Journal of Multidisciplinary Studies
Vol. 9, Issue No. 1, pp. 67-92, August 2020
ISSN 2350-7020 (Print)
ISSN 2362-9436 (Online)
doi: http://dx.doi.org/10.7828/jmds.v9i1.1451

Differential Item Functioning of Mathematics Joint Mock Multiple-Choice Test Items in Kwara State, Nigeria

Mohammed Idris Jimoh[1], Dorcas Sola Daramola[2], Jumoke Iyabode Oladele[3], Adaramaja Lukman Sheu[4], Mayowa O. Ogunjimi[5]

[123]Department of Social Sciences Education,Faculty of Education,
University of Ilorin Ilorin, Kwara State
[4]Department of Educational Foundation, Faculty of Education,
Federal University Gusau, Gusau, Zamfara State
[5]Department of Adult and PrimaryEducation Studies, Faculty of Education,
University of Ilorin, Ilorin, Kwara State
Corresponding author: Jumoke Iyabode Oladele, email: oladele.ji@unilorin.edu.ng

Abstract

Test in educational settings is one of the assessment techniques to measure and compare examinees' abilities. This study examined differential item functioning of mathematics joint mock multiple-choice test items conducted by the Kwara State Ministry of Education and Human Capital Development. The descriptive research design of the survey type was adopted in carrying out this study. The population comprised all senior secondary students, while the target consists of senior secondary II students in Kwara State. The sampling procedure used was multi-stage in stratified and simple random sampling techniques at different selection stages and, therefore, sampled 1,062 examinees. The measuring device used for the data collection was the 2018/2019 academic session Joint Mock Mathematics Multiple-choice items that contained 50 items. To validate the measuring device, the Item Level Content Validity (I-CV) was calculated and obtained a coefficient of 0.91. Two research questions were prepared and answered using Mantel-Haenszel chi-square. The first finding revealed that 16 items were flagged DIF, 12 items were for reference (male), and four were for focal (female) group. The second finding also showed that out of 20 items that were flagged DIF, eight were for reference (urban schools) group, and 12 were for focal (rural schools). While carrying out this research, the researchers observed that the quality of the assessment device being used is another important factor that hinders students' performance in mathematics at the school and external examinations levels. It was recommended that in the construction of any test items, a test developer must ensure that irrelevant clues are avoided that may allow examinees to interpret tests differently.

Keywords: assessment device, Item Level Content Validity, rural school, student’s performance, urban school

Introduction

Testing in educational settings is one of the assessment techniques or tools used to measure educational programs. Test items contain a series of questions, items, or tasks to which a group of examinees reacts independently. The result of such can be used to compare examinees' ability. For any result obtained in the test administered to be valid and reliable, fair test items must be generated so that differences in the examinees' ability could be compared. It must be free from any irrelevant clues that will disallow examinees interpret an item the same way if it is noticed that items in the test favor one group of examinees over the other, it implies that the items function differentially among a group of examinees (e.g., male and female) of the same examinees' ability (theta).

Different types of test formats could be used in the school setting, the free-response, and closed-response. In this study, the emphasis is on a closed-response, such as multiple-choice questions. The multiple-choice questions are usually scored dichotomously of either correct or incorrect. In multiple-choice questions, it is assumed that examinees' performance should be at par since they are exposed to the same materials, curriculum, and comparable ability levels. This assumption implies that test items prepared must be investigated to find out the quality test items in the mathematics multiple-choice questions constructed by the Kwara State Ministry of Education and Human Capital Development, which are assumed to have undergone processes of test standardization.

Differential Item Functioning (DIF) is a measurement term or statistical technique used to detect multiple-choice items or questions functioning differently based on the same ability scale or person's latent variable (often denoted by the Greek letter θ). In other words, DIF occurs when two groups (male and female) on the same latent variable are not able to answer a multiple-choice item correctly (Abedalaziz, 2010). DIF also occurs when one group does not have an equal chance of getting a multiple-choice item right, though its members have a comparable latent trait to the other group (Karami, 2012). DIF is used to detect the same ability level whenever examinees but from different groups have varying probabilities of answering an item correct. However, if an item is found functioning differently, it will be flagged as an item displaying DIF. A reasonable assumption of IRT is that each examinee responds to a test item that possesses some underlying ability. This underlying ability can be in a numerical value that is a score that places a test-taker somewhere on the ability scale (-3 to +3). At each ability level, there is a certain probability that a test-taker with a particular ability level will give a correct answer to an item. It must be noted that a test-taker with a low ability level would have a small probability of answering an item correctly in a typical test item. In the same vein, an examinee with a high ability level will have a high probability of answering an item correctly. The probability of correct response tappers to zero at the lowest levels of ability. Modern Item Response Theory, in general, according to Hambleton, Swaminathan & Rogers (1991, cited in Zumbo (1999), is based on two predictions. They are:

1. that a test item examinees performance can be explained from a set of factors also known as "traits, latent traits, or abilities" which varies from one examinee to another: and
2. that the relationship between examinees' performance on an item and the continuum of changes affecting item responses can be described as an Item Characteristic Curve (ICC) also known as the Item Response Function (IRF).

Parametric ICCs vary in terms of positioning on the X-axis, slope, and intercept with the Y-axis. The X-axis is the latent variable, and the Y-axis is the probability of getting the item correct. The ICC gives detailed information on an item. Figure 1 gives a typical example of a parametric ICC (Zumbo, 1999). It increases until the highest levels of ability, the probability of correct response approaches 1 (Baker, 2001). If this assumption is plotted on functional ability levels, the result would be a smooth S-shaped curve shown in Figure 1.

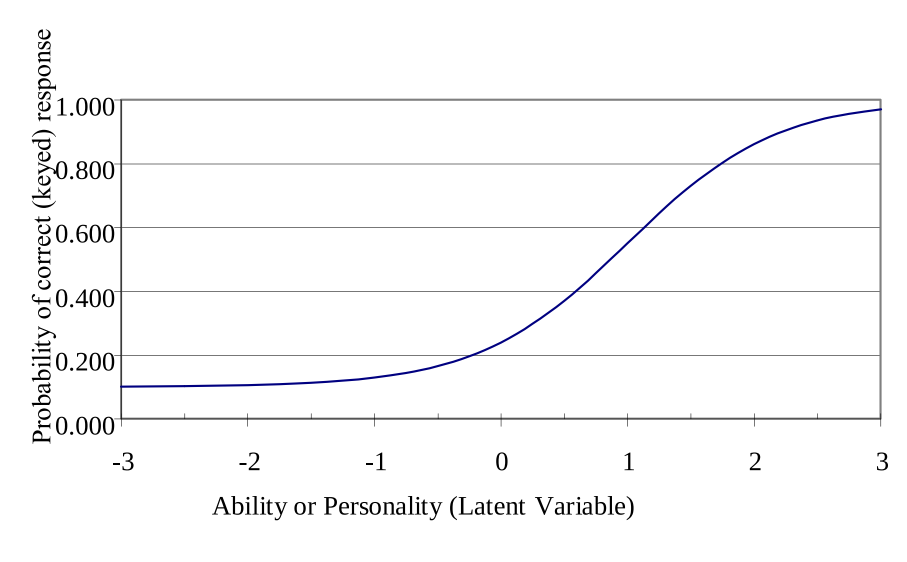

Figure 1. A Typical Item Characteristic Curve Adopted from Zumbo, 1999.

This arched curve describes the relationship between the probability of an item correct response and the ability scale. The ICC is the basic building block of IRT; all the other constructs of the theory depend upon this curve. It must be noted that the item discrimination parameter determines how rapidly the curve rises from its lowest value.

A relatively flat curve indicates that the item does not discriminate among individuals. Item discrimination of a test item determines the extent to which the given item discriminates among test-takers in the function or ability measured by the item. The discrimination index ranges between -1 and +1. Lesser discrimination values are desirable as a highly discriminating item indicates that the students who had high test scores got the item correct. In contrast, students who had low test scores got the item incorrect (Adewole & Ojo, 2017). This finding was elaborated in figure 2.

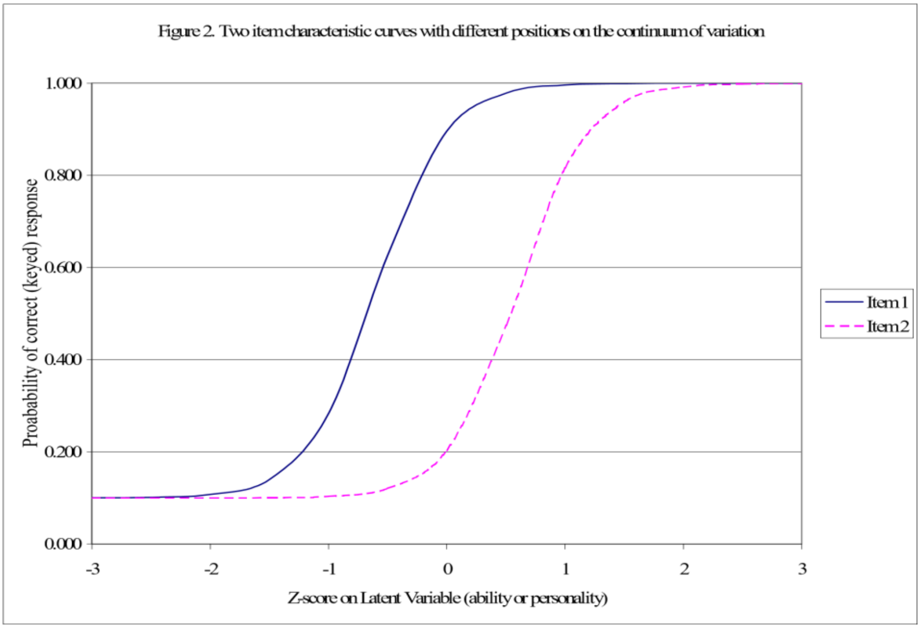

Figure 2. Two ICCs with Different Position on the Continuum variation

The ICCs shown in Figure 2 portray items with similar discrimination indexes among respondents) but different placements on the continuum of variation. More of the latent variable is needed to endorse the item depicted by the dashed line than by the solid line. The dashed line is further to the right on the continuum. The dashed line, item 2, is thus considered more difficult (Zumbo, 1999). Graphically, if the ICCs are identical for each group, or very close to identical, it can be said that the item does not display DIF. If, however, the ICCs are significantly different from one another across groups, then the item is showing DIF. In most contexts, DIF is conceived of as a difference in placement (i.e., difficulty or threshold) of the two ICCS, but as you will see in a few moments, this does not necessarily have to be the case. Some examples of ICCs that do demonstrate DIF and some examples of items that do not demonstrate DIF are presented. Figure 3 presents an item that does not display DIF.

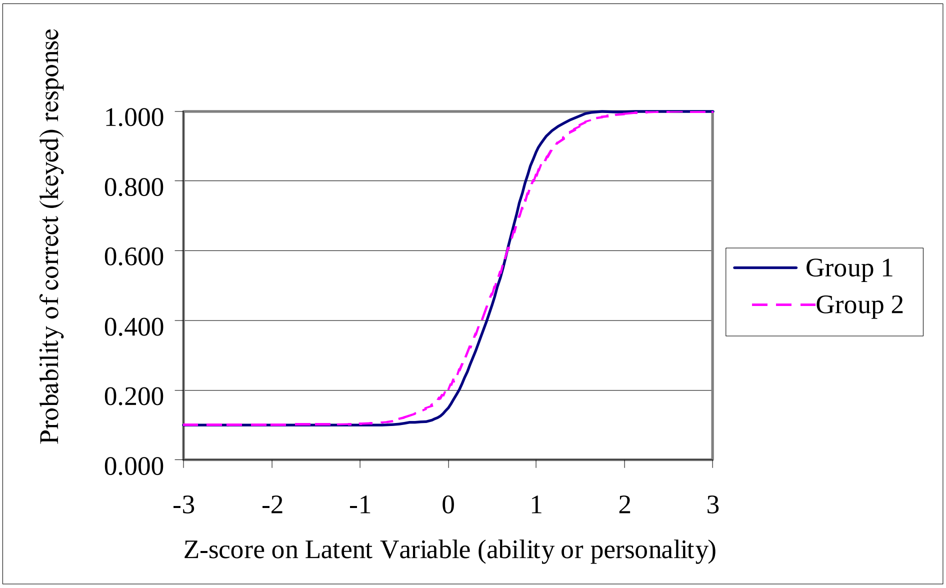

Figure 3. An item that does not display DIF Adopted from Zumbo, 1999.

As shown in figure 3, the area between the curves is minimal, and the parameters for each curve would be nearly equivalent. Figure 4, on the other hand, gives an example of an item that displays substantial DIF with an extensive area between the two ICCs.

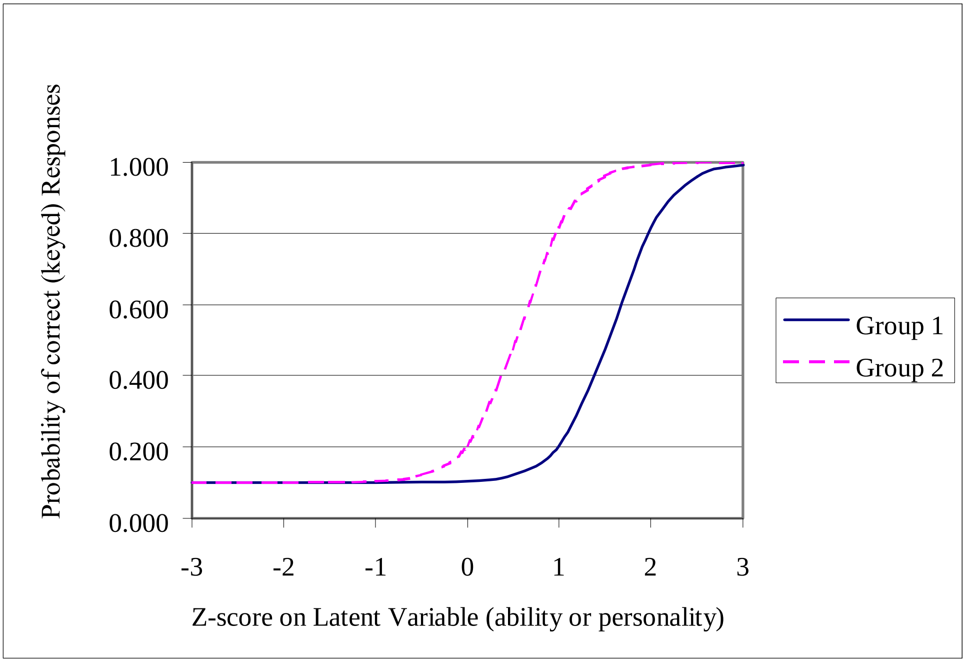

Figure 4. An item that displays substantial uniform DIF Adopted from Zumbo, 1999.

This type of DIF is known as uniform DIF because the ICCs do not cross. An item such as the one shown in Figure 4 may not be an equivalent measure of the same latent variable for both groups. As shown in figure 5, an item that displays substantial non-uniform DIF (i.e., the ICCs cross over one another) depicts non-uniform DIF because those individuals who score at or below the mean (i.e., $z \leq 0$).

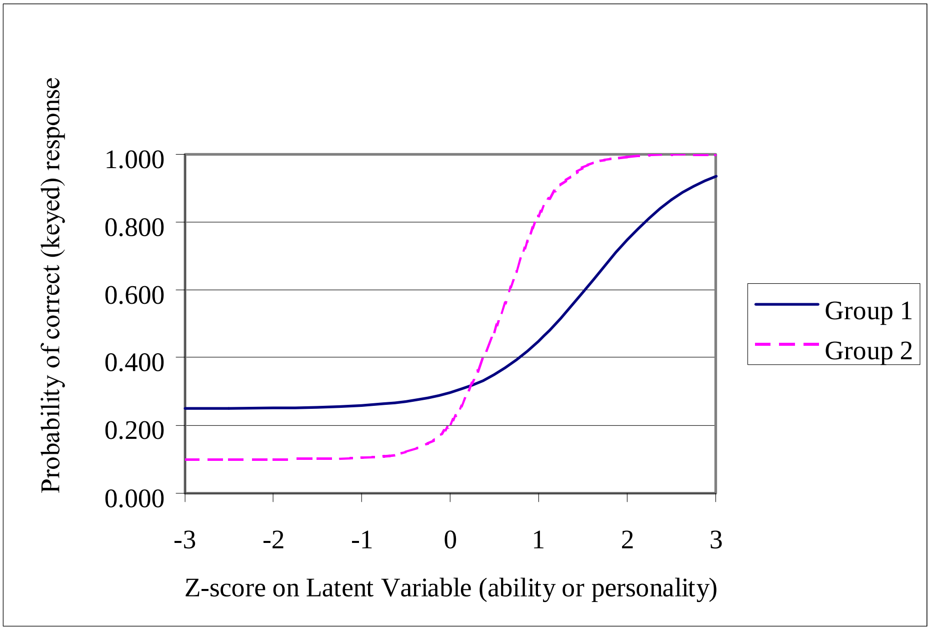

Figure 5. An item that displays substantial Non-uniform DIF Adopted from Zumbo, 1999

Figure 5 shows that group 1 is favored for the scoring below the mean (i.e., z ≤ 0), whereas group 2 is favored for those scoring above the mean (i.e., z > 0).

Several methods have been proposed for detecting Differential Item Functioning (DIF). Among them is the Mantel-Haenszel statistic (1959). Swaminathan and Rogers (1990) proposed a new procedure for detecting both uniform and non-uniform DIF based on a logistic regression model; Jöreskog and Goldberger (1975) proposed the MIMIC model to investigate multi-group differences on a latent construct. Thissen et al., (1986) and Thissen et al., (1988, 1993) proposed the likelihood ratio test to evaluate the significance of observed differences in item responses from different groups. Lord (1977, 1980) proposed a $\chi 2$ statistic to test for

DIF detection under IRT, usually called the Wald test used to compare vectors of IRT item parameters between groups Gao, X. (2019). Guilera et al., (2009), GMH is a procedure widely used to detect DIF used on the standard odds ratio. Simultaneous Item Bias Test (SIBTEST) is a DIF procedure proposed by Shealy and Stout (1993) based on the ratio of the weighted difference in proportion correct (for reference and focal group members) to its standard error Ibrahim, (2017). This study adopted Mantel-Haenszel statistic because it matches groups (e.g., male and female) concerning a binary outcome (right or wrong answer), very simple to use, and has no multilevel equivalent like others methods.

Research activities exist in examining the presence of DIF in a test in different regions of the world. For instance, Driana (2007) examined DIF's presence on the Mathematics section of the Ohio Ninth-Grade proficiency test. The researcher investigated those items that functioned differently between male and female students and between Appalachian and non-Appalachian students. The result of the study did not indicate the presence of regional DIF. Madu (2012) analyzed gender-related differential item functioning in multiple-choice mathematics items administered by the West African Examinations Council. He reported that male and female examinees performed differently in 39 items out of 50 items. In another development, Adehule (2013) carried out a study of DIF in Ekiti State Unified Mathematics Examination for Senior Secondary Schools for 2008-2009 and 2009-2010 academic sessions. One of the findings showed that out of the 40 items examined, seven items displayed DIF based on male and female examinees. Ogbebor & Onuka (2013) researched on differential item functioning method as an Item Bias indicator. The findings revealed that out of 60 items in NECO Economics questions, eight items displayed DIF concerning school location. Amuche & Fan (2014) assessed item bias using differential item functioning technique in NECO Biology conducted examinations in Taraba State, Nigeria. The study examined items biased using a differential item functioning approach concerning school type (private and public schools) and school location (urban and rural schools). It used the 2012 National Examinations Council (NECO) Biology questions. The research outcomes revealed that out of sixty items in NECO Biology questions, ten items were biased regarding school type and eight items with regards to school location.

Furthermore, Mokobi & Adedoyin (2014) identified location biased items in the 2010 Botswana Junior Certificate Examination Mathematics Paper 1 using the item response characteristics curves. They identified location-based items in rural and urban schools in the 2010 Botswana Junior Certificate Examination Mathematics Paper 1. One of the findings revealed that from the 24 items that fit the IRT (3PLM) model, six questions were rural /urban location biased items. The study further found out that three (3) questions were rural /urban location biased regarding males, and six (6) questions were rural /urban location biased with regards to females. In the same vein, Ahmadi & Bazvand (2016) investigated gender DIF across the Ph.D. entrance examination of TEFL (PEET) in Iran. One of the findings indicated that more questions were flagged DIF. There was an equal number of uniform and non-uniform DIF. It was also found that there was female superiority in the test performance. The study's motivation was derived from the researchers' field observations that the failure of students in Mathematics at both internal and external examinations has become recurrent decimal. Before now, series of factors (inadequate reading materials, unsuitable instructional strategies, ill-equipped school environment, home factors, teachers' factors, and students' study habits) have been identified by earlier researchers as being responsible for students' poor academic performance in both internal and external examinations. It is observed in the literature by the researchers that much research studies have not been carried out on the quality of the test items used in assessing students' academic performance. In test construction, the underlying basic aim is to develop items that will be uniform to different across subgroups (male and female, rural and urban, etcetera). Therefore, study aimed at examining the quality of 2018/2019 Mathematics Joint Mock multiple-choice questions in Kwara State. The findings tell the degree of validity and reliability of the test.

The purpose of this study is to examine the DIF in multiple-choice Mathematics test questions constructed by the Kwara State Ministry of Education and Human Capital Development for the year 2017/2018 academic session. The specific research objectives are to assess:

1. Mathematics multiple-choice questions that are flagged DIF for male and female examinees.
2. Mathematics multiple-choice questions that are flagged DIF for rural and urban examinees.

The following research questions were generated and answered by the researchers:

1. What are the multiple-choice Mathematics test questions that are flagged DIF for male and female examinees?
2. What are the multiple-choice Mathematics test questions that are flagged DIF for urban and rural examinees?

Materials and Methods

The descriptive research design of the survey type was adopted in carrying out this study. The population included all public Senior Secondary Schools students in Kwara State, Nigeria, while the target population included all Senior Secondary School II students in Kwara State, Nigeria. At present, there are 25, 954 Senior Secondary School II students across Kwara State, Nigeria. The respondents of this study were drawn from senior secondary school II students. Mock examinations are being written at this level, in preparing for the final senior secondary examinations.

In Kwara State, Nigeria, schools were stratified into three senatorial districts (i.e., Kwara Central, Kwara North, and Kwara South). In each senatorial district, two Local Government Areas (LGAs) were sampled using a simple random sampling technique. A simple random sampling technique was used in each LGA sampled to select six public senior secondary schools. In any school sampled, 30 students were sampled using a simple random sampling technique. Hence, the expected sample size was 1,080, but some examinees could not complete the test items. Because of this scenario, 18 scripts were not used, representing a 1.7% mortality rate. Therefore, 1,062 scripts were used for computation and to answer the research questions.

The typical practice in the DIF studies is to designate a "reference" group as a group that is alleged to have an advantage over the "focal" group. Two categories of sub-groups (male & female, and rural & urban students) have been identified. In the first sub-group (male and female), male examinees are labeled as "reference" and the females as "focal" group. In the same vein, based on school location, examinees from urban schools are named "reference" while examinees from rural schools are called "focal" groups. In this study, the 2018/2019 academic session Joint Mock Mathematics Multiple-choice Questions (JMMMQ) that contained 50 items was used. Permission was obtained from the Kwara State Ministry of Education and Human Capital Development sequel to the consent of the sampled students. The researchers administered the measuring device to the examinees irrespective of their gender and school location. The examinees answered the questions dichotomously and scored as right or wrong.

The measuring tool is assumed to be a product of a standardized test; however, the measuring device's psychometric properties were not disclosed to the public. Three mathematics education teachers were involved in ascertaining that all items in the instrument addressed the defined research objectives. They also helped in finding answers to the questions. Scale content validity (S-CVI) was computed through item-level content validity (I-CVI). The item level content (I-CVI) was computed using frequency counts, the percentage for all items, and averaged to form scale content validity (S-CVI), as reported in table 1.

Table 1. Processes of Content Validation of 3 Experts and 30 Items.

Items	Expert 1	Expert 2	Expert 3	Experts' Agreement	I-CVI
1	√	√	√	3	3/3 = 1.00
2	—	√	√	2	2/3 = 0.67
3	√	√	√	3	3/3 = 1.00
4	√	—	√	2	2/3 = 0.67
5	—	√	√	2	2/3 = 0.67
6	—	√	√	2	2/3 = 0.67
7	√	√	√	3	3/3 = 1.00
8	√	√	—	2	2/3 = 0.67
9	√	√	√	3	3/3 = 1.00
10	√	√	√	3	3/3 = 1.00
---	---	---	---	---	---
---	---	---	---	---	---
---	---	---	---	---	---
50	√	√	√	3	3/3 = 1.00
			S-CVI = $\frac{45.50}{50}$ = 91.0		

Table 1 revealed the summary of the computation; hence, Scale Content Validity (S-CVI) obtained was 0.91. The reliability of the measuring instrument (JMMMQ) was not tested. This is because it is assumed that the instrument has undergone standardization. The instrument was conducted by the Kwara State Ministry of Education and Human Capital Development. The underlying basic aim of this study is to find out the level of standardization of the instrument.

Results and Discussion

There are 50 multiple-choice questions in the test administered. The examinees answered the questions dichotomously and scored as right or wrong.

Demographic Characteristics of the Examinees

The percentage was used to describe the examinees' categories based on gender and school location that participated in the study, as illustrated in table 1. Histogram was drawn via WinGen software to describe test-takers' abilities (Theta) that participated in the study, as shown in table 2 and figure 1.

Table 2. Demographic Characteristics of the Test-Takers.

Categories	Characteristics	Grouping	Frequency	Percentage (%)
	Male	Reference	633	59.60
Gender	Female	Focal	429	40.40
	Total		1062	100.00
	Urban	Reference	596	56.12
School Location	Rural	Focal	466	43.88
	Total			100.00

Table 2 shows the background profile of the test-takers. It reveals that 633 (59.60%) were males tagged as the reference group, while the rest 429 (40.40%) were females, labeled as the focal group. In the same vein, 596 (56.12%) were test-takers from urban schools labeled reference groups, while the rest 466 (43.88%) were from rural schools, called the focal group. In the same vein, test-takers' abilities (Theta) that participated in the study are shown in figure 6.

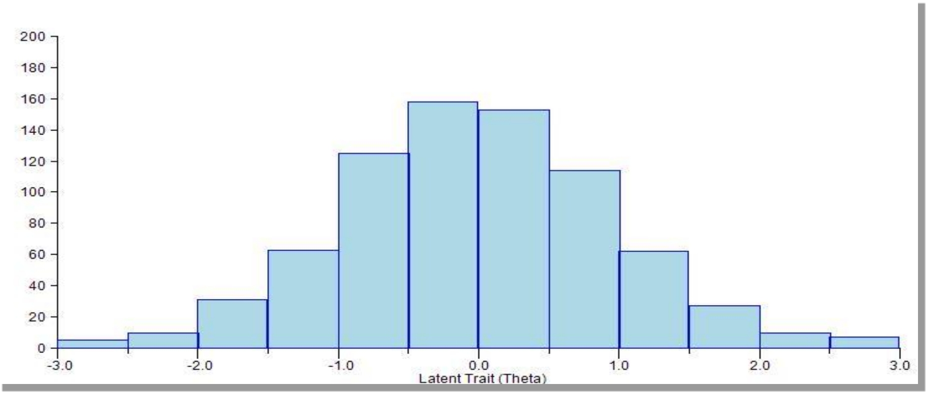

Figure 6. Test-takers' Abilities (Theta) in the Joint Mock Mathematics.

Figure 6 reveals that the examinees' ability (Theta) sampled concentrated and ranged between -2 to +2, as shown in the graph. It, therefore, implies that 86.67% of the students tested are of average ability.

Answering of Research Questions

Before the research question was answered, unidimensionality of test items were considered. The test revealed that 41 out of 50 representing 82.0% tends to be unidimensional. This implies that the mathematics multiple-choice items constructed in 2018/2019 academic session by the Kwara State Ministry of Education and Human Capital Development measured mathematics ability only. The two research questions raised in this study assessed DIF using Mantel-Haenszel (MH D-DIF).

Therefore, to assess DIF, the rule-of-thumb behind for dichotomously scored items is by classifying items into "A," "B," and "C." The classification rules state that items could be classified into "A" if the MH D-DIF value is less than or equal 1 (MH D-DIF $\leq$ 1) having negligible DIF, otherwise known as no DIF. In the category "B," if the MH D-DIF value is less than or equal 1.5 (MH D-DIF $\leq$ 1.5) is termed intermediate, and lastly, an item is classified as "C" if the MH D-DIF value is higher than 1.5 (MH D-DIF >1.5) (Michaelides, 2008).

The odds ratio (αMH) was used to compute the direction of DIF. Items detected having DIF (category "B" or "C") could favor either the reference (male students) or focal (female students) group. Using Mantel-Heanszel Common odd ratio (αMH) greater than one indicates that an item favors the reference group. In contrast, an odds ratio (αMH) less than one favors the focal group (Michaelides, 2008). Those items that are flagged DIF irrespective groups are considered based on gender using odds ratio (αMH) less than one or greater than one to answer research question one.

Research Question 1: What are the multiple-choice Mathematics test items that are flagged DIF for male and female examinees?

Table 3. Multiple-choice Mathematics Test Items that are Flagged DIF for Male and Female Examinees.

Items	αMH	MH D-DIF	DIF Classification	Favouring Group
1	0.85	1.14	B	Focal (Female)
2	0.16	0.51	A	No DIF
3	0.08	1.12	B	Focal (Female)
4	0.81	0.07	A	No DIF
5	0.97	0.25	A	No DIF
6	0.38	0.64	A	No DIF
7	0.74	-0.01	A	No DIF
8	0.68	-0.53	A	No DIF
9	1.38	0.64	A	No DIF
10	1.26	0.49	A	No DIF
11	0.80	0.04	A	No DIF
12	1.03	0.29	A	No DIF
13	0.59	-0.27	A	No DIF
14	1.49	0.69	A	No DIF
15	1.17	0.43	A	No DIF
16	0.67	0.87	A	No DIF
17	0.75	2.33	C	Focal (Female)
18	0.86	1.11	B	Focal (Female)

Continued: Table 3. Multiple-choice Mathematics Test Items that are Flagged DIF for Male and Female Examinees.

Items	αMH	MH D-DIF	DIF Classification	Favouring Group
19	2.53	1.98	C	Reference (Male)
20	1.01	0.34	A	No DIF
21	1.98	1.39	B	Reference (Male)
22	0.45	0.59	A	No DIF
23	2.35	1.74	C	Reference (Male)
24	0.98	1.26	B	Reference (Male)
25	0.59	0.75	A	No DIF
26	1.91	1.23	B	Reference (Male)
27	0.68	0.90	A	No DIF
28	1.95	1.22	B	Reference (Male)
29	0.95	0.20	A	No DIF
30	2.07	2.80	C	Reference (Male)
31	0.88	1.13	B	Reference (Male)
32	0.64	0.85	A	No DIF
33	1.90	1.15	B	Reference (Male)
34	0.97	0.26	A	No DIF
35	1.78	1.04	B	Reference (Male)
36	1.33	0.54	A	No DIF
37	0.40	0.61	A	No DIF
38	1.40	0.61	A	No DIF
39	1.20	0.46	A	No DIF
40	1.10	0.34	A	No DIF
41	0.76	1.03	B	Focal (Female)
42	2.83	0.32	A	No DIF
43	1.15	0.39	A	No DIF
44	1.03	0.35	A	No DIF
45	0.16	0.20	A	No DIF
46	0.86	0.11	A	No DIF
47	1.37	0.58	A	No DIF
48	0.93	0.22	A	No DIF
49	1.40	0.61	A	No DIF
50	1.85	1.10	B	Reference (Male)

NB: αMH = Mantel-Heanszel Common odd ratio
MH D-DIF = Mantel-Heanszel Delta DIF

Result from Table 3 shows that out of 50 multiple-choice questions in 2017 Joint Mock Mathematics test, 34 of the items (2, 4, 5, 6, 7, 8, 8, 10, 11, 12, 13, 14, 15, 16, 20, 22, 25, 27, 29, 32, 34, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48 and 49) representing 68.0% flagged MH D-DIF in category "A" which is termed as negligible DIF otherwise known as no DIF. Twelve (12) items (1, 3, 18, 21, 24, 26, 28, 31, 33, 35, 41 and 50) representing 24.0% were flagged MH D-DIF in category "B" which inferred intermediate DIF; such items needs adjustment or managed to be used and retained in the test format. Lastly, only four items (17, 19, 23, and 30) representing 8.0% were flagged MH D-DIF in category "C"; such items must be removed from the set of questions except if they are there a purpose. It implies that twelve (16) items (1, 3, 17, 18, 19, 21, 23, 24, 26, 28, 30, 31, 33, 35, 41 and 50) representing 32.0% were flagged DIF.

Table 3 also indicates that 16 items (1, 3, 17, 18, 19, 21, 23, 24, 26, 28, 30, 31, 33, 35, 41 and 50) has evidence of DIF, out of which 11 items (19, 21, 23, 24, 26, 28, 30, 31, 33, 35 and 50) were misbehaved items to focal group (female examinees) while only 5 items (1, 3, 17, 18 and 41) were unstable items to reference group (male examinees). Those items that are flagged DIF irrespective groups are considered based on school location using odds ratio (αMH) less than one or greater than one to answer research question one.

Research Question 2: What are the multiple-choice Mathematics test questions that are flagged DIF for urban and rural examinees?

Table 4. Multiple-Choice Mathematics Test Questions that are Flagged DIF for Urban and Rural Examinees.

Items	αMH	MH D-DIF	DIF Classification	Favouring Group
1	0.13	1.03	B	Focal (Rural)
2	1.88	0.87	A	No DIF
3	1.64	0.81	A	No DIF
4	1.28	0.53	A	No DIF
5	1.70	0.82	A	No DIF
6	0.93	0.18	A	No DIF
7	0.77	0.99	A	No DIF
8	0.57	0.71	A	No DIF
9	0.65	0.90	A	No DIF
10	1.10	1.41	B	Reference (Urban)
11	1.04	1.35	B	Reference (Urban)
12	0.72	0.93	A	No DIF
13	0.83	1.07	A	No DIF
14	0.66	0.88	A	No DIF
15	0.93	0.19	A	No DIF
16	0.69	0.88	A	No DIF
17	2.25	2.94	C	Reference (Urban)
18	0.72	0.93	A	No DIF
19	0.75	0.96	A	No DIF
20	0.64	0.87	A	No DIF
21	0.85	1.21	B	Focal (Rural)
22	0.60	0.76	A	No DIF
23	1.04	0.29	B	Reference (Urban)
24	0.94	0.20	A	No DIF
25	0.97	0.24	A	No DIF
26	1.08	0.37	A	No DIF
27	0.97	1.44	B	Focal (Rural)
28	0.90	0.14	A	No DIF
29	1.08	0.39	A	No DIF
30	0.86	1.14	B	Focal (Rural)
31	0.91	1.13	B	Focal (Rural)
32	0.17	1.17	B	Focal (Rural)
33	0.13	0.44	B	Focal (Rural)
34	0.79	1.34	B	Focal (Rural)
35	1.09	1.05	B	Reference (Urban)

Continued: Table 4. Multiple-Choice Mathematics Test Questions that are Flagged DIF for Urban and Rural Examinees.

Items	αMH	MH D-DIF	DIF Classification	Favouring Group
36	0.80	0.40	A	No DIF
37	1.57	1.00	B	Reference (Urban)
38	0.69	2.00	C	Focal (Rural)
39	0.72	0.89	A	No DIF
40	0.73	0.93	A	No DIF
41	0.12	0.96	A	No DIF
42	2.16	0.19	A	No DIF
43	0.64	2.78	C	Focal (Rural)
44	0.09	0.82	A	No DIF
45	0.55	0.11	A	No DIF
46	0.89	0.75	A	No DIF
47	0.74	1.15	B	Focal (Rural)
48	1.59	1.03	B	Reference (Urban)
49	1.07	1.40	B	Reference (Urban)
50	0.92	1.19	B	Focal (Rural)

NB: αMH = Mantel-Heanszel Common odd ratio
MH D-DIF = Mantel-Heanszel Delta DIF

Result from Table 4 reveals that out of 50 multiple-choice items in 2017 Joint Mock Mathematics test, 30 of the items (2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16, 18, 19, 20, 22, 24, 25, 26, 28, 29, 36, 39, 40, 41, 42, 44, 45, and 46) representing 62.0% flagged MH D-DIF "A" item which is termed as negligible DIF otherwise known as no DIF. Seventeen (17) items (1, 10, 11, 21, 23, 27, 30, 31, 32, 33, 34, 35, 37, 47, 48, 49 and 50) representing 32.0% were flagged MH D-DIF "B" which inferred intermediate DIF; such items needs adjustment or managed to be used and retained in the test format. Lastly, only three items (17, 38, and 43) representing 06.0% were flagged MH D-DIF "C" items; such items must be removed in the test format except if they are there for a purpose. This implies that in all, nineteen (20) items (1, 10, 11, 17, 21, 23, 27, 30, 31, 32, 33, 34, 35, 37, 38, 43, 47, 48, 49 and 50) representing 32.0% were flagged DIF.

Table 4 also shows that 20 items (1, 10, 11, 17, 21, 23, 27, 30, 31, 32, 33, 34, 35, 37, 38, 43, 47, 48, 49 and 50) has evidence of DIF, out of which 8 items (10, 11, 17, 23, 35, 37, 48 and 49) misbehaved to reference group (examinees from urban schools) while 12 items (1, 21, 27, 30, 31, 32, 33, 34, 38, 43, 47 and 50) unstable for focal group (examinees from rural schools).

One of the outcomes of this study revealed that 16 items were flagged DIF, 12 of the questions were misinterpreted by the focal (female) group, and four to the reference group (male). This outcome suggests that mathematics is more inclined to males than females. The findings of this study are in congruence with the conclusion reached by Birjandi & Mohadeseh (2007) that in the general reading comprehension, 7 out of the 13 items flagged DIF favored females, and six proved much easier for males. The outcome is also in agreement with Adedoyin (2010), who discovered that out of 16 items flagged DIF, five items were gender-biased. Madu (2012) observed that male and female examinees functioned differently in 39 items out of 50. This study also agrees with Adebule's (2013) result that out of the 40 items investigated, seven items displayed DIF comparing male and female examinees. The present finding disagrees with the finding of Driana (2007), who studied the presence of DIF on the mathematics section of the Ohio Ninth-Grade proficiency test. The result of the study indicated no presence of regional DIF. This finding is not in agreement with Ajeigbe and Afolabi's (2014) outcome, which investigated DIF items of Osun State qualifying Mathematics and English Language test. It was discovered that the test items were unidimensional, assessed one trait or attribute of interest with no DIF items found.

The secondary outcome of this study revealed that out of 20 items that were flagged as DIF, eight items were misinterpreted by reference group (students from the urban schools), and 12 to focal group (students from rural the schools). The interview carried out on the sampled students revealed that urban examinees had been taught most of the topics via extra-mural (evening classes) that allowed them to perform better than the rural examinees. This sole reason allowed the urban examinees to have fewer items flagged as DIF. This outcome agreed with Ogbebor and Onuka (2013) when they concluded that out of 60 items in NECO Economics questions, eight questions were flagged DIF with regards to school location. This finding agrees with Amuche and Fan's (2014) conclusion that eight items were biased regarding school location. This study's outcome was also in support of Mokobi and Adedoyin (2014) when they discovered that 24 items fit the IRT (3PLM) model while 6 of the items were rural /urban location biased.

Conclusion

It was noted that any test constructed by the examiners (internal or external) must be free from any kind of irrelevant clues that can allow examinees to give different interpretations to items in a test. Hence, it makes examinees perform differently in the test. In the 2018/2019 academic session, the Kwara State mathematics multiple-choice joint mock test items were flagged DIF concerning gender and school location. It could be concluded, therefore, that the quality of assessment instrument might be another variable or factor that could hinder students' performance in Mathematics both at the internal and external examinations levels. The misbehavior of any instrument favoring one group over the other group implies that the validity and dependability of that instrument are being threatened.

Recommendations

Stemming from the study findings are the following recommendations:

1. In the construction of any instrument (test), the developer must ensure that instrument (test) has no irrelevant clue that may make an individual or group respond to the instrument differently.

2. Test construction Department of Kwara State Ministry of Education and Human Capital Development should be ready to train and retrain test developers so that the test items being used could be valid and be relied upon to assess examinees' ability.
3. Whatever test items to be used by the Ministry of Education and other allied examination bodies, it is advisable to subject the items to DIF to identify items that could function differently and flagged as DIF before being used for intended purposes.
4. Test items should be written in a straightforward manner, avoiding unnecessary clue that could disallow examinees' straightforward interpretation of the test items.
5. There is a need to introduce Item Response Theory in master's measurement and evaluation course curriculum in item parameters (difficulty, discrimination, and guessing) computation, but retaining item parameters (difficulty, discrimination, and guessing) being computed using Classical Test Theory (CTT) for the undergraduate program.

Literature Cited

Abedalaziz, N. (2010). Detecting gender-related using logistic regression and Mantel-Haenszel approaches. *Procedia Social and Behavioral Sciences, 7* (C), 406-413. Downloaded on 17/01/2020 from www.scencedirect.com.

Adebule, S. O. (2013). A study of differential item function in Ekiti State unified mathematics examination for senior secondary schools. *Journal of Education and practice, 4(*17), 43-46. www.iiste.org.

Adedoyin, O. O. (2010). Investigating the invariance of person parameter estimates based on classical test and item response theories. *International Journal of Education Science, 2* (2), 107-113.

Adewole, P. & Ojo, B. O. (2017). Application of item characteristic curve (ICC) in the selection of Test items. *British Journal of Education*, 5, (2), 21-41. www.eajournals.org

Ahmadi, A. & Bazvand, A. D. (2016). Gender differential item functioning on a national field-specific test: The case of the Ph.D. entrance examination of TEFL in Iran. *Iranian journal of languages teaching research*, 4 (1), 63-82.

Ajeigbe, T. O. & Afolabi, E. R. I. (2014). Assessing unidimensionality and differential item functioning in qualifying for senior secondary school students, Osun State, Nigeria. *World Journal of Education 4* (4), 30-37. Downloaded on 17/01/2020. http://dx.doi.org/10.5430/wje.vn4p30.

Amuche, C. I. & Fan, A. F. (2014). An assessment of item bias using differential item functioning technique in NECO Biology conducted examinations in Taraba State Nigeria. *American International Journal of Research in Humanities, Arts and Social Sciences*, 95-100. Downloaded on 22/01/2020 and available online at http://www.iasir.net

Baker, F. B. (2001). *The basics of item response theory* (2nd Edition). United States of America: ERIC Clearinghouse on Assessment and Evaluation.

Birjandi, P. & Mohadeseh, A. (2007). Differential item functioning (test bias analysis paradigm across manifest and latent examinee groups (on the construct validity of IELTS. *Human Sciences, 55*, 153-172.

Dorans, N. J. & Holland, P. W. (n.d.) DIF detection and description: Mantel-Haenzel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning,* 35-66. Hillsdale, NJ: Lawrence Eribaum Associates, Inc.

Driana, E. (2007). *Gender differential item functioning on a ninth-grade mathematics proficiency test in Appalachian, Ohio.* A dissertation presented to the faculty of the College of Education of Ohio University in partial fulfillment of the requirement for the degree Doctor of Philosophy.

Gao, X. (2019). A Comparison of Six DIF Detection Methods. A thesis submitted in partial fulfillment of the requirements for the degree of Master of Arts at the University of Connecticut.

Ibrahim, A. (2017). An empirical comparison of three methods for detecting differential item functioning in dichotomous test items. *Journal of Teaching and Teacher Education,* 5, (1),

Kanjee, A. (2007). Using logistic regression to detect bias when multiple grows are tested. *South African Journal of Psychology, 37,* 47-61.

Madu, B. (2012). Analysis of gender-related differential item functioning in multiple-choice mathematics items administered by the West African Examinations Council. *Journal of Education and Practice, 3* (8), 71-76.

Mokobi, T. & Adedoyin, O.O. (2014). Identifying location biased items in the 2010 Botswana Junior Certificate Examination Mathematics paper one using the item response characteristics curves. *International Review of Social Sciences and Humanities*, (2), 63-82. Downloaded on 22/01/2020 and available online at www.irssh.com

Michaelides, M. P. (2008). Practical assessment research and evaluation: An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. A peer electronic journal, 13 (7), -13, Downloaded on 11/01/2020 and available online at www.pareonline.net/gettvn.asp?v=13&n=7

Ogbebor, U & Onuka, A. (2013). A differential item functioning method as an item bias indicator. *International Research Journals*, 4 (4), 367-373. Downloaded on 22/01/2018 and available online at http://www.interesjournals.org/ER.

Karami, H. (2012). An Introduction to Differential Item Functioning. *The International Journal of Educational and Psychological Assessment*, 11(2) 012, 59–79.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.