International Journal of Scientific Research in Education, March 2017, Vol. 10(1), 1-11.

International Journal of Scientific Research in Education (IJSRE)
ISSN: 1117-3259

Bridging the Gap in the Current Global Initiative in Validation Process in Psychometrics: Nigerian Perspective

Cyrinus B. Essen[i]
University of Calabar, Calabar, Nigeria
ecyrinus@ymail.com

Id-Basil F. Ukofia[ii]
Federal College of Education (Technical), Omoku, Rivers State.
ibfukofia86@gmail.com

Bassey A. Bassey[iii]
University of Calabar, Calabar, Nigeria
babassey67@gmail.com

Delight O. Idika[iv]
University of Calabar, Calabar, Nigeria
delightoidika@yahoo.com

Abstract

The study investigated types of differential item functioning (DIF) and effect sizes in 2014 State conducted examination in Multiple-choice Mathematics items. Ex-post facto design was adopted. The population of the study consisted of 47, 599 senior secondary two (SS2) students' responses in the three educational zones of Akwa Ibom State, Nigeria. The study sample comprised, 3,066 examinees' responses; 1,533 male and 1,533 female respectively were proportionately selected through stratified sampling procedure. A three-step model logistic regression procedure using IBM SPSS statistics version 20 was used to determine the different types of DIF. The results revealed that all the 50 items displayed uniform and nonuniform differential item functioning (DIF). Also, these items displayed negligible effect size at the uniform and/or nonuniform DIF. It was recommended that DIF analysis should incorporate types of DIF and effect sizes for making valuable psychometrics decision.

Keywords: Bridging, Gap, Validation, Process, Psychometrics, Nigeria.

Reference to this paper should be made as follows:

Essen, C. B., Ukofia, I. F., Bassey, B. A., & Idika, D. O. (2017). Bridging the Gap in the Current Global Initiative in Validation Process in Psychometrics: Nigerian Perspective. *International Journal of Scientific Research in Education, 10*(1), 1-11. Retrieved [DATE] from http://www.ijsre.com

INTRODUCTION

The use of test results for major educational decision without DIF evaluation in its complete dimensions is against the global best initiatives in measurement and assessment. Evaluating DIF is one of the expected modern psychometric analyses of bias measures to ensure equality of opportunity to all examinees. The global concern of measurement evaluators and educators to ensure that items in test and examinations are not only administered to obtain the scores but fairness and equality of opportunity to all examinees irrespective of differences in race, gender and socioeconomic backgrounds is receiving great attention in education and other major .areas that use test information for major decision making.

The progressive drift from traditional conceptualization of validity to the current view of validity is a welcome development in Psychometrics (McNamara & Roever, 2006). Evaluating DIF is one of the expected modern psychometric analyses of bias measures to ensure equality of opportunity to all examinees. Nigerian test developers and users in educational measurement and assessment are yet to key into this global initiative for decades.

The Educational Testing Services (ETS) in providing standard for psychometric bias analysis has evolved into an appreciable improvement in educational measurement. Test developers and evaluators are currently laying emphasis on the investigation of fairness, equality of opportunity and appropriateness of examination items with respect to different groups of examinees with similar abilities conditioned on their responses to construct of interest being measured and the identification of bias items ((Noortage & Boeck, 2005; Roever, 2005).

Differential Item Functioning as a collection of statistical procedures for detecting differences in group performances as a result of bias items among examinees with similar abilities is a global focus in educational assessment. Some of these procedures are Classical Test Theory (CTT) based, while others are Item Response Theory (IRT) based. Various detecting methods are widely used within these assessment frameworks of Classical test theory, Item response theory and Rasch Model in DIF analysis. The most popular methods within classical test theory include: Mantel Haenszel (MH) and Logistic regression (LR) and others. Within the Item response theory are: Lord's Chi-square statistics, Wald statistics, among others. Various computer programmes have been developed to make it feasible and possible. Amongst the software programmes are: BILOG. MG. 30; IRTPRO MG.30; MULTILOG; WINSTEP; RUMM and others (Perrone, 2006; Scherman & Goldstein, 2008).

McNamara and Roever (2006) opined that the contemporary discussion of validity is focused on test fairness through developing procedures that supports rationality of decision based on items not on demographic and social variables of the examinees. In addition, Zumbo (1999, as cited in Salahi, & Tayebi, 2011) and Brown (2005) stressed that the shift from the traditional view which simply relied on computing correlation with another measure, to the current view of validity as a measure to ensure, fairness and equality in testing situation is appropriate. Furthermore, Brown (2005) argued that test consideration should look away from the scores but on the interpretations for some specific purposes and inferences drawn from the test results, the decisions and actions thereafter.

Differential Item Functioning (DIF) is said to occur when a group of examinees with similar abilities, taught and examined on the same construct of interest exhibit differing probabilities of responding to items in the test (Camilli, 2006; Osterlind & Everson, 2009). Steinberg and Thissen (2006) believed that DIF examines the probability of correctly responding or endorsing an item(s) conditioned on the examinee's ability. In other word, when a particular group of examinees with respect to social, gender, race or demographic variables respond to a particular item(s) correctly than the other group, the particular item(s) is said to exhibit bias or differential functioning between the groups of examinees.

However, Differential Item Functioning (DIF) in bias analysis is expressed in two forms: Uniform and Non-uniform (DIF). Uniform DIF is believed to occur when a sub-group of examinees with ability levels, uniformly answer a particular item or subset of items than the other group. Therefore, that particular sub-group is said to be advantaged over the other group and can be considered as having a superior ability over the less favoured group. The advantaged group is termed as the "reference" group, while the less advantaged is the "focal" or the group of focus in bias analysis comparatively (Walker, 2011; Huang & Han, 2012).

Walker (2011) asserted that in uniform DIF, the item favours the advantaged group, while the other group is less favoured with respect to difficulty of the item(s) at different ability levels of the examinees. However, within the item response theory (IRT) framework, uniform DIF, occurs when item characteristic curves (ICCs) for the groups equally discriminate but exhibit differences in the difficulty parameter. Walker's view is supported as Huang and Han (2012) opined that the difficulty level in uniform DIF is different between the groups but the discrimination is the same. To further offer clarification on the uniform DIF, Le (2006) stressed that uniform DIF occurs when there is no interaction between ability level and group membership of the examinees.

In contrast to uniform DIF, Salehi and Teyabi (2012) explained that non-uniform DIF occurs when there is an interaction between test takers' ability level and their performance on an item contributing to change in the direction of DIF along the ability scale. Camilli and Shepard (1994, cited in Le, 2006) expressed that in non-uniform DIF, interaction is found between trait level, group assignment and item responses. In other words, the difference in the probability of responding correctly to item (s) between the groups is not the same at all levels of ability.

De Beer (2004) and Walker (2011) believed that in IRT, there is ICC intersection between the two groups at a point, indicating that a given item exhibits difficulty as well as discrimination. Huang and Han (2012) affirmed that in non-uniform DIF, the discrimination parameter is different; the difficulty may or may not be the same. In summary, the uniform DIF is about the difficulty of an item at a particular ability level or theta between the groups, while the non-uniform DIF is about the discrimination index of an item at a particular ability level between the groups of examinees. However, a given item can exhibit both uniform and non-uniform DIF or uniform or vice-versa.

Various effect size classification measures are used to examine types of DIF and effect sizes. Zumbo and Thomas, (1999 cited in Rezaee & Shabani, 2010); Jodoin & Gierl (2001) among others had proposed these measures in DIF assessment for informed decisions. However, this study used binary logistic regression as one of the mostly preferred CTT procedures for analysis of uniform and non-uniform DIF, and the effect sizes. The effect size for LR DIF is R-squared coefficient used to examine the partial correlation between the dependent (item responses) and each of the independent variable (group and total score). R^2 differences in the model fit of Chi-squared represent the effect levels of DIF. According to the effect size guideline, DIF values can be classified as: negligible (A- level = $R^2 < .035$), moderate (B- leve, = $R^2 \leq R^2.070$) and large (C – level, = $R^2 > .070$) (Alvi, Rezae &Amirian, 2011). The study used proposed classification guidelines of effect size measure by Jodoin and Gierl (2001) as predictor of Nagelkerke's R^2 to quantify uniform and non-uniform DIF.

Statement of the problem

The use of test results for major educational decision without DIF evaluation in its complete dimensions is against the global best initiatives in measurement and assessment. Evaluating DIF is one of the expected modern psychometric analyses of bias measures to ensure equality of opportunity to all examinees. Nigerian test developers and users in educational

measurement and assessment are yet to key into this global initiative for decades. This development has plagued the educational system with respect to quality and equitable assessment and measurement. Test is one of the preferred instruments to appraise examinee's level of proficiency, ability and skill in performing a given task with respect to chosen field of study. Hence, the need to dully evaluate DIF in its dimension is in compliance with Joint Committee on Standards for Educational and Psychological Testing of the AERA, APA, and NCME (2014). Few studies revealed that State and National examination bodies are yet to show commitment to this initiative as some items in various subjects administered displayed DIF. Though, the issues of uniform and non-uniform DIF and effect sizes have not been addressed. The imperativeness of these issues should not be overlooked, if information for practical significance of DIF to interpret results is to be achieved. Based on this gap, this study decided to investigate uniform and non-uniform DIF and effect size using State administered 2014 mock multiple choice Mathematics items in Nigeria, conducted by the Akwa Ibom State Ministry of Education for senior secondary three (SS 3) students. As a major examination, the importance of making major decisions based on the results obtained has educational, professional or employment implications. Thus, the need to key into the current global initiative in validation process in psychometric is imperative for Nigerian education assessors and evaluators.

Purpose of the study

The study was carried out to investigate types of differential item functioning (DIF) and the effect size in 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria. Specifically, the study investigated the extent 2014 Mathematics items displayed:

1. Uniform DIF,
2. Non-uniform DIF, and
3. DIF effect sizes.

Research Questions

The following research questions were formulated to guide the study:

- To what extent do 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria, display uniform DIF?
- To what extent do 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria, display non- uniform DIF?
- To what extent do 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria, display different DIF effect sizes?

METHODOLOGY

The data used for the study consisted of 2014 mock multiple-choice Mathematics students' responses in Akwa Ibom State, Nigeria. Ex- post facto design was adopted. The population of the study consisted of 47, 599 Senior Secondary Two (SS2) students' responses in 2014 state mock multiple choice Mathematics in the public senior secondary schools in three educational zones of Uyo, Eket and Ikot Ekpene. Twenty three thousand, eight hundred and thirty one were male students' responses, while 23,768 were female students' responses. The sample for the study comprised, 3,066 examinees' responses from the three educational zones representing 15.52 percent, selected through stratified sampling procedure. One thousand, five hundred and thirty three (15.54 %) were male candidates' responses, while 1,533

(15.54%) were female candidates' responses. A large sample size was considered to obtain a stable estimate as a requirement in DIF analysis procedure (Broer, Lee, Rizavi & Powers, 2005). The 2014 State mock multiple –choice Mathematics is a 50 item four-optioned: A – D test. The item responses were scored in a binary format of "1" correct and "0" incorrect. The data collection was gathered by the researcher through the permission by the Director of the State Ministry of Education, Examination and Certification Unit, Akwa Ibom State, Nigeria. The reliability of the test instrument was .71 Cronbach's Alpha based on standardized items. The examinees' responses were subjected to a three-step model binary logistic regression statistical analysis proposed by Zumbo (1999, cited in Rezaee & Shabani, 2010), using IBM SPSS statistics version 20.

Step 1 (model 1): Entered the conditioning variable (total test score)
Step 11 (model 2): Entered the group variable (male and female)
Step 111 (model 3): Entered the interaction term between the conditioning variable and group variable (total test score multiply by group variable).

Chi-square test was used in addition to estimate Nagelkerke R^2 effect size as a means of examining practical significance of DIF. The logistic regression equation was analyzed thus:

1. Model 2 minus model 1 = Uniform DIF only
2. Model 3 minus model2 =Non-uniform DIF.

A 1-degree of freedom Chi-square difference between model 1 and model 2 with a p-value less than .05 significant level, indicated uniform DIF. Similarly, 1-degree of freedom Chi-square difference between model 2 and model 3 with a p-value less than .05 significant level, indicated non-uniform DIF. An item was said to be DIF significant when the group differences when the Chi-squared test of significant for a particular item was less than or equal to .05. Jodoin and Gierl (2001) classifications of "negligible", "moderate" and "large" effects respectively were used with Nagelkereke R^2 difference between model 1 and model 2, as well model 2 and model 3.The dependent variable (item responses), while the independent variables(total score(ability) and gender) were used. The reference group (male) was coded 1, while the focal group (female) was coded 0.

RESULTS

The results of the data analysis are presented in Tables, 1, 2, and 3, according to the research questions.

Research Question 1: To what extent do 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria, display uniform DIF?

Table 1: Uniform Differential Item Functioning (DIF) of 2014 Mock Multiple Choice Mathematics Items of Akwa Ibom State, Nigeria.

Item	Model 2 minus Model 1 Difference			Item	Model 2 minus Model 1 Difference			Item	Model 2 minus Model 1 Difference		
	X^2	df	Sig.		X^2	df	Sig.		X^2	df	Sig.
1	.651	1	.000	21	3.925	1	.000	41	3.631	1	.000
2	2.207	1	.000	22	7.542	1	.000	42	1.276	1	.000
3	.792	1	.000	23	3.113	1	.000	43	6.446	1	.000
4	2.808	1	.000	24	2.088	1	.000	44	.006	1	.000
5	-1.62	1	.000	25	4.995	1	.000	45	2.358	1	.000
6	- .067	1	.000	26	.265	1	.000	46	8.318	1	.000

7	.832	1	.000	27	2.798	1	.000	47	4.352	1	.000
8	.04	1	.000	28	1 .13	1	.000	48	7.381	1	.000
9	.134	1	.000	29	.468	1	.000	49	.993	1	.000
10	6.048	1	.000	30	.134	1	.000	50	2.232	1	.000
11	1.936	1	.000	31	1.393	1	.000				
12	.207	1	.000	32	.43	1	.000				
13	4.115	1	.000	33	.818	1	.000				
14	6.033	1	.000	34	1.287	1	.000				
15	7.864	1	.000	35	17.435	1	.000				
16	.061	1	.000	36	1.034	1	.000				
17	.065	1	.000	37	5.183	1	.000				
18	.726	1	.000	38	3.023	1	.000				
19	3.676	1	.000	39	.842	1	.000				
20	3.363	1	.000	40	.532	1	.000				

***** All items display uniform DIF with 1 degree of freedom at .000 significant level

The result in Table 1 indicates that all the items exhibit uniform DIF with 1 degree Chi-square difference with significant value of .000 less than .05. Uniform DIF was assessed by subtracting model 2 (group) Chi-square from model 1 (total score) Chi-square. This indicates that the Mathematics items display uniform differential functioning

Research question 2: To what extent do 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria, display non- uniform DIF?

Table 2: Non-uniform Differential Item Functioning (DIF) of 2014 Mock Multiple- Choice Mathematics Items of Akwa Ibom State, Nigeria

Item	**Model3 minus Model 2 Difference**			**Item**	**Model 3 minus Model 2 Difference**			**Item**	**Model 3 minus Model 2 Difference**		
	X^2	**df**	**Sig.**		**X^2**	**df**	**Sig.**		**X^2**	**df**	**Sig.**
1	2.527	1	.000	21	1.358	1	.000	41	.044	1	.000
2	4.7	1	.000	22	.163	1	.000	42	.021	1	.000
3	.081	1	.000	23	.943	1	.000	43	.18	1	.000
4	2.108	1	.000	24	5.211	1	.000	44	.314	1	.000
5	.143	1	.000	25	.128	1	.000	45	1.527	1	.000
6	.067	1	.000	26	.327	1	.000	46	.606	1	.000
7	.004	1	.000	27	.379	1	.000	47	.472	1	.000
8	.004	1	.000	28	.484	1	.000	48	5.334	1	.000
9	2.676	1	.000	29	6.671	1	.000	49	2.224	1	.000
10	1.612	1	.000	30	.848	1	.000	50	12.167	1	.000
11	1.329	1	.000	31	.724	1	.000				
12	2.263	1	.000	32	.295	1	.000				
13	.939	1	.000	33	.827	1	.000				
14	.149	1	.000	34	.779	1	.000				
15	.834	1	.000	35	2.235	1	.000				
16	.523	1	.000	36	.874	1	.000				
17	.556	1	.000	37	.082	1	.000				
18	. 081	1	.000	38	.082	1	.000				
19	.833	1	.000	39	.001	1	.000				
20	1 .262	1	.000	40	.278	1	.000				

Note: All items display Non-uniform DIF with1 degree of freedom at .000 significant level.

The result reveals that the entire all the 50 items display non-uniform differential item functioning (DIF) with 1 degrees of freedom at .000 significant less than .05. Therefore, the 2014 mock multiple- choice Mathematics items of Akwa Ibom State, Nigeria, display non-uniform DI. This result indicates that non- differential item functioning occurs as a result of the interaction between the group membership and total score.

Research question 3: To what extent do 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria, display different DIF effect sizes?

Table 3: Differential Item Functioning (DIF) Effect Size Analysis

Uniform DIF [Model 2 (R^2) – Model 1(R^2)]						Non-uniform DIF [Model 3 (R^2) – Model 2(R^2)]					
Item	Sig.	Remark	Item	Sig.	Remark	Item	Sig.	Remarks	Item	Sig.	Remark
1	.000	Negligible	26	.000	Negligible	1	.002	Negligible	26	.000	Negligible
2	.002	Negligible	27	.002	Negligible	2	.002	Negligible	27	.000	Negligible
3	.000	Negligible	28	.001	Negligible	3	.000	Negligible	28	.000	Negligible
4	.003	Negligible	29	.000	Negligible	4	.001	Negligible	29	.003	Negligible
5	-.002	Negligible	30	.000	Negligible	5	.000	Negligible	30	.000	Negligible
6	.000	Negligible	31	.000	Negligible	6	.000	Negligible	31	.000	Negligible
7	.04	Negligible	32	.000	Negligible	7	.000	Negligible	32	.000	Negligible
8	.000	Negligible	33	.001	Negligible	8	.000	Negligible	33	.000	Negligible
9	.002	Negligible	34	.001	Negligible	9	.002	Negligible	34	.000	Negligible
10	.003	Negligible	35	.008	Negligible	10	.000	Negligible	35	.001	Negligible
11	.001	Negligible	36	.001	Negligible	11	.000	Negligible	36	.000	Negligible
12	.000	Negligible	37	.002	Negligible	12	.001	Negligible	37	.000	Negligible
13	.002	Negligible	38	.001	Negligible	13	.000	Negligible	38	.000	Negligible
14	.003	Negligible	39	.000	Negligible	14	.000	Negligible	39	.000	Negligible
15	.004	Negligible	40	.000	Negligible	15	.000	Negligible	40	.000	Negligible
16	.000	Negligible	41	.002	Negligible	16	.000	Negligible	41	.000	Negligible
17	.000	Negligible	42	.001	Negligible	17	.001	Negligible	42	.000	Negligible
18	.001	Negligible	43	.003	Negligible	18	.000	Negligible	43	.000	Negligible
19	.001	Negligible	44	.000	Negligible	19	.001	Negligible	44	.000	Negligible
20	.002	Negligible	45	.001	Negligible	20	.003	Negligible	45	.001	Negligible
21	.001	Negligible	46	.004	Negligible	21	.001	Negligible	46	.000	Negligible
22	.003	Negligible	47	.002	Negligible	22	.001	Negligible	47	.000	Negligible
23	.001	Negligible	48	.004	Negligible	23	.001	Negligible	48	.002	Negligible
24	.000	Negligible	49	.000	Negligible	24	.003	Negligible	49	.001	Negligible
25	.000	Negligible	50	.001	Negligible	25	.002	Negligible	50	.007	Negligible

A (negligible) DIF: $R^2 < .035$; B (moderate) DIF: $R^2 \leq R^2 \leq .070$; C (large) DIF: $R^2 > .070$; Items display negligible effect size DIF at Uniform and Non- uniform DIF

The result reveals that by 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria, display uniform and non-uniform DIF with negligible effect size at Nagelkereke $R^2 < .035$. The negligible DIF effect size is considered small. The result indicates that the items did not display the other two types (moderate: $R^2 \leq R^2 \leq .070$ and large: $R^2 > .070$) of the effect size classifications.

DISCUSSION OF FINDINGS

The result of research question one revealed that all the 50 items of 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria, displayed uniform DIF. The finding of this study is in consonant with various studies that identified uniform DIF in items of different examinations using logistic regression procedure. For instance, Abedalaziz (2010) found in a study using logistic regression that 10 of the 30 items of the tenth grade students' Mathematics in Jordan at the end of the First semester, school year 2009 – 2010, displayed uniform DIF that favoured the male group. Similarly, Rezaee and Shabani (2010) discovered uniform DIF in 11 out of the 39 items in a partial requirement admission examination for Ph.D. programme of the University of Tehran English Proficiency test. Six of the items favoured males, while five items favoured females. Besides, Alavi, Rezaee and Amirian

(2011) in the study of the University of Tehran English proficiency test for master's degree holders in humanities, science and engineering revealed that logistic regression flayed 14 items as exhibiting uniform DIF. However, Alavi, Rezaee and Amiriam attributed the source of the uniform DIF to group variable not the interaction effect. Furthermore, Cormier (2012) discovered the influence of uniform differential item functioning (DIF) for race and gender on STAR Mathematics in all items with logistic regression procedure. The consistency in the use of logistic regression analysis to examine uniform DIF has all been supported in the cited researches. It is imperative that type of DIF be examined in addition to DIF with respect to race, gender among others.

The result on research question two revealed that all the 50 items of 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria, displayed non- uniform DIF. The result supports other previous researches such as the study by Cormier (2012) that revealed non-uniform differential item functioning (DIF) for race and gender on STAR Mathematics in all items using logistic regression procedure amongst male and female; White; Black and Hispanic students. Similarly, Alavi, Rezaee and Amirian (2011) discovered only 5 items that exhibited non-uniformed DIF in the University of Tehran English proficiency test in master's degree humanities, science and engineering. In contrast to the findings of this study, Abedalaziz (2010) discovered that 8 of the 30 items administered to tenth grade students' in Mathematics in Jordan at the end of the First semester school year of 2009 – 2010 displayed non-uniform DIF in a study using logistic regression. Therefore, the study of Non-uniform DIF is a considerable issue in DIF evaluation in educational assessment which is a global concern to ensure fair assessment.

The result that provided answer to research question three revealed that all the 50 items 2014 mock multiple choice Mathematics of AkwaIbom State, Nigeria, displayed negligible DIF effect size. The negligible DIF effect size is considered small. The result indicated that the items did not display the other two types of the effect size classifications of moderate and large DIF effect sizes. This finding is in agreement with some previous studies on classifications of DIF effect sizes when logistic regression analysis was used. Similarly, Salehi and Tayebi (2012) found that 35 items of the University of Tehran English proficiency test (UTEPT) using 3,398 male and female test takers in the partial requirement Ph.D. entrance examination revealed negligible effect size when Jodion and Gierl (2001) NagelkerekeR^2 was used. It was concluded that the items did not favour any particular group of examinees regarding gender. Therefore, those items were considered fair to all. Furthermore, Park (2006) found DIF across language and gender groups in Michigan English Language Assessment Battery (MELAB) that exhibited effect size that was far too small to have an important group effect. However, Parked concluded that with the effect size being too small, the group differences could be attributable to item impact rather than group difference. However, Alavi, Rezaee and Amirian (2011) found that out of the 100 items of University of Tehran English proficiency test, only item 47 displayed moderate or type B effect size when using Jodoin and Gierl (2001) NagelkerekeR^2 classification guidelines, but the other items displayed negligible effect size. Also, Fidalgo, Alavi and Amirian (2014) found 4 items that displayed negligible effect size at non-uniform DIF, while 13 items played moderate effect size at uniform DIF level when Jodoin and Gierl, (2001) classification guidelines was used. But when other method of classifying effect size such as Wald test was used, there was increase in the number of moderate DIF. Also, Cormier (2012) found negligible effect sizes in 554 STAR Mathematics items at uniform and non-uniform DIF when Jodoin and Gierl (2001) classification was used. Cormier concluded that since the effects size was negligible in value, the result demonstrated that the items analyzed did not exhibit bias towards a particular gender or race.

The plausible explanations as to why such results were arrived at in this present study may be attributable to some possible external influences at some centres that must had assisted the students in solving the mathematics problems. Secondly, some correct answers must have been provided that influenced the results.

CONCLUSION

The study revealed that all the 50 items of the 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria, exhibited uniform as well as non-uniform DIF when three step comparative logistic regression models were used. Jodoin and Gierl (2001) classification guidelines of effect size revealed negligible DIF, with no item displayed moderate or large effect size. It was concluded that on the account of the effect size result the items were considered to be fair to all groups since the manifested DIF of the items had negligible value. However, the implication of this result to valid, fair and equality of opportunity in testing is the information that the findings provided should engender further studies that offer some insight into the use of uniform and/ or non-uniform DIF in bias assessment for practical significance.

Recommendations

Based on the f the findings, the following recommendations were made:

- That a comparative analysis of 2014 mock multiple choice Mathematics items of Akwa Ibom State, Nigeria, be made using other DIF detection procedures.
- That other DIF effect size classification guidelines should be used in conjunction with DIF evaluation;
- Other papers than Mathematics at the Junior Secondary School Three (JSS3) and national examinations should be used for comprehensive view of the testing situation in Nigeria.

REFERENCES

Abedalaziz, N. (2010). A gender-related differential item functioning of mathematics test items. *The International Journal of Educational and Psychological Assessment, 5*(2), 101-116.

Acar, T., & Kelecioğlu, H. (2010). Comparison of differential item functioning determination techniques: HGLM, LR and IRT-LR. *Educational Sciences Theory & Practice,10*(2), 639-649.

Alavi, S. M., Rezaee, A. A., & Amirian, S. M. (2011). Academic discipline DIF in an English Language proficiency test. *Journal of English Language teaching and Learning, 7*, 39-65.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English Language assessment.* New York: McGraw-Hill.

Camilli, G. (2006). *Test fairness.* Westport, CT: American Council on Education.

Cormier, D. C. (2012). Evaluating the influence of differential item functioning for race and gender on STAR Mathematics items. Retrieved from www.renlearn.com

De-Beer, M. (2004). Use of differential item functioning analysis for bias analysis in test construction. *Journal of Industrial Psychology, 30*(4), 52-58.

Fidalgo, A. M., Alavi, S. M., & Amirian, S. M. R. (2014). Strategies for testing statistical and practical significance in detecting DIF with logistic regression models. *Language Testing, 31*(4), 433-451.

Huang, J., & Han, T. (2012). Revisiting differential item functioning: Implications for fairness investigation. *International Journal of Education, 4*(2), 195- 250.

Jodoin, G. M., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349.

Joint Committee on Standard for Education and Psychological Testing of AERA, APA, NCME (2014). Retrieved from: http/www.teststandards.org/files/standards.

Le, J. (2006). Analysis of differential item functioning. Paper prepared for annual meetings of the American Educational Research Association in San Francisco.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. New York: Blackwell publishing.

Monaham, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio. Delta ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioural Statistic, 32*, 92-109.

Noortgate, W. V. D., & Boeck, P. D. (2005). Assessing and examining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics, 30*(40), 443-464.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning.* Thousand Oaks, CA: Sage Publishing.

Park, T. (2006). Detecting DIF across Different Language and Gender Groups in the MELAB Essay Test using the Logistic Regression Method. Spain: *Working Papers in Second Foreign Language Assessment, 4*, 81-96.

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Columbia University Working Papers in TESOL & Applied Linguistics, 6*(2), 1-3.

Rezaee, A. A., & Shabani, E. (2010). Gender differential item functioning analysis of the university of Tehran English proficiency test. *Pazhuhesh-e ZaaabanhayeKhareji. Special Issue 56*, 89-108.

Roever, C. (2005). That's not fair! Fairness, bias, and differential item functioning in language testing. Retrieved from http://www2.hawaii.edu/~roever/brownbag.pdf.

Salehi, M., & Tayebi, A. (2012). Differential item functioning (DIF) in terms of gender in the reading comprehension subtest of a high stakes test. *Iranian Journal of Applied Language Studies, l4*(1), 135-168.

Scherman, C. A., & Goldstein, H. W. (2008). Examining the relationship between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement, 68*, 537-553.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*, 402-415.

Walker, C. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psycho educational Assessment, 29*, 364-376.

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretical comparison of methods. Retrieved from http/www.edusci.um.se/digitalAsserts/59/59534_em_no-60pdf.

---

[i] **Cyrinus B. Essen** holds a Ph.D. degree in Educational Measurement and Evaluation, Research and Statistics from the University of Calabar, Cross River State, Nigeria, after a Masters' degree in the same discipline from the same institution. Dr. Essen is a member of Association of Researchers and Evaluator of Nigeria (ASSEREN). His research focus on Test Theories: Classical Test theory (CTT) and Item Response Theory (IRT). He can be reached via email at ecyrinus@ymail.com.

[ii] **Id-Basil F. Ukofia is of the** Department of Educational Psychology, Guidance and Counselling, Federal College of Education (Technical), Omoku, Rivers State. He can be reached via email at ibfukofia86@gmail.com.

[iii] **Dr. Bassey A. Bassey** is of the Department of Educational Foundations, University of Calabar, Calabar, Nigeria**.** He can be reached via email at babassey67@gmail.com.

[iv] **Dr. Delight O. Idika** is of the Department of Educational Foundations, University of Calabar, Calabar, Nigeria**.** He can be reached via email at delightoidika@yahoo.com.