Assessment of the Unidimensionality and Differential Item Functioning in the 2017 West African Examination Council (WAEC) November/December Mathematics Multiple Choice Test Items

Chinelo Blessing Oribhabor, Ph.D
Department of Guidance and Counseling, Faculty of Arts and Education,
University of Africa, Toru-Orua, Bayelsa State
Phone Number: 08034982069
E-mail Address: chiblessing42004@yahoo.co.uk; chinelo.oribhabor@uat.edu.ng

Abstract
This study assessed unidimensionality and occurrence of Differential Item Functioning (DIF) in the 2017 November/ December West African Examination Council (WAEC) Mathematics test items administered in Edo State. The population for the study consisted of all the responses of students who answered the 2017 WAEC November/ December Mathematics multiple choice Examination. A sample of 1,238 Senior School III students' responses to 50 multiple-choice WAEC 2017 Mathematics multiple choice items selected using stratified and random sampling techniques was used in the study. Research instrument used for the study were adopted 50 multiple-choice 2017 WAEC November/ December Mathematics multiple choice test items. Cronbach Alpha technique was used to determine the reliability of the instrument and the reliability coefficient of 0.87 was gotten. Data collected were analysed using Raju Area Measure technique, Chi-square and Principal Component Analysis. The results showed that the 2017 WAEC November/ December Mathematics multiple choice test items is unidimensional. Also, there was occurrence of DIF items in the 2017 WAEC November/ December Mathematics multiple choice test items. Twelve items representing 24% of the 50 items in the Mathematics examination exhibited DIF. Based on the findings, it was recommended among other things that examination bodies should take a deliberate decision to intensify reviewing of items including multiple choice items to determine the extent to which each item meets the assumptions of the IRT model under consideration.

Keywords: Assessing, Unidimensionality, Differential Item Functioning, Mathematics, Multiple Choice Test Items.

Introduction

Testing has become one of the most important parameters by which a society adjudges the product of her educational system. The essence of testing is to reveal the latent ability of the examinee. Testing has been fully accepted in most modern societies as the most objective method of decision making in schools, industries, and government establishments. It is now used for admission, recruitment, promotion, placement, evaluation, guidance, research, and teaching purpose among others (Emaikwu, 2011). Construction of a test or examination is aimed at deriving scores which are interpreted as manifestations of a trait or ability. In order to test whether such an interpretation is feasible, different psychometric criteria have been suggested (Campbell & Fiske, 2009; Loevinger, 2007).

The often neglected aspect which is relevant to the interpretation of test scores is the dimensionality of the underlying items. One of the critical and basic assumptions of measurement theory is that a set of items forming an instrument all measure just one thing in common. This assumption provides the basis of most Mathematical measurement model. Further, to make psychological sense when relating variables, ordering persons in some attribute, forming groups on the basis of some variables or making comments about individual differences, the variable must be unidimensional; that is, the various items must measure the same ability, achievement, attitude, or other psychological variables. Unidimensionality refers to the existence of one underlying measurement construct (dimension) that accounts for variation in examinee responses.

Violating this assumption could severely bias item and ability parameter estimation. Unidimensionality of the items comprising a test score is essential for the soundness of the assessment processes the score is being used in. Without testing for unidimensionality, an interpretation of the test score as representing one dimension is potentially risky.

Moreover, testing for unidimensionality of the items provides general information regarding factorial validity of the test score interpretation and might reveal that the test score needs to be separated into several scores (Stout, 2007).

One of the assumptions of Item Response Theory (IRT) is unidimensionality. Unidimensionality means that the items in a test measure one and only one area of knowledge or ability. A set of items testing bits of knowledge which are logically and sequentially related may be expected to be unidimensional. A unidimensional test may be defined therefore as a test in which all items are measuring the same thing (Lumsden, 2007). Item response models that assume a single latent trait are known as unidimensional. The assumption of a unidimensional latent space is a common one for test constructors to make because they usually want to construct unidimensional tests in order to enhance the interpretability of a set scores (Hambleton, 2009).Therefore, misdiagnosis or misrepresentation of the dimensional structure can impact model parameter estimates including person ability estimates (i.e., student scores). The dimensional structure of a test is also used to provide one type of validity evidence based upon the internal structure of a test. Validity refers to the degree to which evidence and theory support the interpretations of test scores, it is a fundamental consideration in test development. The various methods of assessing unidimensionality of items are Cronbach Alpha, Raju's alpha, Item-test correlation, Principal Component Analysis (percent variance, Number of given values>1), Number of residuals > 0.00, chi-square (1 factor), Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA),Theta, Nonlinear factor analysis, one and two parameter latent trait model and Lord's chi-square.

The study of DIF has become an integral part of determining the validity and reliability of standardized tests. In measurement, an item is biased if "its construction, setting, language, idea or interest portrayed, picture/diagram used, relevance and illustration are giving an undue advantage or disadvantage to a particular group of testees over the other group" (Nenty, 2008). These are the most likely sources of differential item functioning. In the context of tests, Test items should not behave differently for particular subgroups of test takers. If an item functions differently for certain groups, the item reduces the validity of the measure for that construct, and test fairness is threatened.

The Research measurement model enables the detection of test items which are biased toward different subgroups according to construct irrelevant factors, such as ability, gender, and ethnicity, subgroups, by calculating differential item functioning (DIF) measures. DIF occurs when people who have the same ability level but from different groups have a different probability of a correct response. According to Item Response Theory (IRT), DIF occurs when item characteristic curves (ICC) of two groups are not identical or do not have the same item parameters after rescaling (Baştuğ, 2016).If, for example, in a mathematics test, boys display higher probability of answering correctly more often than girls of equal ability level because the contents in the test items are biased against girls, then the items are said to exhibit DIF and should be considered for modification or removal from the test. Differential item functioning of an item can therefore be understood as a lack of conditional independence between an item response and group membership (often gender, location or ethnicity) given the same latent ability or trait.

When standardized tests are administered on test takers, the test-taking population could vary on a number of personal and educational characteristics such as age, gender, first language, environment, and academic discipline. From the researcher's personal experience and observations, some test developers do not always take into cognizance the diversities that characterized the test takers before administering such test. This could result into various kinds of errors especially scoring error that inflates scores for one group at the expense of the other. Consequently, such test may be regarded as unreliable or lack test fairness. There are several methods to detect if items have DIF effects. Some of the methods include the Classical Test theory methods which include standard mean difference (SMD) techniques, Generalized Mantel-Haenszel (GMH)methods, chi-squares techniques, analysis of variance methods, methods of comparing plots of transformed item difficulties, factor analysis methods, correlation, logistic regression, log-linear method, methods based on experimental manipulations (Obinne and Amali, 2004) and methods based on Item Response Theory which include Item Characteristic Curve (ICC) method, Raju Area Index, b-parameter method (Oribhabor, 2015). Item response theory (IRT) techniques are theoretically preferred procedures for detecting DIF because they least confound real mean differences in group performance with bias (Obinne and Amali, 2004). There are several advantages of using the IRT approach in testing DIF effects. IRT approaches represent an improvement over the classical approaches in latent trait parameter in variance. With the traditional approach, changes in the examinee sample yield unpredictable differences in the item statistics.

A second advantage is that item response theory is less likely to artificially label an item as biased, unlikely in the Classical Test Theory (CTT) approach where a large p-value difference and item by group interaction may label an item as biased when in fact no bias exists.

Öztürk-Gübeş and Kelecioğlu (2016) examined the impact of dimensionality, common-item set format, and different scale linking methods on preserving equity property with mixed-format test equating. Item response theory (IRT) true-score equating (TSE) and IRT observed-score equating (OSE) methods were used under common-item non-equivalent groups design. A simulation study was conducted based on actual item parameter estimates obtained from the TIMSS 2011 8th grade mathematics assessment. The results showed that: (i) The FOE and SOE properties were best preserved under the unidimensional condition, were poorly preserved when the degree of multidimensionality was severe. (ii) The TSE and OSE results, which were provided by using a mixed-format common-item set, preserved FOE better compared to equating results, which provided only a multiple-choice common item set. (iii) Under the unidimensional and multidimensional test structure, characteristic curve methods performed significantly better than moment scale linking methods in terms of preserving FOE and SOE properties. Hagell (2014) in his study tested the unidimensionality of test items within the Rasch model. The researcher explored the impact of sample size and method of estimating the 95% binomial CI upon conclusions according to recommended conventions. From the researcher's finding, he opined that the PCA/*t*-test protocol should not be viewed as a "definite" test of unidimensionality and does not replace an integrated quantitative/qualitative interpretation based on an explicit variable definition in view of the perspective, context and purpose of measurement. Samantha (2008) analyzed the dimensional structure of

mathematical achievement tests aligned to National Council of Teachers of Mathematics (NCTM) content strands using four different methods for assessing dimensionality. The effect of including off-grade linking items as a potential source of dimensionality was also considered. The result indicates that although mathematical achievement tests for Grades 3-8are complex and exhibit some multidimensionality, the sources of dimensionality are not related to the content strands or the inclusion of several off-grade linking items. The complexity of the data structure along with the known overlap of mathematical skills suggest that mathematical achievement tests could represent a fundamentally unidimensional construct.

Ubi, Joshua and Umoinyang (2012) sampled from a pool of examination scripts of candidates who sat for the Joint Admissions and UME in Cross River State, Nigeria for the years 2002 and 2003. The purpose of the study was to assess the dimensionality of Mathematics items using factor analysis. Results showed that JAMB-UME test revealed five significant dimensions and they concluded that examinations designed for selection of candidates might not be purely unidimensional, especially when items are fielded from a wide syllabus. Robin, Zenisky and Hambleton (2003) study was (1) to identify gender DIF in a large scale science assessment, and (2) to look for trends in the DIF and non-DIF items due to content, cognitive demands, item type, item text, and visual-spatial/reference factors. To facilitate the analyses, DIF study was conducted at three grade levels, and for two randomly equivalent forms of the science assessment at each grade level (administered in different years). A variant of the standardization procedure was applied to very large sets of data (six sets of data, each involving 60,000 students), and has the advantages of being easy to understand and to explain to practitioners. Adedoyin (2010) carried out a study using IRT approach to detect gender biased items in public examinations. The author randomly selected a sample of 4,000 students' (2000 males and 2000 females) response to Mathematics Paper 1 of the Botswana Junior Certificate Examination which were selected from the 36, 000 students who sat for the examination. The examination paper consisted of 38 items. To detect gender bias items, test generated the item characteristics curves (ICC for the male/female). The study compared the ICC curves for the male and female groups, and found that, out of 16 test items that fitted the 3PL item response theory (IRT) statistical analysis, 5 items were gender biased.

The importance of Mathematics in national development is so high that the Federal Republic of Nigeria enshrined Mathematics in the National Policy on Education as a core (compulsory) subject for all secondary schools students in Nigeria (FRN, 2014). Adedayo (2017) stated that knowledge of Mathematics promotes the habit of accuracy, logical, systematic and orderly arrangements of facts in the individual learner. According to him, it also encourages the habit of self-reliance and assists learners to think and solve their problems themselves. Mathematical knowledge indeed equips individuals with the skill to solve a wide range of practical tasks and problems they may encounter in life. Mathematics is a major and pre-requisite subject for gaining admission into higher institution of learning these days, it is important to examine the unidimensionality and DIF technique that can be used to determine the degree to which the subject is free of DIF across different groups of examinees (male and female students). This may be necessary at this time especially considering the major challenges faced by students in passing the subject. There are many methods for DIF detection proposed over the past two decades. This study focused on Raju Area Index method because of its strength and power in detecting DIF items; moreover, Raju Area Index method is one of the IRT techniques which has least confound real mean differences in group performance with bias. Raju (1988) formula for area index between two curves is shown below:

$$Area = \left| 2\frac{(a_2 - a_{1)}}{Da_1\, a_2}\, L_n\left[1 + e^{Da_1 a_2}\, \frac{b_{2\_} b_1}{a_2 - a_1}\right] - (b_2 - b_1)\right|$$

Where:

a_1: discrimination parameter for males (reference group)

a_2: discrimination parameter for females (focal group)
b_1: difficulty parameter for males (reference group)
b_2: difficulty parameter for females (focal group)
D = 1.7 (constant: scaling factor)

The difference is obvious if the area is larger than 0.22 (Raju, 1988). Raju Area index is positioned on the premise that when an item is not revealed differential item functioning, the item characteristic curves for two subgroups are identical and the area between the curves is zero.

Statement of the Problem

The critical nature of educational challenges in Nigeria is evident in the increasing poor performance of students in the national examination. West African Examination Council (WAEC GCE) November/December revealed that less than 30% of school students passed Mathematics in 2017 (WAEC, 2018). If this is allowed to continue, the fear is that the country may not achieve the vision 2020, which is basically anchored on education. This is consequent upon the fact that without a credit pass in mathematics, learners will not be able to proceed to higher educational institutions where highly skilled work force in Science, Technology, and Engineering needed for today's global economy are produced. In addition, despite the huge amount being expended by the Federal Government, students' performance in public examinations has been generally unsatisfactory, especially Mathematics which is a core subject. Given that the teachers and students have put in efforts in academic preparation because of the high stake attached to the examination, it is important to address the quality of the test items used for the state exam by examining its unidimensionality. In unidimensionality, all the items on a test must measure a single latent trait of the examinee, and violation of this assumption would lead to serious misleading results and moreover, it will make the test items not to be fair to group of the examinees.

A test is supposed to measure students/examinees ability/performance or other traits of interest irrespective of certain factors such as gender, ethnicity, geographical location, social status and others. In other words, a test item by IRT standards is supposed to be invariant in nature. This is not always the case, for psychometricians have often found some test items to have interactions with the characteristics of the sample (examinee/students). Therefore, it is pertinent for this study to direct attention towards examining the characteristics of the test items and also to find out the differential functioning (DIF) of the test items administered by the West African Examination Council (WAEC).

Purpose of the Study

The study was designed to assess the unidimensionality and occurrence of Differential Item Functioning (DIF) in the 2017 WAEC November/ December Mathematics multiple choice items. This was with a view to improving the quality of test items to ensure valid decisions. The objectives of the study are to:

(a) determine the dimensionality of the items in 2017 WAEC November/ December Mathematics multiple choice test items.

(b) establish the occurrence of DIF in the 2017 WAEC November/ December Mathematics multiple choice test items in terms of gender.

Research Questions

The following research questions were raised from the above stated objective.

1. What is the dimensionality of the 2017 WAEC November/December Mathematics multiple choice Examination?

2. Does DIF Exist in the2017 WAEC November/ December Mathematics multiple choice Examination?

3. Is there a difference in the number of items functioning differentially in the 2017 WAEC November/ December Mathematics multiple choice Examination in terms of gender?

Hypothesis

There is no significant difference in the number of items functioning differentially in the 2017 WAEC November/ December Mathematics multiple choice in terms of gender

Methodology

The research design adopted for the study was ex-post-facto. The population for the study consisted of all the responses of students who answered the 2017 WAEC November/December Mathematics multiple choice Examination. A sample of 1,238 Senior School III students' responses to 50 multiple-choice WAEC 2017 Mathematics multiple choice items was used in the study. The sample size for the study was selected using stratified and random sampling techniques. Research instruments used for the study were adopted 50 multiple-choice 2017 WAEC November/December Mathematics multiple choice test items. The researcher administered the instruments to the students and their answer scripts were collected and scored. The correct responses were coded "1", while the wrong options were coded "0". Cronbach Alpha technique was used to determine the reliability of the instruments and the reliability coefficients of 0.87 was gotten. Raju Area Measure technique was used to establish the presence of DIF in the items. In Raju technique, an item is reported to possess DIF when the area index is greater than a critical value of 0.22, while an item does not possess differential item functioning when the area index is zero or close to zero (De Beer, 2004). Also, according to Ling and Lau (2003), when the b parameter (item difficulty) for one group (for example, Male) is greater than the other group (for example, Female), this shows that the item is more difficult for the male group and the item is said to favour the other group (that is, female), and vice versa. Chi-square and Principal Component Analysis were used to analyze the data using Microsoft excel and SPSS version 22.

Presentation of Data

Research Question One: What is the dimensionality of the 2017 WAEC November/ December Mathematics Examination?

Figure 1 shows the scree plot for the 50 multiple-choice 2017 WAEC November/December Mathematics Examination items. The factor analysis that was performed on the items using extraction method of principal component analysis (PCA) showed that the first factor having the initial given value (15.580) which clearly exceeded that of the second factor (7.582) is also revealed in Table 1. From Figure one, the Scree plot showed a visual of the total variance associated with each factor. The steep slope showed the large factors associated with the loading greater than the given value of 1. The gradual trailing off (scree) showed the rest of the factors lower than the given value of 1. There are nine factors whose values are greater than given value of l and one extracted commonality factor distinctly higher than others, showing that the test is unidimensional in nature. It can therefore be concluded that the 50 multiple-choice mathematics items are unidimensional.

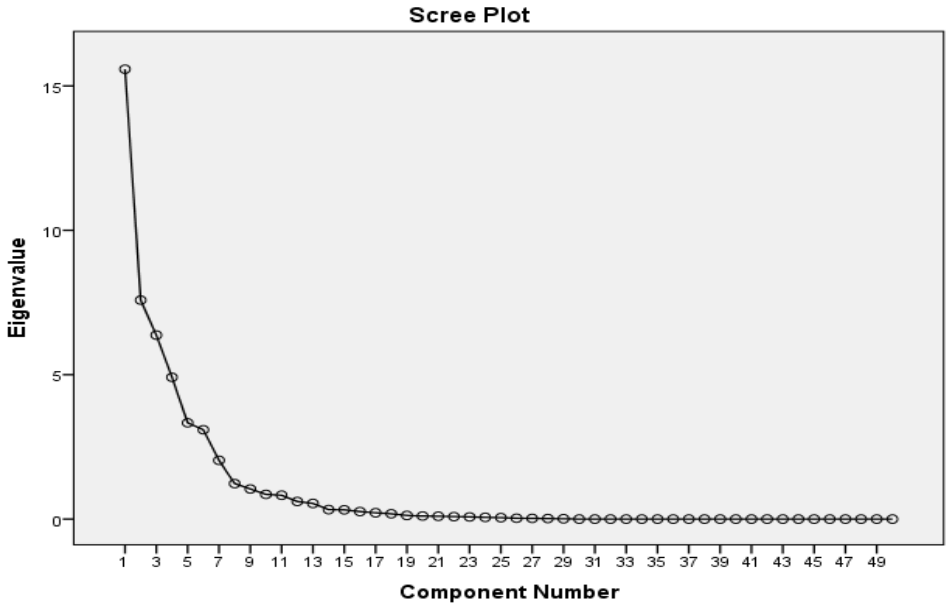

Figure 1: Scree plot of 2017 WAEC November/ December Mathematics multiple choice Examination

Table 1:Total Variance Explained

Component	Initial Given Values			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	15.580	31.160	31.160	15.580	31.160	31.160
2	7.582	15.164	46.324	7.582	15.164	46.324
3	6.371	12.742	59.066	6.371	12.742	59.066
4	4.910	9.821	68.887	4.910	9.821	68.887
5	3.333	6.666	75.553	3.333	6.666	75.553
6	3.091	6.183	81.736	3.091	6.183	81.736
7	2.034	4.068	85.804	2.034	4.068	85.804
8	1.229	2.458	88.262	1.229	2.458	88.262
9	1.043	2.086	90.348	1.043	2.086	90.348
10	.855	1.710	92.058			
11	.826	1.652	93.710			
12	.608	1.217	94.926			
13	.541	1.083	96.009			
14	.328	.656	96.665			
Component	Initial Given Values			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
15	.321	.642	97.307			
16	.260	.520	97.827			
17	.218	.436	98.263			
18	.184	.368	98.631			
19	.127	.253	98.884			
20	.106	.212	99.096			
21	.097	.195	99.291			
22	.087	.173	99.464			
23	.076	.151	99.615			
24	.061	.121	99.736			
25	.048	.097	99.832			
26	.029	.058	99.890			

27	.024	.047	99.937			
28	.020	.040	99.978			
29	.011	.022	100.000			
30	5.493E-015	1.099E-014	100.000			
31	2.335E-015	4.671E-015	100.000			
32	1.917E-015	3.834E-015	100.000			
33	1.516E-015	3.032E-015	100.000			
34	1.085E-015	2.169E-015	100.000			
35	9.433E-016	1.887E-015	100.000			
36	7.434E-016	1.487E-015	100.000			
37	3.127E-016	6.255E-016	100.000			
38	1.869E-016	3.737E-016	100.000			
39	8.489E-017	1.698E-016	100.000			
40	2.386E-017	4.772E-017	100.000			
41	4.289E-018	8.577E-018	100.000			
42	-2.303E-017	-4.606E-017	100.000			
43	-3.355E-017	-6.710E-017	100.000			
44	-3.130E-016	-6.259E-016	100.000			
45	-3.814E-016	-7.629E-016	100.000			
46	-5.875E-016	-1.175E-015	100.000			
47	-8.826E-016	-1.765E-015	100.000			
48	-1.249E-015	-2.498E-015	100.000			
49	-1.702E-015	-3.403E-015	100.000			
50	-3.042E-015	-6.085E-015	100.000			

Extraction Method: Principal Component Analysis.

Moreover, according to Reckase (1979), the variance explained by the first factor should be greater than 20% as to be indicative of unidimensionality. The variance explained in this study (Table 1) exceeded the requirement of this criterion, demonstrating a unidimensional trait of the data.

Research Question Two: Does DIF Exist in the 2017 WAEC November/ December Mathematics multiple choice Examination?

To answer this question, Raju Area Index method with critical value of 0.22 was used to establish the presence of DIF in the 2017 WAEC November/ December multiple choice Mathematics examination.

Table 2: Summary of Results from the Raju Area Index Method of Detecting Differential Item functioning in the 2017 WAEC November/December multiple choice Mathematics examination

Item	b1(male)	b2(female)	Index Area	Decision	Favoured Group
Item 1	-2.920	-2.410	5.78	DIF	Male
Item 2	2.890	3.110	0	NO DIF	-
Item 3	-0.970	-0.640	0.92	DIF	Male
Item 4	1.520	3.340	1.86	DIF	Male
Item 5	-0.180	-0.500	0	NO DIF	-
Item 6	1.080	1.790	-0.04	NO DIF	-
Item 7	2.490	1.240	0	NO DIF	-
Item 8	2.100	6.960	0	NO DIF	-
Item 9	4.460	5.800	7.14	DIF	Male
Item 10	1.140	0.550	-7.84	NO DIF	-
Item 11	-2.150	5.090	-5.68	NO DIF	-

Item 12	-5.550	-4.750	3.59	DIF	Male
Item 13	-3.110	-2.600	3.51	DIF	Male
Item 14	-5.550	-4.750	3.59	DIF	Male
Item 15	0.860	0.460	-0.94	NO DIF	-
Item 16	-0.280	-0.620	-13.52	NO DIF	-
Item 17	0.050	-0.280	0	NO DIF	-
Item 18	2.590	1.150	0.71	DIF	Female
Item 19	0.130	-0.200	0	NO DIF	-
Item 20	-2.110	-0.700	-0.97	NO DIF	-
Item 21	-4.030	-5.310	0	NO DIF	-
Item	**b1(male)**	**b2(female)**	**Index Area**	**Decision**	**Favoured Group**
Item 22	-1.930	-1.350	1.6	DIF	Male
Item 23	-1.590	2.980	-3.19	NO DIF	-
Item 24	4.950	7.630	0	NO DIF	-
Item 25	-0.870	-1.360	0	NO DIF	-
Item 26	-0.700	2.720	-2.74	NO DIF	-
Item 27	0.610	0.730	3.25	DIF	Male
Item 28	0.940	-0.050	0	NO DIF	-
Item 29	1.140	1.250	0	NO DIF	-
Item 30	2.100	6.960	0	NO DIF	-
Item 31	0.130	0.380	0	NO DIF	-
Item 32	2.500	2.730	0.41	DIF	Male
Item 33	-0.180	-0.430	0	NO DIF	-
Item 34	-3.870	-1.840	0	NO DIF	-
Item 35	5.150	8.330	-0.85	NO DIF	-
Item 36	-2.150	5.090	-5.68	NO DIF	-
Item 37	3.200	2.000	0	NO DIF	-
Item 38	0.130	-0.280	-1.88	NO DIF	-
Item 39	-0.280	0.080	0	NO DIF	-
Item 40	0.610	0.720	0	NO DIF	-
Item 41	-0.180	0.600	-0.3	NO DIF	-
Item 42	-0.490	3.290	0	NO DIF	-
Item 43	0.630	0.020	-0.58	NO DIF	-
Item 44	6.340	3.990	-5.04	NO DIF	-
Item 45	0.690	2.120	0	NO DIF	-
Item 46	4.950	5.130	9.21	DIF	Male
Item 47	-0.280	-0.510	-12.24	NO DIF	
Item 48	1.140	1.400	0	NO DIF	-
Item 49	-0.180	0.520	0	NO DIF	-
Item 50	0.130	-0.200	0	NO DIF	-

Table 2 revealed that there is occurrence of DIF in the 2017 WAEC November/December Mathematics multiple choice test items. Out of 50 items, 12 items possess DIF (24%) while 38 items (76%) do not possess DIF. The items that possess DIF are items 1, 3, 4, 9, 12, 13, 14, 18, 22, 27, 32 and 46. Out of the 12 items that possess DIF, 11 items favoured male students while only one item (item 18) favoured female students.

Hypothesis One: There is no significant difference in the number of items functioning differentially in the 2017 WAEC November/ December Mathematics multiple choice in terms of gender.

To test hypothesis one, chi-square statistics was used to analyze the data.

Table 3: Chi-square summary of Differential Item Functioning in favour of males and females

Variable	Observed	Expected	df	Chi square	Sig(2-tailed)
Male	11	6			
Female	1	6	1	8.33	0.004
Total	12	12			

$\alpha = 0.05$

The data in Table 3 showed chi-square value of 8.33 and p-value of 0.004. Testing at an alpha level of 0.05, the p-value is less than the alpha value, consequently the null hypothesis is rejected. Therefore, there is a significant difference in the number of items functioning differentially by gender in the 2017 WAEC November/ December multiple choice Mathematics examination.

Discussion of Findings

Differential Item Functioning analysis is recommended only when the test scores are unidimensional (Clauser & Mazor, 1998). Principal component analysis (PCA) method was used to test the unidimensionality of the 2017 WAEC November/December multiple choice Mathematics examination. According to Reckase (1979), the variance explained by the first factor should be greater than 20% for it to be indicative of unidimensionality. The variance explained in this study was 31.16% which exceeded the requirement of this criterion, demonstrating a unidimensional trait of the data. Result from research question two revealed that there is occurrence of DIF in the 2017 WAEC November/December Mathematics multiple choice test items in which out of 50 items, 12 items possess DIF (24%) while 38 items (76%) do not possess DIF. The result of this finding is in agreement with the findings of Omorogiuwa and Iro-Aghedo (2016), who examined the presence of DIF on the 2015 NABTEB Mathematics multiple choice items in terms of gender. The Raju Area Measure technique was used to determine items that functioned differentially. The finding showed that there is existence of DIF in the test items and seventeen out of fifty test items representing 34% exhibited DIF. Result from hypothesis one showed that there is a significant difference in the number of items functioning differentially by gender in the 2017 WAEC November/ December multiple choice Mathematics examination in favour of male group while the findings of Omorogiuwa and Iro-Aghedo (2016) revealed that that there was no significant difference in the number of items functioning differentially between the male and female students.

Conclusion

The study concluded that the 2017 WAEC November/December multiple choice Mathematics examination measured a single construct which showed evidence of unidimensionality. The study also revealed that the 2017 WAEC November/ December multiple choice Mathematics examination exhibited DIF items. The study showed that undimensionality of test items is a necessary condition for DIF analysis. It also showed that

the detection DIF in multiple-choice items will help test developers to generate quality items that will subsequently ensure correct interpretations of test scores. Test practitioners should endeavour to perform DIF analysis from a pilot study before administration of test(s) so that items that function differently for different test taking groups can be identified for possible replacement.

Recommendations

Based on the conclusion, the following recommendations were posed:

- Examination bodies should take a deliberate decision to intensify reviewing of items including multiple choice items to determine the extent to which each item meets the assumptions of the IRT model under consideration. This would enable them to produce quality items for criterion-referenced decision which is what the examination bodies are currently using in grading students.

- Teacher training institutions should expose pre-service teachers into the test development which meet IRT assumptions particularly unidimensional. It should be stressed to trainee teachers that the syllabus is central to all assessment and advocate IRT test analysis which will also detect DIF items.

- The Nigerian government through the Ministry of Education and Skills Development should solicit for donor funding that would be specially for teachers and examiners on test construction or item writing and modern test analysis. This would enable them to be in position to produce valid, reliable and fair assessment tools which are bias free.

References

Adedayo, O. (2017). *Mathematics Phobia, Diagnosis and Prescription*. National Mathematical Centre, 1st Annual Lecture, Abuja.

Adedoyin, O. O. (2010) Using IRT Approach to Detect Gender Biased Items in Public Examinations: A Case Study From the Botswana Junior Certificate Examination in Mathematics. *Educational Research and Reviews, 5*(7), 385-399.

Baştuğ, O.Y.O. (2016). A Comparison of Four Differential Item Functioning Procedures in the Presence of Multidimensionality. *Educational Research and Reviews*, *11*(13): 1251-1261.

Campbell, D. T. & Fiske, D. W. (2009). Convergent and Discriminant Validation by the Multitrait-Multi-Method Matrix. *Psychological Bulletin*, 56, 81-105.

Clauser, B.E. & Mazor, K.M. (1998). Using Statistical Procedure to Identify Differentially Functioning Test Items. *Educational Measurement: Issue and Practice, 17,* 31-44.

De Beer, M. (2004). Use of Differential Item Functioning (DIF) Analysis for Bias Analysis in Test Construction. *SA Journal of Industrial Psychology,30*(4): 52-58.

Emaikwu, S.O. (2011). *Evaluation of Students' Ability in Schools*. A Paper Presented at a Workshop on Teaching Practice on Friday, 29th July in the College of Agricultural and Science Education, Federal University of Agriculture, Makurdi, Benue state.

Federal Republic of Nigeria (2014). National Policy on Education (Revised) Lagos: National Educational Research Council Press.

Hagell, P. (2014) Testing Rating Scale Unidimensionality Using the Principal Component Analysis (PCA)/*t*-Test Protocol with the Rasch Model: The Primacy of Theory over Statistics. *Open Journal of Statistics*, 4, 456-465.

Hambleton, R. K. (2009). Translating and Adapting Tests for Cross-Cultural Assessment. Retrieved from www.researchgate.net on May 10, 2018.

Linacre, J. M. (1998). Structure in Rasch Residuals: Why Principal Components Analysis? *Rasch Measurement Transactions, 12*(2), 636.

Linacre, J. M. (2010). *A Users' Guide to Winsteps Rasch Model Computer Program: Program Manual 3.70.* Chicago: Winsteps.

Ling, S.E. & Lau, S.H. (2004). Detecting Differential Item Functioning (DIF) in Standardized Multiple-Choice Test: An Application of Item Response Theory (IRT) Using Three Parameter Logistic Model. *Journal of Applied Psychology, 94*(7): 452-459.

Loevinger, J. (2007). Objective Tests as Instrument of Psychological Theory. *Psychological Reports*, 3, 631-694.

Lumseden, S. E. (2007). A Priori Considerations in Choosing an Item Response Model. In R. K. Hambleton (Ed.), *Applications of Item Response Theory* (pp. 57-70). Vancouver, Canada: Educational Research Institute of British Columbia.

Nenty, H.J. (2008). Cross-cultural Bias Analysis of Cattell Culture-Fair Intelligence Test. *Perspectives in psychological researches,9*(7): 1-16.

Obinne, A. D. E. & Amali, A. O. (2004). Differential Item Functioning: The Implication for Educational Testing in Nigeria. *International Review of Social Sciences and Humanities, 7*(1):52-65.

Omorogiuwa, K. O. & Iro-Aghedo, P. E (2016). Determination of Differential Item Function by Gender in the NABTEB Mathematics Multiple Choice Examination. *International Journal of Education, Learning and Development, 4*(10): 25-35.

Oribhabor, C. B. (2015). *Determination of Differential Item Functioning in Edo State Basic Education Certificate Examination Mathematics Test Items Using Item Response Theory Analytical Procedure.* A Ph.D thesis, Department of Educational Psychology and Curriculum Studies, Faculty of Education, University of Benin, Benin City.

Öztürk-Gübeş, N. & Kelecioğlu, H. (2016). The Impact of Test Dimensionality, Common-Item Set Format, and Scale Linking Methods on Mixed-Format Test Equating. *Educational Sciences: Theory & Practice,* 16, 715-734.

Raju, N.S. (1988). The Area Between Two Item Characteristic Curves. *Psychometrika,* 53, 495-502.

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multi-Factor Tests:Results and Implications. *Journal of Educational Statistics*, 4, 207-230. Du

Robin, F.; Zenisky, A.L & Hambleton, R. K. (2003). *DIF Detection and Interpretation in Large Scale Science Assessments: Informing Item Writing Practices*. University of Massachusetts, Amherst and Frederic Robin Educational Testing Service.

Samantha, S. B. (2008). *An Investigation of Dimensionality Across Grade Levels and Effects on Vertical Linking for Elementary Grade Mathematics Achievement Tests*. Durham: MetaMetric Inc.

Sick, J. R. (2009a). Rasch Measurement in Language Education Part 3: The Family of Rasch Models. *SHIKEN, 13*(1), 4-10.rha

Stout, W. (2007). A nonparametric Approach for Assessing Latent Trait Dimensionality. *Psychometrika, 52,* 589-617.

Ubi, I. O.; Joshua, M. T. & Umoinyang, I. E. (2012) Assessment of Dimensionality of Mathematics Tests of University Matriculation Examination in Nigeria: Implications for Regional Development. *Journal of Educational Assessment in Africa,* 7, 122-130

WAEC (2018). WAEC releases Nov/ Dec 2017 GCE Results. Retrieved from www.pmnewsnigeria.com/2017/11/21/waec-releases-novdec-2017-gce-results/on May 9, 2018.

Wright, B. D.; Linacre, J. M.; Gustafson, J. E. & Martin-Löf, P. (2004). Reasonable Mean-Square Fit Values. *Rasch Measurement Transactions, 8*(3), 370.m