COMPARATIVE ANALYSIS OF MATHEMATICS ACHIEVEMENT TEST BETWEEN PUBLIC EXAMINING BODIES USING ITEM RESPONSE THEORY

Tobih, Deborah O.
&
Ajayi, Olufemi Abiodun
&
Aghanti, Zachaeus U.
Department of Counselling Psychology and Educational Foundations
College of specialized and Professional Education
Tai Solarin University of Education
Ijagun, Ogun State, Nigeria.

Abstract
This study compared Mathematics achievement test between two public examining bodies in Nigeria, using Item Response Theory (IRT). It adopted non-experimental design of survey research type. Three hundred randomly chosen SS3 students participated in the study. Mathematics multiple choice test (MMCT) composed of NECO and WAEC past questions was used for data collection. The data were analyzed using the software BILOG-MG, and the Stout test of essential uni-dimensionality. The findings revealed that NECO and WAEC 2018 Mathematics objective test items were not uni-dimensional; WAEC 2018 multiple choice items had four distracters and a key, whereas NECO had three distracters and a key; the item parameters for NECO and WAEC 2018 multiple choice items were not comparable; and 43 items out of 50 NECO 2018 items used, tested male and female students differently, whereas 37 items out of 60 WAEC 2018 items tested male and female students differently. The results also revealed that 43 items out of 50 administered by NECO function differently between male and female, accounting for 86 percent of the total number of items administered, whereas 37 items out of 60 administered by WAEC function differently between male and female, accounting for 61.7 percent of the total number of items administered. The items of both examinations tested male and female students differently and that the item parameter for NECO and WAEC 2018 multiple choice items were not comparable. Based on the findings, it was suggested that the public examining bodies should be more meticulous with the procedure in test construction making sure that the process is never compromised.

Keywords: *Comparative analysis, public examining bodies (NECO and WAEC), Mathematics achievement test, Item Response Theory, Gender, Senior Secondary Schools.*

Introduction

Mathematics is a compulsory subject at every level of education in Nigeria except in the university where it is a field on its own. It relates to other school subjects in areas like number and numeration, variations, graphs, fruitions, solutions of equation, and area and volumes. The place of Mathematics in secondary school curriculum in Nigeria is paramount for scientific and human development as it serves both as a tool for academic progress in a chosen career and as a tool for preparing the individual for useful living (Science Teachers Association of Nigeria, 1992). As part of achieving the objectives of Mathematics education, there is the need to conduct external examinations at the terminal class of the Senior Secondary Schools, through the use of different assessment formats including essay and objective tests by the National Examinations Council (NECO). Multiple choice tests have 'problems' (questions) which are called the stem and a list of alternative responses (the correct answer is the key while the incorrect ones are called distracters). The scores obtained from the multiple-choice questions are used to assess the competence of the students in Mathematics (Okoro, 2006). Item Response Theory (IRT) is the most significant development in psychometrics. It explains what happens when an individual encounters a multiple-choice test item. The model simply says that the outcome of such an encounter is governed by the product of the ability of the person and the easiness of the item and nothing more. On the other hand, ITR has become important in the development, interpretation and evaluation of tests and test items (Nenty, 2004). According to Ojerinde (2013), IRT has only one parameter ascribed to the trait level of the individual, the task or item is often characterized by the three parameters. The individual trait level is often designated by theta (θ), which represents the amount of ability, trait or attribute level possessed by an individual. The three parameters associated with the item are 'a' discrimination power, 'b' the difficulty parameter, and 'c' the guessing parameter. IRT has three models which are known as three, two or one parameters. The simplest of model is the one parameter model and it is recommended that it is better to start from the most complex, which is the three parameter IRT model which is represented as:

$$P_i(\theta) = C_i + (1 - C_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

Where Ci is represented the guessing factor
ai is the item discrimination parameter which is generally known as item slope.
bi is the item difficulty parameter commonly known as the item location parameter.
D is the arbitrary constant (normally D = 1.7) and θ represents the ability level of particular examinees.

The location parameter of an item is on the same scale of ability θ and takes the value of θ at the point at which a test taker with the ability level has 50/50 probability of answering the item correctly. The slope of the target line of the item characteristics curve at the point of the location parameter is known as item discrimination. The guessing factor is represented as zero. The two parameters need to be estimated as stated:

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

If there are stipulations that all the items have equal and fixed discrimination, then "a" will become a constant rather than a variable, thus the parameter does not need estimation and the IRT model is further reduced to:

$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}$$

Thus, the one parameter IRT model, constraints have been composed on two of the three possible item parameter that needs to be estimated, the three parameter model is the most general model, and the other two models (two and one parameter models) can be considered as models subsumed under the three parameter model (Lord, 1980; Hambleton & Swaminathan 1985; Hambleton Swaminathan, & Rogers, 1991). The three IRT models are based on the logistic cumulative distribution function. These logistic equations when plotted on a graph, produce plots that is called item characteristics curve (ICC) and when the ICC is plotted the ability of the examinee is denoted by theta (θ) on the x-axis, while the probability of an examinee's correct answer to the questions is represented by P(θ) on the y-axis. This is represented in figure 1

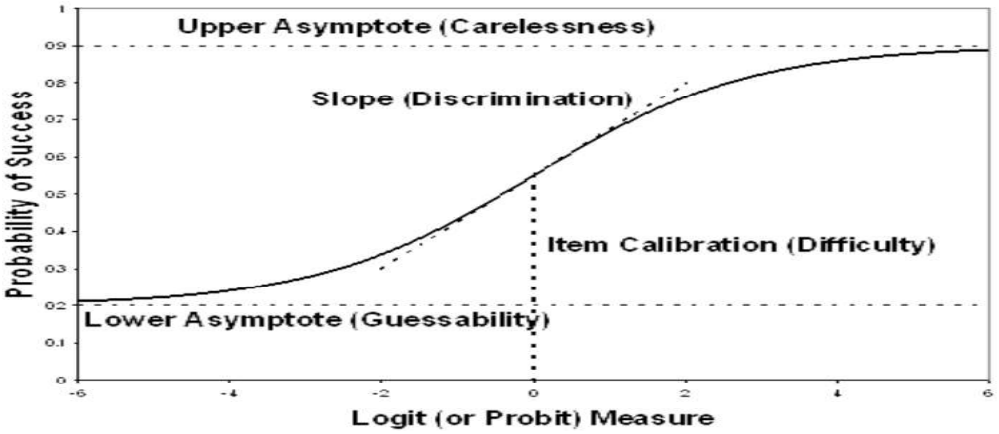

Figure 1: Item Characteristics Curve. Source: (Ojerinde. 2013).

The ICC typically takes the shape of an S-shaped curve called ogive (∫). As shown in figure 1.

IRT is said to have three basic assumptions which are uni-dimensionality, local item independence and item characteristic curves (Ayanwale, 2017). Uni-dimensionality means that the items measure one and only one area of knowledge or ability (Ojerinde, 2013). The assumption of uni-dimensionality states that items of a test measure one's ability. Local item independence is the probability of an examinee getting a test item correct must not be dependent on the response given to other items in the test. The issue of local item independence is that the items on the examinee's performance on different items on the test should be independent. Another assumption is that of Item Response Function (IRF), also referred to as Item Characteristic Curve (ICC). It is a graphical display of student proficiency (ability) level based on the student ability level (Θ). The graph displayed takes the form of a normal *ogive* (normal distribution curve). After the probability of giving the correct answer across different levels of Θ are combined, the relationship between the probabilities and Θ are thus presented as an item characteristic curve.

In Nigeria, public examining bodies including the National Examinations Council (NECO), construct test items on Mathematics (among others) which they administer to students for certification. Students that take this examination are supposed to perform without bias to gender, age and so on. However, candidates who participated in the examinations conducted by these examination bodies are in different settings and therefore differently toned for

personal and environmental reasons. As a result of this, the problem of test item bias cannot be ruled out in these examinations. It has been claimed that some of the national examinations unfairly favour examinees of some particular group (Emaikwu, 2012). Test score validity is of primary importance in a Certification Programme. The National Policy on Education (FRN, 2014) stated that the National Examination tests should be as valid as possible and as fair as possible to all students. Also, there have been conflicting findings by researchers (Donnellan, 2003; Hazari and Potvin, 2005; Laura, 2006) on the influence of gender on students' achievement. Iroegbu (2008) discovered that gender effect is significant and went further to posit that male students performed significantly better than female students. In his view, Erinosho (2005) stated that science is a male enterprise while David and Stanley (2000), Arigbabu and Mji (2004) in their findings stated that there were no longer distinguishing differences in the cognitive, affective and psychomotor skills' achievements of students in respect of gender.

In addition to the above, over the years, performance of students in Mathematics examinations has not been encouraging and this has been a source of concern to the government, parents, teachers, school administrators, other stakeholders and the general public. Students fail due to the low-quality teaching staff, nature of the subject, inadequate preparation of students for Mathematics examinations as well as cut in education budgets leading to shortages of facilities and equipment needed for effective teaching and learning. Public examinations in Nigeria, particularly the Secondary School Certificate Examinations, have been afflicted by examination malpractices and mass failure. It is in view of the above background that this study is set to investigate the comparability of Mathematics questions of NECO June/July and WAEC May/June 2018 using Item Response Theory (IRT) measurement framework.

Observations have revealed some criticisms against NECO Mathematics paper in which some stated that its questions are tougher than those of WAEC. In addition to this is the observation that results from public examination bodies in Mathematics showed that students are not performing well in external Mathematics tests compared with their performance in teacher made tests, and several reasons among which are inadequate teaching methods, inadequate instructional materials, students' fear of Mathematics, teacher-centred/textbook directed rather than learner-centred methods, to mention a few, have been advanced as some of the challenges. However, an important area which still needs to be looked into critically is the area of assessment of

the items making up the objective tests. By Item Response Theory (IRT) Standards, test items should be invariant in nature but unfortunately some items are found to be interacting with the characteristics of the examinees which ought not to be. Invariably, the fairness of the examination items constructed by NECO and WAEC in Mathematics should be examined for comparison. It is in view of this that this study comparatively analyzed public examining bodies mathematics achievement test using Item Response Theory measurement framework.

The general purpose of this study was to carry out Item Response Theory (IRT) analysis of Mathematics questions of NECO June/July and WAEC May/June 2018. The specific objectives of this study were to:

i. Examine if the Mathematics test items of NECO and WAEC 2018 are uni-dimensional;

ii. Determine the item parameters (difficulty, discrimination and guessing) of Mathematics test items of NECO and WAEC 2018 using Item Response Theory framework;

iii. Examine how comparable are the item parameters (difficulty, discrimination and guessing) of Mathematics test items of NECO and WAEC 2018 using Item Response Theory framework;

iv. Find out if NECO Mathematics test items function differentially between male and female;

v. Find out if WAEC Mathematics test items function differentially between male and female?

vi. Examine how comparable are the differential item functioning of Mathematics test items of NECO and WAEC 2018 based on gender using Item Response Theory framework?

The following questions were answered in the study:

1. Are the Mathematics test items of NECO and WAEC 2018 uni-dimensional?

2. What are the item parameters (difficulty, discrimination and guessing) of Mathematics test items of NECO and WAEC 2018 using Item Response Theory framework?

3. How comparable are the item parameters of NECO and WAEC 2018 mathematics test items.
4. Do NECO Mathematics test items function differentially between male and female?
5. Do WAEC Mathematics test items function differentially between male and female?
6. How comparable are the differential item functioning of Mathematics test items of NECO and WAEC 2018 based on gender using Item Response Theory framework?

Method

The study employed non-experimental design of survey research type. The population for the study comprised all SS3 students in public senior secondary schools in Ijebu-Ode Local Government Area, Ogun State. The sample of the study was made up of three hundred (300) SS3 students taken from ten randomly selected schools (30 students from each school). The study instrument "Mathematics Multiple Choice Test (MMCT)" comprised of 60 objective test items adapted from May/June WAEC Mathematics examination and 50 objective items adapted from 2018 June/July NECO examination respectively. Bilog-Mg software and Stout test of essential uni-dimensionality were used for the analysis.

Results

Table 1: Factor Loading of NECO 2018 Multiple Choice Test Items

ITEMS	MR2	MR4	MR3	MR5	MR7	MR8	MR1	MR9	MR10	MR12	MR6	MR13	MR11
1	0.642												
2	0.533												
3													
4					0.698								
5					0.502								
6								0.713					
7								0.471					
8												0.328	
9											0.367		
10								0.458					
11													
12			0.82										
13			0.463									-0.361	
14													
15		0.748											
16		0.552											

17											-0.381		
18	0.36			0.403							-0.311		
19													
20													0.37
21										0.308			0.337
22											0.31		
23	0.361												
24							0.588						
25							0.567						
26													
27						0.321							
28						0.534							
29						0.742							
30						0.45							
31												0.602	
32												0.371	
33													
34			0.36				-0.315				0.3		
35		0.311									0.539		
36													
37				0.355									
38				0.843									
39				0.506									
40													
41										0.361			
42										0.673			
43										0.54			
44	0.359				-0.414		0.331						
45							0.331						
46									0.359				
47									0.792				
48									0.429				
49													
50													0.521

Table 1 shows that NECO 2018 Mathematics objective test items has 13 factors into which the 50 multiple choice items are divided. This shows that the test is not uni-dimensional in nature. The table also shows that some items do not load on any of the 13 factors, which means they do not add anything to the testing that was done. However, some items load under more than one factor which means that they are testing more than one thing at a time. This could be items such as word problem. The items to be retained here are those items that have positive values from 0.3 and above.

Table 2: Factor Loading of WAEC 2018 Multiple Choice Test Items

ITEMS	MR1	MR5	MR2	MR6	MR3	MR4	MR7
1	0.609						
2							0.389
3	0.678						
4	0.427						
5	-0.486						
6							
7	0.518						
8	0.473						
9							
10				-0.306			
11	0.337		0.338		-0.381		
12	0.349						0.385
13	0.572						
14	0.388						0.381
15	0.505						
16	0.324						
17		0.323					
18				0.483			
19							
20		0.39	0.336				
21		0.584					
22							
23							
24							
25		0.57					
26		0.463					
27							
28							
29							
30		0.413					
31							
32			0.353			0.361	
33					0.537		
34					0.47		
35							
36							
37			0.369				
38							
39							
40			0.306				
41							
42							
43							
44			0.581				
45			0.44			0.327	0.323
46							
47							
48							
49							
50							

51				0.427			
52							
53							
55		0.461					
56						0.345	
57							
58						0.6	
59							
60							0.303

Table 2 shows the factor loading of WAEC 2018 Mathematics multiple choice test items. The table shows there are seven factors been examined from the 60 items administered on the candidates that sat for the examination which indicates that the test is not uni-dimensional in nature. The table shows that some items do not test any of the seven factors identified from the analysis which means they should be removed from the set of items or restructured to be in line with other items that comprised the test. Meanwhile, some items load on more than one factor which suggests they are testing more than what they were intended to test. Those items which are loading on more than a single factor are to be screened and will be retained under the factor that has positive value which is 0.30 and above while the items that has negative loading will be removed because it implies the item has negative contribution to the factor it loaded on.

Table 3: NECO and WAEC 2018 3PL Item Parameters Estimates

ITEMS	**GUESSING**		**DIFFICULTY**		**DISCRIMINATION**	
	WAEC	**NECO**	**WAEC**	**NECO**	**WAEC**	**NECO**
1	0.0000	0.4021	-0.6325	0.5009	1.6811	1849.4017
2	0.0099	0.4669	-55.2126	0.1352	-0.0424	4103.3626
3	0.0000	0.5665	-0.5749	1.0756	2.5891	2.0608
4	0.0237	0.0000	0.4368	-0.5351	1.6179	1.0021
5	0.0010	0.0000	-1.4890	-1.0400	-1.2063	0.9278
6	0.0000	0.6697	0.8945	1.5818	0.8165	204.1595
7	0.0000	0.7389	-0.3072	-1.6711	1.3757	-5.7682
8	0.0000	0.7390	0.4841	-1.2306	0.5619	-88.7023
9	0.0002	0.6147	-5.3246	-1.8102	-0.2869	-47.2326
10	0.0000	0.6347	3.9107	1.3603	0.2569	90.3635
11	0.0000	0.6166	0.2100	-1.3595	0.4248	-620.5106
12	0.1472	0.0000	0.4959	-4.5383	2.5609	0.1420

13	0.0453	0.0000	0.0470	-2.0939	2.9570	0.3630
14	0.1330	0.0000	0.7928	6.4728	2.0660	-0.1153
15	0.0046	0.0000	-0.2379	-4.9231	1.3786	0.2059
16	0.0159	0.0000	0.3308	7.6650	2.2420	-0.1895
17	0.2251	0.0000	0.9898	-3.5096	1.4023	0.1713
18	0.0829	0.6850	1.4662	1.4546	8.9027	473.1786
19	0.1641	0.6703	1.6904	-1.4853	1.4563	-42.4855
20	0.2450	0.0000	0.4447	-1.0406	4.4949	0.4801
21	0.2149	0.3802	0.1865	0.1307	1.4402	1.9123
22	0.0527	0.6003	5.0451	0.6782	0.7399	1512.0430
23	0.1038	0.5509	2.0059	0.7587	1.7452	1.2157
24	0.0000	0.0000	-6.6519	-1.8513	-0.3772	0.5847
25	0.1039	0.0000	0.7311	-1.3988	2.0635	0.6126
26	0.0497	0.0000	0.8905	-4.5237	1.7594	0.1818
27	0.0000	0.0000	-0.3818	-4.4569	1.4119	0.2166
28	0.0046	0.0000	-2.2232	-5.8931	-0.6904	0.1168
29	0.2592	0.5887	1.9126	-0.8310	2.4331	-914.8406
30	0.3665	0.6499	1.3167	-1.1148	18.9692	-489.2578
31	0.0425	0.0000	1.3971	6.6166	2.7732	-0.1337
32	0.1552	0.0000	-2.3881	54.8184	-0.8201	-0.0244
33	0.0000	0.0000	2.7198	-2.4561	1.1086	0.3083
34	0.0000	0.0000	6.7473	-1.7091	0.4302	0.3147
35	0.1768	0.2401	-11.6792	-1.3451	-0.2789	0.2737
36	0.2980	0.0000	1.4029	6.0445	16.5011	-0.1466
37	0.1068	0.3843	-2.7175	1.6053	-1.6863	-0.0759
38	0.0786	0.5331	1.8084	-0.8797	2.6560	-2.0019
39	0.1158	0.0000	1.0822	1.4068	1.5801	-0.4242
40	0.0941	0.0000	1.2042	2.3057	1.8130	-0.4896
41	0.0000	0.0000	2.1832	2.7915	0.6683	-0.6008
42	0.0000	0.7142	10.2218	-1.4346	0.2020	-12.7388
43	0.0660	0.0000	2.0099	-2.3279	0.7789	0.3881
44	0.1544	0.7433	1.3649	1.0007	17.2575	148.5677
45	0.3369	0.0000	1.4715	-1.7782	30.1922	0.5164
46	0.1218	0.5886	1.9545	1.7585	1.9763	9.1181
47	0.1904	0.6295	0.9724	-1.3571	3.0055	-506.8605
48	0.2877	0.7179	1.3144	2.0852	5.9627	13.7824
49	0.1180	0.6520	-42.8319	1.3565	-0.0570	286.5923

50	0.1289	0.0000	1.4552	-2.1464	12.6754	0.2804
51	0.1417		1.4488		14.4618	
52	0.1122		0.7445		1.4865	
53	0.0000		7.4153		0.3549	
54	0.8613		11.5563		-2.3295	
55	0.1969		0.7473		4.1542	
56	0.0000		-1.5754		-0.4495	
57	0.0000		-4.7069		-0.5286	
58	0.0024		-11.1313		-0.0808	
59	0.0000		-10.0759		-0.3086	
60	0.0000		1.8575		1.0955	

WAEC 2018 multiple choice items have three distracters with a key while the NECO had four distracters and a key. Therefore, the guessing threshold for guessing for WAEC is 0.25 since there are four options while for NECO it is 0.20 since there are five options. The result further shows that most of the items have high guessing propensity for both WAEC and NECO multiple choice test items which could be because the distracters are not convincing enough and make it very easy for the candidates to guess right. Again, the acceptable range for difficulty index falls within 0.2 and 0.6 ($0.2 \leq p \leq 0.6$) while discriminating index $D \geq 0.2$. The difficulty index for both WAEC and NECO 2018 multiple choice items for most of the items were not outside the permitted range. However, majority of the items in both examinations discriminate sufficiently between the high and low achievers.

The item parameter for NECO and WAEC 2018 multiple choice items are not comparable because items found to be good in NECO 2018 are only two items which are 4 and 5 whereas WAEC 2018 has only 11 items out of 60 items administered to be good. Therefore, WAEC 2018 items can be said to be better than NECO 2018 multiple choice items.

Table 4: NECO 2018 Differential Item Functioning

<	**Statistics**	**P-value**	**Code**	**Items detected as DIF Items**
1	-157.923	0.0000	***	DIF
2	-123.813	0.0000	***	DIF
3	1.2353	0.2167		No DIF
4	-56.7388	0.0000	***	DIF

5	-53.8852	0.0000	***	DIF
6	32.2262	0.0000	***	DIF
7	86.2519	0.0000	***	DIF
8	7.5961	0.0000	***	DIF
9	78.1379	0.0000	***	DIF
10	52.9403	0.0000	***	DIF
11	85.0488	0.0000	***	DIF
12	0.5299	0.5962		No DIF
13	-10.0655	0.0000	***	DIF
14	-7.1817	0.0000	***	DIF
15	-4.0936	0.0000	***	DIF
16	-8.4343	0.0000	***	DIF
17	-65.2284	0.0000	***	DIF
18	-32.7942	0.0000	***	DIF
19	-1.9716	0.0487	*	No DIF
20	-57.8068	0.0000	***	DIF
21	0.6307	0.5282		No DIF
22	2.0788	0.0376	*	DIF
23	-28.1211	0.0000	***	DIF
24	-0.9392	0.3476		No DIF
25	1.1349	0.2564		No DIF
26	11.0983	0.0000	***	DIF
27	17.2749	0.0000	***	DIF
28	61.5106	0.0000	***	DIF
29	136.9829	0.0000	***	DIF
30	89.5224	0.0000	***	DIF
31	43.3653	0.0000	***	DIF
32	4.4911	0.0000	***	DIF
33	-9.5018	0.0000	***	DIF
34	-17.07	0.0000	***	DIF
35	-35.4481	0.0000	***	DIF
36	-4.649	0.0000	***	DIF
37	-28.39	0.0000	***	DIF
38	20.1975	0.0000	***	DIF
39	0.0141	0.9887		No DIF
40	95.3641	0.0000	***	DIF
41	72.0305	0.0000	***	DIF

42	69.0252	0.0000	***	DIF
43	7.277	0.0000	***	DIF
44	-0.0007	0.9995		No DIF
45	-0.0239	0.9809		No DIF
46	61.523	0.0000	***	DIF
47	18.1635	0.0000	***	DIF
48	72.7212	0.0000	***	DIF
49	5.6073	0.0000	***	DIF
50	5.2178	0.0000	***	DIF
Signif.	0 '***'	0.001 '**'	0.01 '*'	0.05 '.' 0.1 "
Detection Threshold		-1.96 to 1.96 (significance level: 0.05)		

Table 4 above shows that 43 items out of 50 items used in testing the students test male and female differently which are items 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40, 41, 42, 43, 46, 47, 48, 49, and 50. Meanwhile, only seven items do not function differently between male and female which are items 3, 12, 21, 24, 25, 44, and 45. Items are identified as not flagging DIF when the statistics value is between -1.96 and 1.96 whereas any item with statistics outside the specified range is said to be flagging DIF.

Table 5: WAEC 2018 Differential Item Function According to Gender

ITEMS	Statistics	P-value	Codes	Items detected as DIF items
1	-14.1269	0.0000	***	DIF
2	0.9871	0.3236		No DIF
3	-15.3143	0.000	***	DIF
4	-11.2439	0.000	***	DIF
5	7.6705	0.000	***	DIF
6	-4.1078	0.000	***	DIF
7	-11.2524	0.000	***	DIF
8	-2.0221	0.0432	*	DIF
9	1.5139	0.13		No DIF
10	-0.3794	0.7044		No DIF
11	-0.4884	0.6252		No DIF
12	-12.9769	0.000	***	DIF
13	-18.0824	0.000	***	DIF

14	-9.7069	0.000	***	DIF
15	-11.063	0.000	***	DIF
16	-13.7644	0.000	***	DIF
17	-6.954	0.000	***	DIF
18	-5.1742	0.000	***	DIF
19	-1.5499	0.1212		No DIF
20	-15.3898	0.000	***	DIF
21	-10.3364	0.000	***	DIF
22	-1.107	0.2683		No DIF
23	-0.5467	0.5846		No DIF
24	2.654	0.008	**	DIF
25	-12.0086	0.000	***	DIF
26	-10.3085	0.000	***	DIF
27	-13.0397	0.000	***	DIF
28	2.4603	0.0139	*	DIF
29	-0.9456	0.3444		No DIF
30	-3.3084	0.0009	***	DIF
31	-3.9239	0.0001	***	DIF
32	1.6041	0.1087		No DIF
33	-1.9005	0.0574	.	No DIF
34	-0.6319	0.5274		No DIF
35	1.53	0.126		No DIF
36	-1.5378	0.1241		No DIF
37	1.1448	0.2523		No DIF
38	-2.3432	0.0191	*	DIF
39	-5.7036	0.000	***	DIF
40	-2.198	0.028	*	DIF
41	-3.278	0.001	***	DIF
42	0.1202	0.9043		No DIF
43	-2.2237	0.0262	*	DIF
44	-3.4433	0.0006	***	DIF
45	-1.2142	0.2247		No DIF
46	-1.2547	0.2096		No DIF
47	-9.2394	0.000	***	DIF
48	-5.4082	0.000	***	DIF
49	1.472	0.141		No DIF
50	-2.3756	0.0175	*	DIF
51	-1.1369	0.2556		No DIF

52	-8.8992	0.000	***	DIF
53	-0.5153	0.6063		No DIF
54	0.000	1		No DIF
55	-13.0587	0	***	DIF
56	3.7762	0.0002	***	DIF
57	3.1187	0.0018	**	DIF
58	1.1359	0.256		No DIF
59	0.7391	0.4599		No DIF
60	-2.5169	0.0118	*	DIF
Signif.	0 '***'	0.001 '**'	0.01 '*'	0.05 '.' 0.1 "
Detection		-1.96 to 1.96 (significance level: 0.05)		

Table 5 above shows that 37 items out of 60 items used in testing the students test male and female differently which are items 1, 3, 4, 5, 6, 7, 8, 12, 13, 14, 15, 16, 17, 18, 20, 21, 24, 25, 26, 27, 28, 30, 31, 38, 39, 40, 41, 43, 44, 46, 47, 48, 50, 52, 55, 56, 57 and 60. Meanwhile, only seven items do not function differently between male and female which are items 2, 9, 10, 11, 19, 22, 23, 29, 32, 33, 34, 35, 36, 37, 42, 45, 46, 49, 51, 53, 54, 58 and 59. Items are identified as not flagging DIF when the statistics value is between -1.96 and 1.96 whereas any item with statistics outside that range is said to be flagging DIF.

From the 50 items administered by NECO (Table 4), it showed that 43 items function differently between male and female and those items that function differently account for 86% of the total number of items administered whereas out of 60 items administered from WAEC, 37 items function differently between male and female which amount to 61.7% of the items administered. This therefore means that larger proportion of the items administered by the two examination bodies are biased and favour one gender over the other.

Discussion

Findings showed that the mathematics test items of NECO and WAEC 2018 are not unidimensional in nature since some items did not load on any of the 13 factors examined. This means that they did not add anything to the testing that was done. Ayanwale (2017) reported that IRT had three basic assumptions which were uni-dimensionality, local item independence and item characteristic curves. The assumption of uni-dimensionality states that items of a test measure one's ability. Uni-dimensionality means that the items measure one and only one area of knowledge or ability (Ojerinde, 2013). A set

of test items testing bit of knowledge which it logically and sequentially relates to may be expected to be uni-dimensional. Uni-dimensionality does not mean that items must correlate positively with each other. In fact, it is conceivable for all items to correlate negatively with each other and still be unidimensional.

Findings from research question two which asked what were the item parameters (difficulty, discrimination and guessing) of Mathematics test items of NECO and WAEC 2018 using Item Response Theory framework, showed that most of the items in both examinations had high guessing propensity which could be because the distracters were not convincing enough and make it very easy for guessing for the candidates. The difficulty and discrimination index for both WAEC and NECO 2018 multiple choice items for most of the items were not within the range allowed. On item difficulty, Adegoke (2012) reported that the first item characteristic should be determined. Item difficulty is simply the proportion of examinees taking the test, who got an item or answer it correctly. The larger the percentage getting an item correctly, the easier the item is. The higher the difficulty value, the easier the item is understood to be. To compute the item difficulty index, divide the number of examinees answering the item correctly by the total number of examinees answering item. An item answered correctly by 75% of the examinees would have a difficulty index or p-value, of .75, whereas an item answered correctly by 40% of the examinees would have a lower item difficulty or p-value, of .40. A general guideline for the interpretation of an item difficulty index is provided in Table 3. Courville (2004) reported that for the item difficulty, a group that answered the item correctly, and one that did not is created. This statistic focuses on determining the correct respondents or the examinees that get the item right or wrong in a test. In essence, the aim of item discrimination is to eliminate or modify items that do not function well in the tested group.

Findings from research question three which asked how comparable are the item parameters (difficulty, discrimination and guessing) of Mathematics test items of NECO and WAEC 2018 using Item Response Theory framework, showed that the item parameter for NECO and WAEC 2018 multiple choice items were not comparable because items found to be good in NECO 2018 are only two items out of 50, whereas WAEC 2018 has only 11 items out of 60 items administered to be good. Therefore, WAEC 2018 items can be said to be better than NECO 2018 multiple choice items. A study by Obinne (2008) reported that items from the two examination bodies were equally reliable and valid. A study conducted in Nigeria on the effect of guessing on the scores of a sample of students reported consistencies with the position of Lord (1977)

who reported that the classic theory of testing cannot provide a good framework for the reliability of tests (William & Amini, 2012).

Results showed that ordinary scoring had stronger effects on test reliability than negative scoring. In a study, one group was allowed to guess answers, but the other group was not allowed. In the first group, negative scores raised the test reliability. However, no change was observed in the other group by using negative scores (Imam & Mohammadreza, 2016). In another study, negative scores lowered students' scores and led them to be rejected in the interested subject. Also, a significant correlation was observed between students' score before and after applying negative scores. The analysis of students' performance also disclosed that there were factors other than guessing, affecting the choices (Gholami, Panah, & Derakhshan, 2013).

Findings from research question four which asked if NECO Mathematics test items function differentially between Male and Female, showed that 43 items out of 50 items used in testing the students, tested male and female differently. Similarly, findings from research question five which asked if WAEC Mathematics test items function differentially between Male and Female, showed that 37 items out of 60 items used in testing the students, tested male and female differently. On gender issues, Davis (2002) reported that sometimes, items were found to behave differently in distinct groups such as gender or language (such as loading on different dimensions in a multi-dimensional factor analysis, or having largely different mean item scores). In other words, two examinees with the same latent trait value but differing in other characteristics may have differing probabilities of response. The findings of Madu (2012) concluded in a study that thirty-nine (39) items in the mathematics test (stared) were identified as significantly exhibiting differential item functioning between male and female examinees at .05 level of significance while 11 items did not show differential function between male and female examinees.

Findings from research question six which asked how comparable were the differential item functioning of Mathematics test items of NECO and WAEC 2018 based on gender using Item Response Theory framework, showed that from the 50 items administered by NECO, only 43 items function differently between male and female and those items that function differently account for 86% of the total number of items administered whereas out of 60 items administered from WAEC, 37 items function differently between male and female which amounted to 61.7% of the items administered. These therefore

mean that larger proportion of the items administered by the two examination bodies were biased and favoured one gender over the other. For ability testing, Dodeen (2004) reported that DIF was used as an item level performance difference between groups of examinees matched on ability. DIF is typically identified using inferential DIF detection methods. Inferential DIF detection methods used a significance test to determine if an item possesses DIF. The numerical value obtained from the inferential DIF method indicated that an item is more difficult for a particular sub-group than originally intended (Camilli & Shepard, 1994). DIF indicates that a particular subgroup responded more positively to an item than another subgroup. Dodeen & Johansson (2003) reported that the correct answer in the cognitive context was similar to the positive effect of attitude toward the item. Items on ability and attitude assessments may exhibit DIF for several reasons. Assessment developers must evaluate the DIF item to determine the cause of DIF (that is, the source of DIF).

Both NECO and WAEC 2018 Mathematics objective test items were not uni-dimensional. The items of both examinations tested male and female students differently and that the item parameter for NECO and WAEC 2018 multiple choice items were not comparable.

Recommendation

The public examining bodies should be more meticulous with the procedure in test construction, ensuring that the process is never compromised.

References

Adegoke, B. A. (2012). Comparison of Item Statistics of Physics Achievement Test using Classical Test and Item Response Theory Frameworks. *Journal of Education and Practice, 4*(22), 87-96.

Arigbabu, A. A & Mji, A. (2004). Is gender a factor in Mathematics performance among Nigerian pre -service teachers? *Sex Rol* 51(11&12), 749.

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

Courville, T. G. (2004). An Empirical Comparison of Item Response Theory and Classical Test Theory Item/Person Statistics. *Unpublished Ph.D. Dissertation,* Texas A & M University.

Dodeen, H., & Johansson, R.R. (2003). On the consistency of individual classification using short scales. *Psychological Methods*, 12(1), 105-120.

Donnellan, C. (2003). Does sex make a difference? An equalities peak for young people on international women's day. *The Gender Issues*, 64, 14-17.

Emaikwu, S. O. (2012). Issues in Test Item Bias in Public Examinations in Nigeria and Implications for Testing. *International Journal of Academic Research in Progressive Education and Development* 1 (1) 175-187.

Erinosho, Y. E. (2005). Women and science. 36th inaugural lecture, Olabisi Onabanjo University, Ago-Iwoye.

Federal Republic of Nigeria (2014): National Policy on Education, 4th ed., Lagos, NERDC.

Gholami, A., Mojdehipanah, H., and Derakhshan, 2013, the effect of negative score on test results: reporting a case, *Journal of Medical Education and Development,* 1, 49-53.

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

Hambleton, R. K. (2005). Applications of item response theory to improve health outcomes assessment: Developing item banks, linking instruments, and computer-adaptive testing. In: Lipscomb J, Gotay CC, Snyder C, (eds). *Outcomes Assessment in Cancer: Measures, Methods and Applications.* Cambridge: Cambridge University Press. p. 445–464.

Iman, P. & Mohammadreza, P. (2016). The Effect of Guessing on the Parameters of Abilities and Items in Multiple-Choice Tests. *International Journal of Humanities Social Sciences and Education (IJHSSE*), 3, (1), 24-30.

Iroegbu, T. O. (1998). Problem based learning, numerical ability and gender as determinants of achievements problems solving line graphing skills in senior secondary physics in Ibadan. *PhD. Thesis.* University of Ibadan, Ibadan.

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

Laura, A. R. (2006). Why are there so few female physicists? *The Physics Teacher*, 44, 177-180

Madu, B. C. (2012). Analysis of Gender-Related Differential Item Functioning in Mathematics Multiple Choice Items Administered by West African Examination Council (WAEC). *Journal of Education and Practice*, 3 (8), 12.

Nenty, H. J. (2004). *From Classical Test Theory (CTT) to Item Response Theory (IRT): An introduction to a desirable transition. In: OA Afemikhe, JG Adewale (Eds.): Issues in Educational Measurement and Evaluation in Nigeria.* Institute of Education, University of Ibadan, Ibadan, Nigeria, pp.372- 384.

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

Okoro, O.M. (2006). *Measurement and evaluation in education*. Uruowulu-Obosi: Pacific Publishers Ltd.

William, J. U. & Amini, C. M. (2012). The effect of guessing on the test scores. *Mathematical theory and modelling.* 2, 6.