an Item Bias indicator.Journal of International Education Research. pp. 367-373.http://www.interesiials.org/ER

Onuka A.O.U. & Ogbebor, U.C (2016). An Introduction to Assessment. Public examining in Sub-Sahara Africa: Issues, Challenges and Prospects. Vol 1. 145-245. *A book in honour of Professor Dibu Ojerinde published by JAMB, Nigeria.*

Ostini, R. and Nering, M. L. (2006). Polytomous Item Response Theory Models. Thousand Oaks: Sage Publication.

Schumacker, R.E. Si C.B R. & Mount R. (2003). Ability Estimation under Different Item Parameterisation and Scoring Models. *American Educational Research Association* Chicago, Illinois USA.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. Psychometric Monograph, No.17.

Samejima, F. (1979). A New Family of Models for the Multiple-Choice Item (Research Report No. 79-4).

Knoxville, TN: University of Tennessee, Department of Psychology.
West Africa Examinations Council, Nigeria .(2013). Chief Examiners Report. Available: http://waeconline.org.ng/e-learning/
West Africa Examination Council, Nigeria (2014). Chief Examiners Report. Available http://waeconline.org.ng/e-learning/
West Africa Examination Council, Nigeria (2015). Chief Examiners Report. Available http://waeconline.org.ng/e-learning/