Differential Item Functioning in Senior Secondary School Certificate Examination Multiple-Choice Items: Implication for Technology Adoption in large scale Assessment

Ibrahim Shuaibu
Department of Education, Sule Lamido University kafinHausa
Jigawa State Nigeria
Email: ibahim.shuaibu@slu.edu.ng

Aminu Idris Harbau(PhD)
Deparment of Economics, Saadatu Rimi Colledge of Education Kumbotso Campus
harbau98@gmail.com

Yalwaji Shehu
Department of Psychology. Aminu Sale Colledge of Education Azare
Email: ayshehu53@gmail.com

Dr. Sani Ahmd Katsayal
Department of Education, Sule Lamido University KafinHausa

Abstract

This research investigated differential item functioning of Senior Secondary School Certificate Examination Multiple-choice Mathematics items in Kano state Nigeria, with a view to finding out items that exhibits differential item functioning (DIF) on gender basis. The *ex-post-facto research design was adopted. The population comprised of 65,899 candidates who sat for the examination in 2017. A sample size of 1,000 students was selected using three stage a cluster sampling technique. The population was naturally divided into schools and zones; the instrument for data collection was a sixty-item multiple-choice examination constructed and administered by National Examination Council (NECO) in 2017. The quality estimates of the instrument were ensured by NECO. The data were analyzed using two DIF detection methods (i.e. logistic Regression model, and Mantel-Haenszel statistics). The major findings reveal that items exhibit gender-based uniform and non-uniform DIF with small and moderate effect sizes in NECO 2017 multiple-choice Mathematics questions. It was recommended that, NECO which is involved in development, construction and administration of large-scale examinations should put their hands on deck to prioritize subjecting tests items in to modern statistical evaluations of DIF for a better effect.*

KEY WORDS*:* Differential Item Functioning; Logistic Regression; Senior Secondary School Certificate Examination

Introduction

The plethora of students involved in large scale assessment world over and Nigeria in particular, occasioned by population explosion and desire for qualitative education, coupled with advancement in technology, necessitated the technological adoption in large scale assessment. This is increasingly part of the common discourse within the psychometric community.

Large scale assessment is a form of assessment administered to a large population of students at national or cross-national testing level that provides a snapshot of learning achievement for a group of learners in a given year and in a limited number of learning outcomes. The use of these assessments has been increasing over time and broadening in scope.

According to Emaikwu (2012), there exist a number of national examination bodies and they include National Examination Council (NECO), West African Examinations Council (WAEC), National Business and Technical Examination Board (NABTEB), and Joint Admission and Matriculation Board (JAMB). These bodies cater for candidates of various backgrounds all over the country. Candidates who participate in the examinations conducted by these examination bodies are in different settings, and therefore differently toned for personal and environmental reasons. As a result of this, the problem of test item bias cannot be ruled out in these examinations.

To achieve test fairness, item analysis methods should be used to design reliable, valid and usable test. Item analysis helps to make better decisions about the students (test takers), the instruction, and the test items. Various methods have been designed for item analysis, either in the Classical Test Theory (CTT) or Item Response Theory (IRT). Nevertheless, as test bias became a sensitive concern to the community of test makers or developers (Emaikwu, 2012), several procedures are generated to eliminate biases in tests, among such procedures, is Differential Item Functioning approach or method (DIF).

Differential Item Functioning (DIF) also referred to as “measurement bias” occurs when people from different groups (commonly gender, or ethnicity) with same latent traits (ability/skills) have a different probability of giving a correct response on a test. It is a statistical characteristic of an item that shows the extent to which the item might be measuring different abilities for members of separate subgroups.

There are two types of DIF, which are uniform and non-uniform DIF. Uniform DIF is said to occur when differences in correct response probability are found across all levels of ability for a particular item. Non-uniform DIF on the other hand, occurs when there is interaction between the ability and group membership, such that an item may seem difficult for those at the higher level in one group; and after a particular point, it becomes more difficult for those at lower level in the other group (Abedlaziz, 2010)

Uniform DIF is the simplest type of DIF, where the magnitude of conditional dependency is relatively invariant across the latent trait continuum (θ). The item of interest consistently gives one group an advantage across all levels of the ability (θ) (Walker 2011). Uniform DIF emerges when a sub-group of examinees with ability levels, uniformly endorses a given item or subset of items than the other group. Hence, that particular sub-group is said to be advantaged over the other group and can be underscored as having a striking 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).

Since DIF analysis was put into light of the measurement industry, there has been extensive research and method development for detecting DIF. Hence there is no single “best method” of item bias analysis for all purpose (Anastasi & Urbina, 2009) this is because different methods provide somewhat different kinds of information; it is desirable to employ a combination of methods. However, according to Umoinyang (2013:3), there are several methods of detecting differential item functioning item bias depending on definition of the concept.

Logistic regression is based on statistical modeling of the probability of responding correctly to an item by group membership (i.e. reference group and focal group). Its procedure uses the item response (0 or 1) as the dependent variable, with grouping variable (dummy coded as 1=reference, 2=focal).

Another widely accepted and probably the most popular statistics in use for dichotomous DIF detection is the Mantel-Haenszel (MH) technique (Mantel & Haenszel, 1995) especially when the sample size is small. MH procedure is a chi-squared contingency table-based approach which examines differences between the reference and focal groups on items of a given test. It assumes that the ratio of answering a particular item correctly is equal between reference and focal groups across all ability levels (Lai, Teresi & Gershon, 2005).

Shuaibu and Abba (2019) examined the incidence of differential item functioning (DIF) on items in NECO/SSCE English language multiple-choice examination in Dawakin Kudu Education Zone, Kano State, Nigeria. The results from the study show that, 42% of the items has negligible level of DIF; 20% has moderate level of DIF while 38% of the items has magnitude level of DIF. Noteworthy, it is not certain whether NECO has incorporated element of DIF into its process of test construction, because it has been claimed that some of the national examinations unfairly favour examinees of some particular groups to the extent that it is now believed that a particular section of the country performs most woefully in these national examinations (Emaikwu, 2012). Abedlaziz (2010) conducted a study on gender-related DIF that females showed a statistically significant and consistent advantage over males on numerical ability, whereas males showed a consistent advantage over females on spatial and deductive abilities. Similarly, Madu and Bassey (2010) found that Mathematics multiple-choice test items set and administered by National Examination Council (NECO) function differently between male and female students. Therefore, attempts to find out how well NECO 2017 multiple-choice Mathematics test items function among senior secondary school students in Kano state, Nigeria.

Research Questions

Based on the objectives of this study, the following research questions guided the study:

i. which items on NECO 2017 June/July SSCE multiple-choice Mathematics examination display gender-based uniform and non-uniform DIF?

ii. which items on NECO 2017 June/July SSCE multiple-choice Mathematics examination display gender-based DIF effect levels?

Methods

This paper adopted the ex-post-facto research design. The population consists (65,899) students from public and private secondary schools that sat for NECO 2017 Mathematics multiple-choice test in Kano state, Nigeria. The population is naturally divided into clusters of zones and schools. Eight Education zones were randomly selected in the first stage. In the second stage, two schools were equally selected from each zone. In the third stage, examinees were further selected in the proportion as they exist in the school's population. The instrument for data collection was NECO multiple-choice Mathematics test package (2017) which contains 60 items. Each item consists of a stem and five response options lettered A, B, C, D and E. The items were generated from Senior Secondary School Mathematics syllabus.

Findings

Research Question One: Which item on NECO 2017 SSCE Multiple-Choice Mathematics examination display gender-based uniform and non-uniform DIF?

To address research question one, Logistic regression analysis was done for a total of sixty Mathematics multiple-choice test items using SPSS IBM version 25. The output of the analysis was extracted and presented in the table 1. The decision guide line in this study states that, an item reveals uniform DIF when the significant odd ratio is for the group, whereas the item reveals non-uniform DIF when the significant odd ratio is for the interaction between the group and total score. (Swaminathan & Rogers, 1990, Abedalaziz, 2010).

Table 1: *Gender-Based Uniform DIF & Non-Uniform DIF.*

Item	Variable	Statistical significance	Odd Ratio	DIF Form
Q2	Group	0.788	1.065	Non-Uniform
	Interaction	0.000*	0.418	
Q6	Group	0.024*	1.545	Uniform
	Interaction	0.700	1.104	
Q9	Group	0.380	0.822	Non-Uniform
	Interaction	0.000*	1.108	
Q13	Group	0.557	1.241	Non-Uniform
	Interaction	0.000	0.954	
Q15	Group	0.199	1.354	Uniform
	Interaction	0.000	1.179	
Q16	Group	0.118	1.506	Uniform
	Interaction	0.000	1.186	
Q21	Group	0.223	0.738	Uniform
	Interaction	0.320	1.008	
Q24	Group	0.759	1.093	Non -Uniform
	Interaction	0.000	1.199	
Q25	Group	0.523	1.166	Non-Uniform
	Interaction	0.000	0.978	

Q26	Group	0.912	0.971	Non-Uniform
	Interaction	0.861	0.999	
Q28	Group	0.685	0.928	Non-Uniform
	Interaction	0.000	1.050	
Q29	Group	0.423	0.844	Non-Uniform
	Interaction	0.000	1.071	
Q32	Group	0.071	0.618	Non-Uniform
	Interaction	0.000	1.016	
Q33	Group	0.555	1.193	Non-Uniform
	Interaction	0.000	1.271	
Q34	Group	0.405	1.244	Non-Uniform
	Interaction	0.000	1.26	
Q36	Group	0.149	0.737	Non-Uniform
	Interaction	0.000	1.142	
Q42	Group	0.254	1.339	Uniform
	Interaction	0.880	1.116	
Q43	Group	0.946	1.013	Non-Uniform
	Interaction	0.000	1.231	
Q44	Group	0.155	1.347	Non-Uniform
	Interaction	0.127	0.991	
Q45	Group	0.409	1.263	Uniform
	Interaction	0.658	1.068	
Q46	Group	0.806	0.936	Non-Uniform
	Interaction	0.000	1.172	
Q48	Group	0.077	1.479	Non-Uniform
	Interaction	0.000	1.229	
Q50	Group	0.312	0.749	Uniform
	Interaction	0.880	0.898	
Q52	Group	0.117	1.391	Non-Uniform
	Interaction	0.000	1.066	
Q55	Group	0.554	.889	Non-Uniform
	Interaction	0.000	1.094	
Q60	Group	0.788	1.065	Non-Uniform
	Interaction	0.140	0.855	

Group/Gender = Male & Female

TABLE 2: *Summary Percentage of Gender-Based uniform and non-uniform DIF*

Uniform DIF	7	27%
Non uniform DIF	19	73%
Total	26	100

Table 2 depicts the summary results of the Logistic Regression method to identify uniform and non-uniform DIF on the Mathematic test for each of the sixty items. It could be seen that, 26 items, i.e 43% of the 60 items revealed DIF, while 34 items accounting to 57% did not. Seven items (i.e. items 6, 15, 16, 21, 42, 45 &50) display uniform DIF; whereas 19 items (i.e. items 2, 9, 13, 24, 25, 26, 28, 29, 32, 33, 34, 36, 43, 44, 46, 48, 52, 55 and 60) revealed non-uniform DIF.

Research Question Two: Which item on NECO 2017 SSCE Multiple-Choice Mathematics examination display gender-based DIF effect levels?

To address this research question, Logistic regression procedure on SPSS version 25 was used. Analysis was done for a total of sixty (60) Mathematics multiple-choice test items. The output of the analysis was extracted and presented in the Table 3.

Table 3: *Gender-Based DIF (Effect Size)*

Items	G	Item Measure	G	Item Measure	X^2 Chi-Square	df	Effect Levels
Q2	M	2.13	F	1.39	2.420	1	0.003
Q6	M	-0.30	F	0.21	37.590	1	0.049
Q9	M	0.53	F	0.16	0.757	1	0.001
Q13	M	4.48	F	3.79	0.811	1	0.001
Q15	M	-0.75	F	-0.40	22.339	1	0.030
Q16	M	-1.03	F	-0.58	25.423	1	0.034
Q21	M	2.96	F	2.24	1.944	1	0.003
Q24	M	-0.88	F	-0.45	25.962	1	0.034
Q25	M	0.66	F	0.23	0.216	1	0.001
Q26	M	0.48	F	-0.16	0.757	1	0.001
Q28	M	1.48	F	1.05	0.230	1	0.001
Q29	M	1.25	F	0.63	0.496	1	0.001
Q32	M	0.13	F	-0.46	0.346	1	.000
Q33	M	-0.45	F	-0.01	30.727	1	.040
Q34	M	-0.51	F	-0.11	27.580	1	.036
Q36	M	0.31	F	0.66	29.200	1	.038
Q42	M	0.74	F	0.45	1.899	1	.003
Q43	M	-0.41	F	-0.07	24.917	1	.033
Q44	M	0.27	F	-0.15	.264	1	.000
Q45	M	0.48	F	0.97	40.086	1	.053
Q46	M	-0.52	F	-0.18	23.970	1	.032
Q48	M	0.13	F	0.46	26.913	1	.035
Q50	M	0.73	F	1.05	27.551	1	.036
Q52	M	-1.58	F	-2.20	.771	1	.001
Q55	M	1.41	F	1.77	27.619	1	.036
Q60	M	1.15	F	0.67	.046	1	0.001

A (negligible) DIF: $R^2 < .035$; B (moderate) DIF: $R^2 \leq .070$; C (large) DIF: $R^2 > .070$;

TABLE 4: *Summary Percentage of Gender-Based effect levels*

Moderate effect level	8	31%
Negligible effect level	18	69%
Total	26	100

From the results of the analysis on Tables 3 and 4 above, it could be seen that 26 items, which is43% of the 60 items on NECO 2017 Mathematics multiple-choice examination display different levels of effect size; 18 items (i.e. items 2, 9, 13, 15, 16, 21, 24, 25, 26, 28, 29, 32, 42, 43, 44, 46, 52 and 60,) with Nagelkerke: R^2 <.035; display negligible effect size at uniform and non-uniform DIF; while eight (8) Items (i.e. items 6, 33, 34, 36, 45, 48, 50, and 55,) display moderate effect size at uniform and non-uniform DIF with Nagelkerke: $R^2 \leq .070$. The result indicates that the items did not display the other large: (R^2> .070) of the effect size classifications.

Discussion

Based on the results presented in tables 1 and 2, 23 items on NECO 2017 Mathematics multiple-choice test were found to reveal uniform and non-uniform DIF. Seven (7) out of 23 items displayed uniform DIF; while 19 items exhibited non-uniform DIF. The finding of this study is in harmony with various studies that identified uniform and non-uniform DIF 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, while eight (8) items revealed non-uniform.

Twenty three (23) items with different effect size levels were revealed. Sixteen (16) items revealed negligible effect size levels; whereas seven (7) items were found to have revealed moderate effect size level and that no item was flagged with large effect size level. This supports the findings of Essen at-al (2014) in their research on mock multiple-choice Mathematics questions of Akwa-Ibom State, Nigeria, which displayed negligible DIF effect size. 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.

Conclusion and Recommendations

From the findings of this study, it can be concluded that NECO multiple-choice Mathematics items of 2017 display gender-based uniform and non-uniform DIF, and that 16 items reveal gender-based negligible effect size at uniform and non-uniform level, while seven items reveals moderate effect size, whereas no item displayed large gender-based effect size at uniform and non-uniform level. The following recommendations were therefore made:

i. Since it is inappropriate to judge the adequacy of any test item without subjecting it to the process of item analysis, National Examinations Council (NECO) should consider and give priority to review of item analysis before the final administration of each test. This is in addition to the usual judgmental review process.

ii. The detection of DIF items should be an important factor to be considered in any examination. This is because, if an assessment tool is biased, then obviously performance of the students would be influenced by some factors such as gender, School type, specialization etc.

iii. Detection of forms/types of DIF should always be carried out alongside DIF analysis at all times; this will help in revealing the items which display DIF, uniform or non-uniform and their respective effect sizes.

REFERENCES

Anastasi, A. & Urbina, S. (2009). *Psychological Testing.* 7th Edition. New Delhi: PHI Learning Private Ltd.

Abedlaziz, N. (2010). A gender-related differential item functioning of Mathematics test items. *The International Journal of Educational and Psychological Association 5,* 101-116.

Brown, G. T., & Hattie, J. (2012). The benefits of regular standardised assessment in childhood education: guiding improved instruction and learning in Contemporary debates in childhood education and development 301-306. Routledge

Clauser, B. E., & Mazor, K. M. (1 998). Using statistical procedures to identify differential item functioning test items. Educational Measurement: Issues and Practice, 17, 31-44.

Emaikwu, S. O. (2012) Issues in Test Item Bias in Public Examinations in Nigeria and Implications for Testing. *International Journal of Academic Research in Progressive Education and Development* Vol. 1, No. 1 ISSN: 2226-6348

Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education promises and implications for teaching and learning. *The Center for Curriculum Redesign,* Boston, MA

Lai, J. S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation & the Health Professions. 28* (3), 283-294

Madu, B.C & Bassey, H. V.I (2010). Gender related differential item functioning in Mathematic Multiple Choice Test Items Set and Administered by National Examination Council (NECO*)* *Journal of Mathematical Sciences and Education 1* (1), 101-114

Moyinoluwa, T. D. (2015). Analysing the Psychometric Properties of Mathematics in Public

Examinations in Nigeria, *Research on Humanities and Social Sciences* (Paper)2224-5766 ISSN (Online) 2225-0484. Vol.5, No.7, 2015

Oladele,J.I.,(2022). Review of Fasttest for Electronic Item Banking For Standardised Assessments: Implications For The Fourth Industrial Revolution and Covid-19 Interjections. *Journal of Research and Reviews in Social Sciences Pakistan, 5 (1),* 1498-1518.

Ridwan, S. M., Felix, T. S., & Mohammed, M. F. (2019). Assessment of ICT competencies and use of electronic information resources by lecturers in universities in Benue state, *Nigeria. International Journal of Information Management Sciences* (IJIMS),1–20.

Shuaibu, I., & Abba, B.(2019) Identification of school type differential item functioning of 2010 senior school certificate (NECO) English language examination among students of Dawakin Kudu Education Zone-Kano state, Nigeria. *International Journal of Advanced Educational Research,* 4(3), 31-35.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

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