International Journal of Advanced Education and Research
www.alleducationjournal.com
ISSN: 2455-5746
Received: 01-05-2021, Accepted: 16-05-2021, Published: 02-06-2021
Volume 6, Issue 3, 2021, Page No. 18-22

Determination of differential item functioning by gender in the national business and technical examinations board (NABTEB) 2017 physics multiple choice examination in Delta state

Romy O Okoye, Benedicta I Fejokwu
Department of Educational Foundations, Faculty of Education, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria

Abstract
The purpose of the study was to investigate the differential item functioning (DIF) by gender in National Business and Technical Examinations Board (NABTEB) 2017 physics multiple choice test items examination in Nigeria. This was conducted by determining the items that functioned differentially for male and female examinees. A survey research design was employed. A sample of 3,425 examinee responses which comprised 2,310 males and 1,115 females was selected from two Local Government Areas of Delta State, Nigeria out of 14,284 of the total population of examinees in 2017. A 50-item multiple choice physics test item was used for data collection. To detect the items that functioned differentially for male and female examinees, Area Index (Raju) method which is one of the item response theory methods of DIF detection was applied. The results of the analysis revealed that male and female examinees functioned differentially in seventeen items (34%) and no difference in 33 items (66%). Out of the seventeen items, six items were in favour of male students while 11 items were in favour of the female students. Based on the result of the findings, it was then recommended among others that for bias-free items to be produced, examination bodies, test experts and developers should make certain that activities and connotations reflected in the test are relevant to the construct being measured and explore the use of Area Index method of DIF to detect the items that function differentially by gender.

Keywords: national business and technical examinations board, physics multiple, differential item functioning

Introduction

In education, test is crucial in determining students' academic achievement. The test could be used for promotion, certification, recruitments, placement, and so on depending on the purpose, using an instrument with valid and reliable items. The qualities of measuring instruments depend mainly on the quality of items used in the instruments. There is need to ensure that items are not only valid and reliable but also fair to all across the subgroup of examinees (male and female). Indeed, the Federal Republic of Nigeria (2014), in the National Policy on Education stated that national examinations tests should be valid and fair to all students to which the test is set to measure the attributes needed from them. Test fairness is a crucial issue in testing and it reflects the same constructs for all examinees and scores have the same meaning for all individuals in the intended population.

An important step in the construction of assessment instruments is to ensure that no individual or group responding to the instrument is disadvantaged in any way (Kanjee, 2017) [7]. For instance, in an achievement test, students of equal ability drawn from the same population but belonging to different subgroups such as male or female, should have the same probability of getting an item correct. This can only be hindered when the item is biased. Biased test items are those that differentially inhibit individuals from showing their true abilities and thereby measuring irrelevant construct. Such items are said to be displaying differential item functioning (DIF) which according to Reynolds (2016) [13] systematically underestimates or overestimates the value of the variable the items are designed to measure. DIF exists in a test item when, despite controls for overall test performance, examinees from different groups have a different probability of getting an item correct or when students from two sub-populations with the same ability level have different expected scores on the same item (Penfield & Camilli, 2017) [12].

Indeed, DIF occurs when examinees from different groups have different likelihoods of success on an item, after they have been matched on the ability of interest (Clauser & Mazor, 2018) [5]. The presence of DIF is as a result of some characteristics in an item that result in differential performance for individuals of equal ability but from different group. Items may be judged relatively more or less difficult for a particular group by comparison with the performance of another group drawn from the same population. Differential item functioning of an item can therefore be understood as a lack of conditional independence between an item response and group membership (often gender, location or ethnicity) given the same latent ability or trait (Taiwo & Eyitayo, 2014)

It is crucial to match groups, since the comparison should establish a distinction between differences in item responses from divergences between two groups. For example, in a physics test which needs calculation ability, experimentation and English-reading comprehension, consider examinees with the same calculation ability. However, one group is more competent in English-reading comprehension than the other group. If the two groups show differences in the probability of answering some of the items of the test correctly, due solely to differences in English proficiency, the items can be said to possess DIF.

It is essential for test developers or test users to investigate whether items influence examinees' performance in systematically biased ways for some particular subgroups due to any extraneous sources of variance. Thus, if there are

DIF items, it means that irrelevant factors which probably have effects on the responses, but are not interested in, are driving the responses beyond the latent variable that is purportedly measured (Ackerman, 2012). If some items function unfavorably over specific groups, the explanations made from the test cannot be thought of as valid and fair.

One of the ways to investigate bias of items at the item level is through DIF analysis. A DIF analysis is a means of statistically identifying unexpected differences in performance across matched groups of examinees. It compares the performance of matched majority (or reference) and minority (or focal) group examinees. There are several methods of detecting DIF. Some of these methods are based on Classical Test Theory (CTT) which include Mantel-Haenszel (MH), Logistic Regression (LR) and Simultaneous Item Bias (SIBTEST). Other methods such as Lord's chi square test, Raju's area measures and IRT-Likelihood Ratio (IRT-LR) are examples of DIF detection methods based on Item Response Theory (IRT). Most of these methods provide similar but not identical information about DIF. This study focused on the Raju's Area method. The Raju's area measure is based on quantifying the gap between item characteristics curves functions. According to Oshima and Morris (2013), the approach provides an intuitive and flexible methodology for assessing differential item functioning.

Statement of the Problem

Recurrent poor academic performances and achievement difference in physics across Nigerian secondary, technical and business schools seem to be unabated. Some researchers attribute the poor performance of students in Physics to factors such as teachers, environment, parents, and facilities. However, there seems to be a persistent better performance of male groups over female groups in physics test items. Ogbebor and Onuka (2013) [10] found that male students perform better in Physics than their female counterparts. Could these differences in performance be as a result of the nature of test items used that make one gender to perform better than the other? One would wonder if test items are fair to all test takers or if some items are in favour of one group over the other. Thus, the researchers deem it necessary to investigate if the multiple choice Physics test items administered by National Business and Technical Examinations Board (NABTEB) function differentially by gender using Raju method of Item Response Theory based approach. NABTEB is one of the certification bodies in Nigeria responsible for conduct of examinations leading to the award of the National Technical Certificate (NTC), Advanced National Technical Certificate (ANTC), National Business Certificate (NBC), and Advanced National Business Certificate (ANBC).

Concept of Area Index for Two-Parameter Logistic Model (Raju's Area Method)

Area index for two-parameter logistic model is used to measure the area between the two item characteristic curves (ICCs) of the reference and the focal groups as an index of the difference between the performances of the two groups matched on ability. The larger the area, the larger the difference between the two curves (Abedlazeez, 2010) [1]. An item is said to possess differential item functioning when the area index is greater than a critical value of 0.22, while an item does not possess differential item functioning when the area index is zero or close to zero (De Beer, 2014). Also, according to Ling and Lau (2013) [8], when the b parameter (item difficulty) for one group (for example, male) is greater than that of the other group (for example, female), this shows that the item is more difficult for the male group and consequently is said to favour the other group (that is, female), and vice versa.

Raju formula for area index between two curves is as follows*:*

$$\text{Area} = \left| 2\frac{(a_2 - a_1)}{Da_1a_2} L_n\left[1 + e^{Da_1a_2}\frac{b_2 - b_1}{a_2 - a_1}\right] - (b_2 - b_1) \right|$$

Where

a_1: discrimination parameter for males (reference group)
a_2: discrimination parameter for females (focal group)
b_1: difficulty parameter for males (reference group)
b_2: difficulty parameter for females (focal group)
D = 1.7 (constant: scaling factor)

Purpose of the Study

The objective of this study is to determine the differential item functioning of the 2017 NABTEB Physics test items in Delta State, Nigeria. Secondly, this study is aimed to find out if there is a difference in the number of items functioning differentially by gender in the 2017 NABTEB multiple choice Physics examination test.

Research Questions

To carry out this study, the following research questions were posed:

1. What percentage of items in the 2017 NABTEB multiple choice Physics examination function differentially by gender?
2. How many of the items that functioned differently were in favour of each of the two gender groups?

Hypothesis

1. There is no significant difference between number of items that functioned differently in favour of males and those functioning differently in favour of male and females

Method

The research design adopted for this study was the survey research design. This design was considered appropriate because only a part of the population was studied and findings from this was used to generalize for the entire population. The population of the study comprised 14,284 candidates that enrolled and sat for the National Business and Technical Examinations Board (NABTEB) 2017 May/June Physics multiple choice examination in Delta State, Nigeria. The total number of candidates sampled for this study was 3,425, made up of 2,310 male and 1,115 female students. To obtain this number, cluster sampling technique was used. Thus, the population was stratified according to local government areas. Two local government areas were obtained through simple random sampling, and the scores of all candidates in those two local government areas were used for analysis.

The instrument used to collect data was a 50-item May/June 2017 NABTEB Physics multiple-choice question paper. The responses of the candidates to these items were already existing in NABTEB office by the time of the study. The researchers simply went to NABTEB office and obtained the gender and responses of the candidates, as well as the keys to the various items.

Principal Component Analysis was performed to test for uni-dimensionality of the multiple choice Physics test items. Item parameters were estimated using the computer program Xcalibre Version 4.2.0.1 (Assessment Systems Corporation, 2013). Raju Area Measure method was used to determine the presence of differential item functioning. An item is said to possess differential item functioning when the area index is greater than a critical value of 0.22, while an item does not possess differential item functioning when the area index is zero or close to zero (De Beer, 2004). Also, according to Ling and Lau (2003), when the b parameter (item difficulty) for one group (for example, male) is greater than that of the other group (for example, female), this shows that the item is more difficult for the male group and the item is said to favour the other group (that is, female), and vice versa. The hypothesis was tested using chi-square statistic at 0.05 alpha level of significant.

Results

Unidimensionality of the test items was considered before analyzing DIF. This was done because undimensonality is an assumption of item response theory (IRT). The method used in this study for assessing the undimensonality was principal component analysis which was done on the dichotomous items using a sample size of 3,425 students. The result of the scree plot of eigen values is shown in Fig 1.The examinees' performance in the Physics examination was accounted for by a single latent trait/ ability due to the dominating factor. The scree plot shows that the unidimensionality assumption was not by the items.

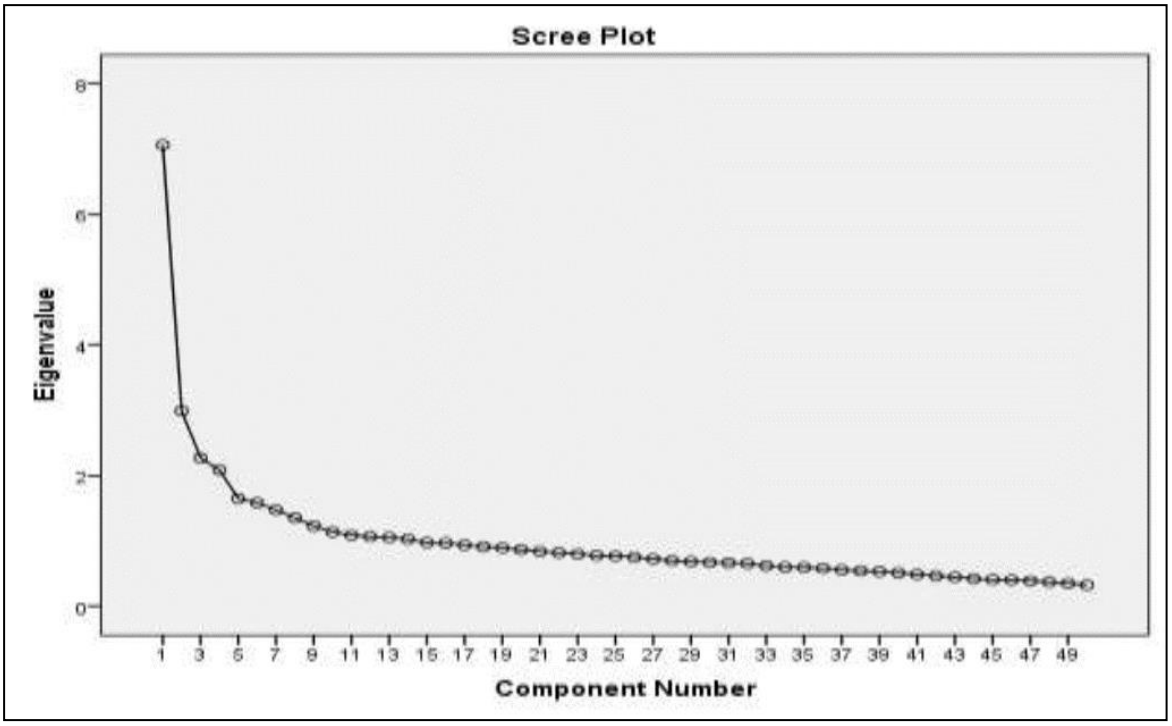

Fig 1: Scree plot of Eigen value (2017 NABTEB Physics)

Research Question 1 was concerned with the percentage of items that functioned differently according to gender. The result is shown in Table 1.

Table 1: Summary of Area Index of 2017 NABTEB Physics multiple choice Items

Item	a1 (male)	b1 (male)	a2(female)	b2(female)	Area Index	Decision	Favoured Group
1	0.453	-3.308	0.365	-3.503	0.291	DIF	Female
2	1.859	-0.12	0.682	-0.828	1.060	DIF	Female
3	0.803	-2.227	0.811	-2.183	0.043	NON-DIF	
4	0.876	-2.855	0.828	-2.844	0.025	NON-DIF	
5	0.831	-2.72	0.813	-2.921	0.228	DIF	Female
6	0.55	-2.917	0.71	-2.482	0.229	DIF	Male
7	3.003	-0.382	0.818	-1.548	3.239	DIF	Female
8	0.74	-1.863	0.686	-2.031	0.207	NON-DIF	
9	0.602	-1.602	0.868	-1.217	0.233	DIF	Male
10	0.622	-2.083	0.709	-1.963	0.104	NON-DIF	
11	1.293	-1.008	1.252	-1.094	0.122	NON-DIF	
12	1.147	-0.901	1.063	-1.081	0.297	DIF	Female
13	0.707	-2.23	0.846	-2.03	0.157	NON-DIF	

14	0.831	-2.074	0.837	-2.024	0.049	NON-DIF	
15	3.166	-0.328	0.971	-1.184	3.783	DIF	Female
16	0.804	-1.833	0.793	-1.85	0.017	NON-DIF	
17	1.052	-1.483	1.146	-1.383	0.096	NON-DIF	
18	0.969	-1.423	0.874	-1.639	0.310	DIF	Female
19	1.077	-1.554	0.979	-1.767	0.341	DIF	Female
20	0.619	-2.317	0.657	-2.223	0.082	NON-DIF	
21	1.007	-1.871	1.152	-1.818	0.212	NON-DIF	
22	0.477	-1.46	0.871	-0.774	0.167	NON-DIF	
23	0.752	-1.935	0.839	-1.803	0.112	NON-DIF	
24	0.516	-2.292	0.598	-1.789	0.335	DIF	Male
25	1.167	-1.044	1.065	-1.187	0.183	NON-DIF	
26	0.807	-1.324	0.885	-1.141	0.144	NON-DIF	
27	0.903	-1.48	1.097	-1.407	0.316	DIF	Male
28	1.017	-1.248	1.224	-1.535	0	NON-DIF	
29	1.251	-0.405	0.745	-0.851	0.282	DIF	Female
30	1.468	-0.579	1.13	-0.843	0.071	NON-DIF	
31	0.695	-1.6	0.85	-1.197	0.243	DIF	Male
32	0.926	-1.519	0.976	-1.305	0.177	NON-DIF	
33	1.161	-1.481	1.237	-1.446	0.075	NON-DIF	
34	0.79	-1.397	0.897	-1.325	0.094	NON-DIF	
35	0.93	-1.674	1.33	-1.446	0	NON-DIF	
36	2.394	-0.236	1.451	-0.458	1.561	DIF	Female
37	2.353	-0.694	1.252	-1.21	1.455	DIF	Female
38	1.26	-1.187	1.349	-1.118	0.096	NON-DIF	
39	0.654	-1.447	0.775	-1.158	0.195	NON-DIF	
40	1.088	-1.472	1.031	-1.541	0.074	NON-DIF	
41	1.2	-0.886	1.462	-0.892	0	NON-DIF	
42	0.081	2.667	0.106	2.694	0.020	NON-DIF	
43	1.192	-1.286	1.152	-1.399	0.161	NON-DIF	
44	1.474	-1.15	1.62	-1.176	0	NON-DIF	
45	0.999	-1.295	0.944	-1.353	0.059	NON-DIF	
46	0.527	-0.75	0.618	-0.161	0.384	DIF	Male
47	0.779	-1.568	0.883	-1.34	0.168	NON-DIF	
48	0.925	-1.339	1.013	-1.325	0.064	NON-DIF	
49	1.026	-1.415	1.057	-1.402	0.017	NON-DIF	
50	0.243	2.667	0.298	2.694	0.056	NON-DIF	

Table 1 shows the summary of area index of the items. It displays the items that exhibited DIF and the group each of them favours. The result shows that 17 out of 50 items, representing 34% functioned differently according to gender, each with an area index greater than the critical value of 0.22, and 33 items, representing 66%, did not function differently, each with an area index less than 0.22.

Hypothesis 1

Table 2: Chi-square Summary of Differential Item Functioning in Favour of Male and Female Students

Gender	Item favoured due to DIF	Df	Chi-square	Sig.)2-tailed)
Male	6 (8.5)	1	1.47	0.225
Female	11 (8.5)			

$\alpha = 0.05$

The results shows a chi-square value of 1.47 and a p-value of 0.225 at .05 alpha level. Since the p-value (0.225) is greater than the alpha level (.05), the null hypothesis which state that there is no difference in the number of items functioning differently in favour of males and those functioning differently in favour of females is not rejected.

Discussion of findings

Results show that 17 out of the 50 multiple choice Physics items functioned differentially for male and female students. The finding of this study agrees with that of Adedoyin (2010) [3], who in his study investigated gender biased items in public examinations, and found that out of 16 items that fitted the 3PL item response theory statistical analysis, 5 items were gender biased. The finding also agreed with that of Adebule (2013) [2] that out of the 40 items examined for the first factor program structure in computer science, only seven items representing 17.5% displayed DIF, comparing male and female examinees. The finding is also in agreement with the report of Birjandi and Mohadeseh (2017) [4] that in the general reading comprehension, 7 out of the 13 DIF flagged items favoured females and 6 proved much easier for males.

The result further showed that there was no significant difference in the number of items functioning differentially in favour of males and those in favour of females. Cognitive skills assessed by items seem the most effective factor that produced gender DIF. The result is in agreement with the finding of Ling and Lau (2014) who investigated the gender DIF in multiple choice and open- response science item types for elementary, middle and high school levels and found out that the possible sources of DIF is due to the differences in content category, visual-spatial component and item type dimensions.

The findings did not corroborate with the findings of Adebule (2013) [2] who investigated DIF in a 3-20 item multiple choice Physics test items selected from Ekiti State

Unified Physics examination for 2008/2009 and 2009/2010 academic sessions. The study concluded that the items did not function differentially among the testees on the basis of gender, age, parental qualification and location. The trend of this study also did not agree with the findings of Madu (2012) [9] who investigated differential item functioning (DIF) by gender in Physics examination conducted by West African Examinations Council (WAEC) in 2011 in Nigeria. Using a sample of 1,671 students and Scheuneuman Modified Chi-square Statistics (SSχ2), the results of the analysis indicated that items significantly function differentially by gender for male and female examinees in 39 items and 11 items did not exhibit DIF.

Conclusion

DIF is an issue that must be properly addressed in examinations and tests designed for heterogeneous groups. Through the application of IRT methodology (Area Index Measure), it was clear that there were presence of DIF in the 2017 NABTEB Physics test items. It is obvious that threat in the validity of test items has been created. Such threats could influence or introduce traits irrelevant to the construct of interest. This could jeopardize classification of subgroup of candidates test scores negatively. It was also concluded that Multiple Choice Physics test items administered by NABTEB in 2017 do not show a significant difference in the number of items functioning differentially by gender in favour of males and those in favour of females. Therefore, test developers, ministry of education and examination bodies should ensure that items are free from differential item functioning (DIF).

Recommendations

On the basis of the findings and conclusion, the following recommendations are made:

1. Test experts and developers should consider the use of Area index measure in determining differential item functioning. This approach provides an intuitive and flexible methodology for detecting DIF.
2. Educational measurement and evaluation experts in Nigeria should rise to the challenges placed by the measurement community and be fully aware of the usefulness of IRT in constructing and scoring of tests or examinations.
3. For bias-free items to be produced, the NABTEB examination developers should make certain that activities and connotations reflected in the test are relevant to the life experiences of examinees responding to the items. Test items should be written in a straight forward, uncomplicated, easily read manner. Excessive wordiness can obviously prevent the examinees from responding appropriately to test items and therefore create bias in the examination.
4. Examination bodies should organize training for item developers on the construction of valid, reliable and fair test especially in the area of DIF. In addition, items flagging DIF should be revised, modified or eliminated from the test.

References

1. Abedlazeez N. Exploring DIF: Comparison of CTT and IRT methods. International Journal of Sustainable Development,2010:1(7):11-46.
2. Adebule SO. A study of differential item functioning in Ekiti State unified physics examination for senior secondary schools. Journal of Education and Practice,2013:4(17):43-46.
3. Adedoyin OO. Investigating the invariance of person parameter estimates based on classical test and item response theories. International Journal of Education Science,2010:2(2):107-113.
4. Birjandi P, Mohadeseh A. Differential item functioning (test bias) analysis paradigm across manifest and latent examinee groups on the construct validity of IELTS. Human Sciences,2017:55:153-172.
5. Clauser BE, Mazor KM. Using statistical procedures to identify differentially functioning test items. Educational Measurement Issues and Practice,2018:17:31-44.
6. De Beer M. Use of differential item functioning (DIF) analysis for bias analysis in test construction. South African Journal of Industrial Psychology, 2014:30(4):52-58.
7. Kanjee A. Using logistic regression to detect bias when multiple groups are tested. South African Journal of Psychology,2017:37:47-61.
8. Ling SE, Lau SH. Detecting differential item functioning (DIF) in standardized multiple-choice test: An application of item response theory (IRT) using three parameter logistic model. Journal of Applied Psychology,2013:94(7):452-459.
9. Madu BC. Analysis of gender-related differential item functioning in physics multiple choice items administered by West African Examination Council (WAEC). Journal of Education and Practice,2012:3(8):71-79.
10. Ogbebor U, Onuka A. Differential item functioning method as an item bias indicator. Educational Research,2013:4(4):367-373.
11. Oshima TC, Morris SB. Raju's differential functioning of items and tests (DFIT). The National Council on Measurement and Evaluation Instructional Module, 2018, 43-50.
12. Penfield RD, Camilli G. Differential item functioning and item bias. In C. R. Rao and S. Sinharay (Eds.), Handbook of statistics,2017:(26):125-167.
13. Reynolds RC. Measurement and Assessment in Education. Boston: Pearson, 2016.