DETECTION OF UNIFORM AND NON-UNIFORM GENDER DIFFERENTIAL ITEM FUNCTIONING IN ECONOMICS MULTIPLE CHOICE STANDARDIZED TEST IN NIGERIA

Amaechi, Chidera Emmanuel
Department of Science Education
Michael Okpara University of Agriculture, Umudike
&
Onah, Frederick Ekene *Ph.D*
Department Of Life Science Education
Imo State University, Owerri

Abstract
The researchers detected uniform and non-uniform gender differential item functioning in Economics multiple choice standardized test in Nigeria. One research question and one hypothesis guided the study. The design of this study is a survey which involved the inferential method. The population of the study was 4,434,979 secondary school students in the 11,875 *public secondary schools in the 36 states including Abuja the Federal Capital Territory (FCT) of Nigeria involving purposive and simple random sampling techniques. The instruments for data collection were Socio-Demographic Inventory (SDI) and a 50 item WAEC General Economics Paper I Multiple Choice Test. The instruments were revalidated by five specialists, three in Educational Measurement and Evaluation and two from Economics Education from Michael Okpara University of Agriculture Umudike and Imo State University, Owerri. The reliability of the test was reestablished using Kuder-Richardson formular 20 (KR-20) statistics with an index of 0.80. In answering the research question, IRT-Binary Logistic Regression method was used while the hypotheses were tested using Wald test associated with binary logistic regression statistics at 0.05 level of significance. The result indicated that, out of the 13 items that have DIF issues, 8 fall under the category of uniform DIF while 5 were non-uniform DIF. The uniform DIF items are 13, 15, 20, 24, 37, 39, 42 and 46; while the non-uniform DIF items are 9, 22, 30, 37 and 44. This shows that 8 items significantly displayed uniform DIF while 5 items significantly displayed non-uniform DIF. It was recommended that examining bodies, evaluators, and all other educational practitioners should be mindful and pay serious attention on gender when setting examination items.*

Introduction

Economics is one of the subjects offered in senior secondary schools in Nigeria. It is one of the social science subjects that heavily utilize statistical

and mathematical models to analyze real-life problems. It is a social science that studies human behaviour in an effort to allocate scarce resources efficiently and effectively in order to minimize cost. As such, it is a subject concerned with the efficient utilization or management of limited or scarce resources for maximum satisfaction of human needs. Given the above information, there is need to add that Economics, as a social science discipline, helps one to understand and manage his scarce resources in order to meet his numerous needs. In line with this, Federal Republic of Nigeria, FRN (2013) advocated that Economics, when taught in secondary schools, will equip the recipients with the knowledge on how to allocate scarce resources, make choices, and take rational decision on pressing economic issues. As such, to achieve the goals and objectives of teaching Economics at the Senior Secondary School level, the assessment of student's academic achievement demands attention.

Academic achievement is used interchangeably with academic performance in this study and it is defined as the degree or level of success attained at the end of an academic endeavour. Chowdhury and Pati (2011) opined that academic achievement is defined by test or examination marks and grades given by teachers' in academic subjects. However, this academic achievement could be affected by assessment methods or testing instruments. As such, some test scores fail to produce the true learning outcome. This could be seen in the poor examination result of Senior Secondary School in Nigeria.

Test is an instrument for systematic measure of a person's behaviour or academic traits with the aid of a numerical scale. It is the easiest instrument to measure the cognitive domain of learners. Testing has become one of the most important procedures and parameters by which a society adjudges the product of her educational system. It has always been an important part of the school system that even the habitual absentees normally turn up in school and present themselves for testing on examination days. Despite these functions of tests and the fact that modern societies have adopted testing as the most objective means of decision making in education, various criticisms have been raised on issues concerning testing in recent times (Amajuoyi, 2015). From the description of test and testing above, a test is supposed to measure students/examinees' ability/performance or other latent traits irrespective of the sub-group they belong. Tests are expected to provide equal opportunities to all examinees to demonstrate their latent construct, abilities and knowledge irrespective of their socio-demographic factors like gender; location; ethnic, cultural and religious groups. Latent construct here means unseen or heeding abilities possessed by a learner.

Buttressing the assertion, Osterlind and Everson (2009) commented that the response to a particular test item is determined by the latent construct of interest, referred to as theta (θ), being measured. The achievement tests, have been faulted (Nworgu, 2015; Moyo & Nenty, 2017) for limitations and bias in what they intend to measure. The existence of bias introduces measurement error and hence decreases the validity of the entire test. Bias is the presence of some characteristic of a test and/or an item in a test that results in differential performance for two individuals of the same ability but from different ethnic, sex, cultural or religious groups. When the whole test is the unit of concern, it is referred to as test bias. Amajuoyi (2015) defined test bias as objective statistical indices that examine the patterning of test scores for relevant subpopulations. Hence, the above violates the unidimentionality assumption of Item Response Theory (IRT). This assumption postulates that only one ability is measured by the items that make up a test.

With the use of Item Response Theory (IRT) framework in the analysis of test items, psychometricians (Moyo & Nenty, 2017) have found that some items in a test may function differently from what the test is meant for. It means that such items have interactions with the characteristics of the sample (examinees/students) taking the test. Differential Item Functioning (DIF) as defined by Angoff cited in Moyo and Nenty (2017) referred to the differences in the statistical properties of an item between groups of equal ability. It is intended to be invariant with respect to irrelevant aspects of the test-takers, such as gender, ethnicity and socio-economic status. It is also expected to be altered by interventions targeted at those items, for instance, the use of calculators in arithmetic tests or the use of assistive device on mobility tests. As such, differential item functioning (DIF) occurs when examinees from different groups show differing probabilities of success on the item after being matched on the underlying ability that the item is intended to measure (Walker, 2011). Lending credence to this, Ajeigbe and Afolabi (2014) contended that Differential item functioning (DIF) referred to a difference in the way a test item functions for comparable groups of test takers. Formally defined, an item displays DIF if subjects of equal proficiency, or equal ability level on the construct intended to be measured by a test but from separate subgroups of the population differ in their expected score on this item. It is noteworthy that when assessing an item for DIF, the groups must be matched on the measured attribute; otherwise this may result in inaccurate detection of DIF (Moyo & Nenty, 2017; Zumbo, 2012).

Two types/categories of DIF may be investigated within the IRT framework; they are uniform (systematic) and non-uniform (unsystematic) DIF. Bao, Dayton and Hendrickson (2009) opined that uniform or systematic

DIF exists when the Item Characteristic Curves (ICCs) for the two groups do not cross over the entire latent trait continuum (θ). Uniform DIF is said to apply when differences between groups in item responses are found at equal/all ability levels. Uniform DIF is said to occur when differences in correct response probability are found across all ability levels for a particular item. This implies that the item of interest consistently gives the reference group an advantage across all the ability θ range (Walker, 2011). Non-uniform or unsystematic DIF presents an interesting case. Rather than a consistent advantage given to the reference group across the ability continuum, the ICCs for the two groups cross at some points on the θ scale showing that the item is at different points on the ability continuum, functioning differently against the focal group who has a low ability level and also functioning differently against the reference group who has a high ability level (Penfield & Camilli, 2007; Bao, Dayton & Hendrickson, 2009).

Gender is a set of characteristics distinguishing between males and females, particularly, in the case of man and woman which, depending on the context, may vary from sex to social role and to gender identity (Bland, 2013). Nevertheless, Chang (2013) reported that although there is a decrease in the gap in gender difference in students' performance in subjects, female representation in subjects involving calculation like sciences is still low in comparison with their male counterparts. Considering socio-cultural background, Oludipe (2012) observed that in Nigeria, certain vocations and professions have traditionally been regarded as men's and others; women. Some of these vocations are Medicine, engineering, architecture and nursing, catering, typing respectively. The society's socio-cultural construct of females as weaker sex, together with females self perception of themselves as weaker sex, inferior and subordinate to the males, have imposed some socio-cultural limitations on female aspirations and achievement in sciences (Ojobo, 2008). Similarly, Nzewi (2010) inferred that the socio-cultural upbringing of females within most Nigerian homes tends to shape the girl-child away from science and science related disciplines. For instance, in most homes, what are regarded as complex and difficult tasks are allocated to boys whereas girls are expected to handle the relatively easy and less demanding tasks. Consequently, fewer females opt for science subjects thereby creating some differences in the number of males and females in science discipline in favour of the males.

Empirically, Essen, Ukofia, Bassey and Idika (2017) result indicated that all the 50 items displayed uniform and no uniform differential item functioning (DIF). The result also 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. Also, 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.

It has been claimed that some of the national examination bodies such as West Africa Examination Council (WAEC), National examination council (NECO) and National Business and Technical Examination Board (NABTEB) may unfairly favour examinees of some particular groups. This could be associated with item bias or lack of un dimensionality of test items and gender bias. If this persists, it can cause doubt on the validity of such tests and may lead to an unfair advantage or disadvantage for certain subgroups in educational testing, which might also affect decision making in such regard. The study posed as a question is: how many items in Economics standardized test displayed uniform and non-uniform gender DIF? The answers to these questions are what the present study provided.

The general purpose of this study was to detect uniform and non-uniform gender differential item functioning in Economics multiple choice standardized test in Nigeria. Specifically, the study ascertained:

1) the number of uniform and non-uniform gender DIF items using IRT-Binary Logistic Regression method,

The following research questions and hypothesis were posed and they guided the study:

-What is the number of uniform and non-uniform gender DIF items using IRT-Binary Logistic Regression method?

-The number of uniform and non-uniform gender DIF items using IRT-Binary Logistic Regression method is not significant.

Method

The design of this study is a survey which involved the inferential method. The population of the study was 4,434,979 secondary school students in the 11,875 public secondary schools in the 36 states including Abuja the Federal Capital Territory (FCT) of Nigeria involving purposive and simple random sampling techniques. The instruments for data collection were Socio-Demographic Inventory (SDI) and a 50 item WAEC General Economics Paper I Multiple Choice Test. The instruments were revalidated by five specialists, three in Educational Measurement and Evaluation and two from Economics

Education from Michael Okpara University of Agriculture Umudike and Imo State University, Owerri. The reliability of the test was reestablished using Kuder-Richardson formular 20 (KR-20) statistics with an index of 0.80. In answering the research question, IRT-Binary Logistic Regression method was used while the hypotheses were tested using Wald test associated with binary logistic regression statistics at 0.05 level of significance.

Results

Table 1:

IRT Binary Logistic Regression for Uniform and Non-Uniform Gender DIF Analysis

	Non-Uniform		**Uniform**	
Items	**Wald test/χ^2-test**	**Prob.**	**Wald test/χ^2-test**	**Prob.**
1	0.15	0.6993	0.00	0.9436
2	0.36	0.5476	0.06	0.8111
3	1.91	0.7086	0.00	0.9531
4	1.71	0.1910	2.59	0.1077
5	3.51	0.0611	0.02	0.8832
6	2.01	0.0790	0.58	0.4456
7	0.41	0.5209	0.07	0.7861
8	0.29	0.5912	0.42	0.5152
9	4.44	**0.0351**	3.07	0.0799
10	1.16	0.2824	0.76	0.3826
11	0.74	0.3898	2.09	0.1481
12	0.01	0.9090	0.03	0.8590
13	3.37	0.0663	74.41	**0.0000**
14	0.51	0.4744	2.26	0.5390
15	0.52	0.4709	59.56	**0.0000**
16	1.39	0.2382	1.04	0.3077
17	1.98	0.1593	0.00	0.9765
18	1.76	0.1850	1.74	0.8806
19	1.43	0.9037	0.13	0.7208
20	2.47	0.0626	19.04	**0.0000**
21	1.11	0.2925	0.27	0.6027
22	4.20	**0.0376**	2.25	0.1340
23	0.55	0.4582	2.04	0.5444
24	0.03	0.8738	62.44	**0.0000**
25	2.19	0.0739	0.87	0.3500
26	0.07	0.7945	1.57	0.2096
27	0.45	0.6111	2.86	0.2495
28	0.52	0.4713	0.32	0.5732
29	0.61	0.4363	0.06	0.8091
30	4.32	**0.0377**	2.75	0.0974
31	2.02	0.4251	0.00	0.9450
32	1.99	0.5558	1.74	0.1867

33	0.62	0.4305	1.57	0.7007
34	0.34	0.5627	1.03	0.3110
35	0.08	0.8015	3.03	0.0700
36	0.56	0.4534	0.02	0.8853
37	0.19	0.6667	9.57	**0.0020**
38	1.26	0.2608	0.09	0.7670
39	1.57	0.2097	10.55	**0.0012**
40	0.05	0.8175	0.07	0.7955
41	0.07	0.7897	0.50	0.4777
42	1.09	0.2974	23.36	**0.0000**
43	0.00	0.9759	0.00	0.9965
44	17.77	**0.0000**	1.92	0.7601
45	3.83	0.0504	0.00	0.9670
46	2.88	0.0688	24.46	**0.0000**
47	4.23	**0.0314**	1.38	0.7001
48	0.20	0.6533	0.67	0.4130
49	0.42	0.5187	0.83	0.3623
50	0.41	0.5197	0.92	0.3372

In Table 1, the number of WAEC items in Economics that have uniform and non-uniform DIF with respect to gender were detected. With Wald/Chi-Square test significance level below 0.05, the result indicated that, out of the 13 items that have DIF issues, 8 fall under the category of uniform DIF while 5 were non-uniform DIF. The uniform DIF items are 13, 15, 20, 24, 37, 39, 42 and 46; while the non-uniform DIF items are 9, 22, 30, 37 and 44. This shows that 8 items significantly displayed uniform DIF while 5 items significantly displayed non-uniform DIF.

Discussion

The study revealed that 8 items significantly displayed uniform DIF while 5 items significantly displayed non-uniform DIF. This finding proved the effectiveness of Binary logistic regression in detecting uniform and non-uniform DIF. 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. In contrast to uniform DIF, 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. 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. The above finding is in line with Essen, Ukofia, Bassey and Idika (2017) result which indicated that all the 50 items displayed uniform and non uniform differential item functioning (DIF). The result also 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. Also, 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.

Conclusion

The study detected uniform and non-uniform gender differential item functioning in economics multiple choice standardized test in Nigeria. Based on the findings accruing from this study, it was concluded that 8 items significantly displayed uniform DIF while 5 items significantly displayed non-uniform DIF. Hence, they are flagged as biased items.

Recommendations

The following recommendations are made based on the findings and discussions of the study;

a) Examining bodies, evaluators, and all other educational practitioners should be mindful and pay serious attention on gender when setting examination items.
b) Also, WAEC as a board should ensure that items to be used to examine their Economics students are free from any form of DIF.

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.

Ajeigbe, T. O. & Afolabi, E. R. I. (2014). Assessing unidimensionality and differential item functioning in qualifying examination for Senior

Secondary School Students, Osun State, Nigeria. *World Journal of Education, 4*(4), 30-37.

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*(1), 39-65.

Amajuoyi, I. J. (2015). Verification of differential item functioning (DIF) status of West African Senior School Certificate Examination (WASSCE) in Chemistry. *Journal of Educational Foundations, 5*(1), 165-175.

Bao, H., Dayton, C. M., & Hendrickson, A. B. (2009). Differential item functioning amplification and cancellation in reading test. *Practical Assessment, Research & Education, 14*(19), 1-27.

Blaug, M. (2007). The social science: Economics, Macro-Economics. *The New Encyclopedia Britannica, 27*(1), 349-351.

Chang, Y. (2013). Gender Differences in Science Achievement, Science Self Concept and Science Values. [*On line] available @Yuan Christian University @tea.ntue.edutw:* Pp1-6.

Chowdhury, B. R. & Pati, N. L (2011). Achievement motivation for teachers' education: Analysis of attitude about future success for residence in poor urban neighborhood. *Psychological Focus, 35*(2) 2001-2008.

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

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.

Federal Republic of Nigeria (2013). *National policy of education.* Abuja: Federal Government press.

Moyo, S. E. & Nenty, H. J. (2017). Dimensionality analysis of students' performance in 2013 BGCSE agricultural examination: Implications for differential item functioning. *Research Journal of Educational Sciences, 5*(1), 1-10.

Nworgu, B. G. (2015). *Educational research: Basic issues and methodology (Third Edition).* Nsukka, Enugu: University Trust Publishers.

Nzewi, U. M. (2010). It's All in the brain of gender and achievement in science and technology education. *51st Inaugural Lecture of the University of Nigeria Nsukka* pp 18-32.

Ojobo, J. A. B. (2008). Education: A catalyst for women empowerment. *Journal of Education and Science, 4*(1), 5-15.

Oludipe, O. I. (2012). Gender difference in Nigerian junior secondary students' academic achievement in basic science. *Journal of Education and Social Research, 2*(1), 93-99.

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

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias, in S. Sinharay and C.R. Rao, (eds.) *Handbook of statistics: Psychometrics.* New York, NY: Elsevier.

Walker, C. (2011). What's differential item functioning? Why differential item functioning analyses are an important part of instrument development. *Journal of Psycho-Educational Assessment, 29*(1), 364-376.

Zumbo, D. B. (2012). A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and likert-type item scores. Retrieved on 06/06/2012 from: *http://edu.ubc.ca/faculty/zumbo/dif/handbook.pdf.*