DOI: https://dx.doi.org/10.4314/gjedr.v20i1.6

GLOBAL JOURNAL OF EDUCATIONAL RESEARCH VOL 20, 2021: 55-62
COPYRIGHT© BACHUDO SCIENCE CO. LTD PRINTED IN NIGERIA. ISSN 1596-6224
www.globaljournalseries.com; globaljournalseries@gmail.com

TEST FAIRNESS AND ASSESSMENT OF DIFFERENTIAL ITEM FUNCTIONING OF MATHEMATICS ACHIEVEMENT TEST FOR SENIOR SECONDARY STUDENTS IN CROSS RIVER STATE, NIGERIA USING ITEM RESPONSE THEORY.

EFFIOM, ANTHONY PIUS

(Received 26, July 2019 Revision Accepted 9, August 2021)

ABSTRACT

This study used Item Response Theory approach to assess Differential Item Functioning (DIF) and detect item bias in Mathematics Achievement Test (MAT). The MAT was administered to 1,751 SS2 students in public secondary schools in Cross River State. Instrumentation research design was used to develop and validate a 50-item instrument. Data were analysed using the maximum likelihood estimation technique of BILOG-MG V3 software. The result of the study revealed that 6% of the total items exhibited differential item functioning between the male and female students. Based on the analysis, the study observed that there was sex bias on some of the test items in the MAT. DIF analysis attempt at eliminating irrelevant factors and sources of bias from any kind for a test to yield valid results is among the best methods of recent. As such, test developers and policymakers are recommended to take into serious consideration and exercise care in fair test practice by dedicating effort to more unbiased test development and decision making. Examination bodies should adopt the Item Response Theory in educational testing and test developers should therefore be mindful of the test items that can cause bias in response pattern between male and female students or any sub-group of consideration.

KEYWORDS: Assessment, Differential Item Functioning, Validity, Reliability, Test Fairness, Item Bias, Item Response Theory.

INTRODUCTION

Fairness in assessment of students' achievement tests in mathematics in our secondary schools is very fundamental as mathematics is the basis for studying other subjects especially in science related courses. Mathematics is a compulsory subject for every individual to function effectively and efficiently in today's world irrespective of one's profession and hence scores obtained by students in this subject should reflect their true ability (Githua & Mwangi, 2003; Blank, Alas & Smith, 2007; Effiom, 2016; Inameti, 2018).

Fairness is an essential quality of a test; its equitable treatment of all examinees during the testing process, absence of measurement bias, equitable access to the constructs being measured, and justifiable validity of test score interpretation for the intended purpose

Every assessment provides formative and summative data on students' learning and achievements through which specific acquired competencies could be acquired by efficient teaching and learning process. Achievement test in mathematics is generally designed to measure

Effiom, Anthony Pius, Department of Educational Foundations, University of Calabar, Calabar Nigeria.

the students' cognitive task that is, it measures the present proficiency, mastery and understanding of general and specific areas of knowledge of the subject. Therefore, students' performance on an achievement test is an index of his or her mastery of the subject taught by the teacher. The test could be teacher-made tests or test constructed and validated by test experts through the adoption of elaborate procedures and degree of precision. A quality test should be valid, reliable, fair and devoid of item bias. Differential Item Functioning (DIF) is a key component in the evaluation of the fairness and validity of educational and psychological tests.

Testing for differential item functioning is an investigation to know whether performance on any test item differs for certain groups of examinees that is, male and female students. The main idea behind differential item functioning is that if we match two different groups of examinees on a construct of interest, then the probability of endorsing an item should be the same for both groups of examinees. That is, differential item functioning is present when equally able examinees, from different groups, do not have the same probabilities of responding to an item (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). The present study matched males and females on mathematics achievement test, and then the probability of responding correctly to an item should be the same for males and females. However, if we find males with the same mathematics ability as females had a greater probability of responding correctly to an item than females, then the item would be identified as functioning differently across gender. This means the achievement test item is not only measuring mathematics ability, but also measuring a second unrelated factor known as sex.

To prevent inappropriate consequences of interpretation of test scores, bias must be detected and removed. It can be detected through various methods and procedures and one of the most important and currently used procedures of bias detection is the DIF approach.This is a method that investigates the test items, one at a time, for signs of interaction with sample characteristics. Test bias can occur when performance on a test requires sources of knowledge different from those intended to be measured, causing test scores to be less valid for a particular group (Penfield and Lam, 2000; & Adediwura, 2006). Test bias is often examined at the item level, with DIF analyses being part of the framework for probing item bias. If a certain group (that is, sex) performs lower on a specific item, when compared with a reference group (after controlling for the overall differences in their ability scores), then one could say that the item is biased against that particular group. DIF analyses compare the performance of two groups of the same level of ability in order to disentangle the effects of unfairness and ability level. Matching ability level is essential, since different groups may have different ability levels, in which case differences in performance are to be expected. Consistent differences between two groups of the same ability level would suggest that DIF is present. However, results of DIF analyses can only suggest that DIF is present, and not that the items are biased. To consider an item as biased also requires determining the non-target constructs that lead to the between-group differences in performance (Penfield & Lam, 2000). Thus, DIF is a necessary but not sufficient condition for item bias (Clauser & Mazor, 1998). DIF framework has become an integral component of test validation methodology and the study of test fairness. The presence of DIF in a particular item indicates that individuals having the same level of ability, but belonging to different groups, do not share the same expected response to the item (Penfield & Camilli, 2007). The two groups being compared are called the reference and focal groups (historically, the reference group is the group for which the test is expected to favour and focal group is the targeted or disadvantaged group of interest). Of all statistical methods of DIF detection, those based on item response theory, especially the three-parameter IRT model, are regarded as most theoretically sound and common (Lord, 1980; Shepard, Camilli, & Williams, 1985; Adedoyin, 2010). This is because IRT expresses through the item characteristic curve (ICC) the relationship between examinee ability and the probability of answering an item correctly, which is a relationship of particular salience in examinations of the interactions between items and groups. The ICCs for individual items for two groups of examinees should match closely; if they do not, the interpretation is that equally able examinees in the reference and focal groups do not have equal chances of getting the item right, which may be considered a textbook definition of item bias. Because of their theoretical advantages, IRT approaches are widely used where the relatively stringent sample size requirements for applying them could be met. IRT allows for a critical examination of responses to

particular items from a test and major merits of using IRT in DIF detection are:

1). Compared to Classical Test Theory, IRT parameter estimates are not as confounded by sample characteristics.

2). Statistical properties of items can be expressed with greater precision which increases the interpretation accuracy of DIF between groups.

3). Statistical properties of items can be expressed graphically, improving interpretability and understanding of how items function differently between groups.

In relation to DIF, item parameter estimates are computed and graphically examined via item characteristic curves (ICCs) also referred to as trace lines or Item Response Functions (IRF). After the examination of ICCs and subsequent suspicion of DIF, statistical procedures are implemented to test differences between parameter estimates. ICCs represent mathematical functions of the relationship between positioning on the latent trait continuum and the probability of giving a particular response. Figure 1 illustrates this relationship as a logistic function. Individuals lower on the latent trait or with less ability have a lower probability of getting a correct response or endorsing an item, especially as difficulty increases. Thus, those higher on the latent trait or in ability have a greater chance of a correct response or endorsing an item. For instance, individuals with higher mathematics ability have a greater probability of getting mathematics items correct than those with lesser ability. Another critical aspect of ICCs pertains to the inflection point. This is the point on the curve where the probability of a particular response is 0.5 and also represents the maximum value for the slope. This inflection point indicates where the probability of a correct response or endorsing an item becomes greater than 50%, except when a *c* parameter is greater than zero (0) which then places the inflection point at 1 + c/2. The inflection point is determined by the difficulty of the item which corresponds to values on the ability or latent trait continuum. Therefore, for an easy item, this inflection point may be lower on the ability continuum while for a difficult item it may be higher on the same scale.

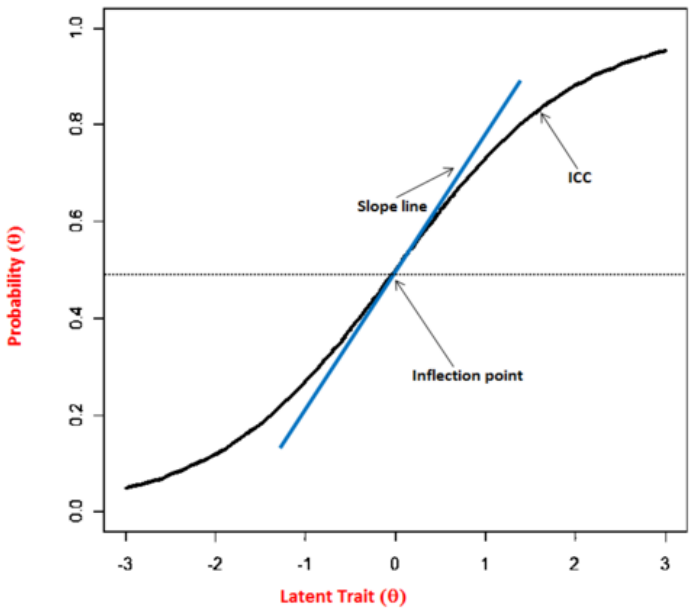

Figure 1. ICC with Inflection point and slope line

In the recent study, it is necessary to provide a general understanding of the different parameter estimation models and their associated parameters. These include the one-, two-, and three-parameter logistic models. These models assume a single underling latent trait or ability that is, they have an item difficulty parameter denoted *b*. For the 1PL and 2PL models, the *b* parameter corresponds to the inflection point on the ability scale. In the case of the 3PL model, the inflection corresponds to 1 + c/2 where *c* is a lower asymptote. Difficulty values, in theory, can range from -∞ to +∞, however in practice they rarely exceed ±3. Higher values are indicative of harder test items. Items exhibiting low *b* parameters are easy test items. Another

parameter that is estimated is a discrimination parameter designated *a*. This parameter pertains to an item's ability to discriminate among individuals. The *a* parameter is estimated in the 2PL and 3PL models. In the case of the 1PL model, this parameter is constrained to be equal between groups. In relation to ICCs, the *a* parameter is the slope of the inflection point. As mentioned earlier, the slope is maximal at the inflection point. The *a* parameter, similar to the *b* parameter, can range from -∞ to +∞; however typical values are less than 2. In this case, higher value indicates greater discrimination between individuals. The 3PL model has an additional parameter referred to as a *guessing* parameter and is denoted by *c*. This corresponds to a lower asymptote which essentially allows for the possibility of an individual to get a moderate or difficult item correct even if they are low in ability. Values for *c* range between 0 and 1.When applying statistical procedures to assess for DIF, the *a* and *b* parameters (discrimination and difficulty) are of particular interest. However, assume a 1PL model was used, where the *a* parameters are constrained to be equal for both groups leaving only the estimation of the *b* parameters. After examining the ICCs, there is an apparent difference in *b* parameters for both groups.

IRT effectively places individuals along the latent trait or ability continuum. Thus, one procedure may indicate DIF for certain items while others do not. Another issue is that sometimes DIF may be indicated but there is no clear reason why DIF exists. This is where reasoned judgment comes into play. It is not enough to report that items function differently for groups; there are needs to give a theoretical reason for why it occurs. Evidence of DIF does not directly translate into unfairness in the test. It is common in DIF studies to identify some items that suggest DIF. This may be an indication of problematic items that need to be revised or omitted and not necessarily an indication of an unfair test. Therefore, DIF analysis can be considered a useful tool for item analysis but is more effective when combined with theoretical reasoning.

Respectively, students are made to take examinations administered by their teachers without detecting the fairness and appropriateness of the examination items with regard to the different groups (e.g., male and female) of examinees. One may ask: are these test items fair enough for all groups? how will one know that such examination or test items are not fair? It is obvious, therefore, that most examination being administered to students may not be fair to one group or the other if methods that will refine test items devoid of gender biases are not taken into consideration, especially differential item functioning (DIF) methods. It is at this instance that the investigator consciously embarked on the recent study to assess the differential item functioning in mathematics achievement test for male and female Senior Secondary Students in Cross River State using Item Response Theory.

RESEARCH QUESTION

What is the Differential Item Functioning of the 50 items instrument on the male and female Senior Secondary Students of Cross River State?

METHOD

This study adopted instrumentation research design to develop and validate a 50-item instrument using the Item Response Theory. The Instrument was titled Mathematics Achievement Test (MAT). A pilot test was administered on 50 SS2 (25 males & 25 females) students randomly chosen from five public schools. The responses obtained were dichotomously scored (0 or 1), and KR_{20} reliability index of 0.92 was obtained. The final version of the 50-item Mathematics Achievement Test was administered to a sample of 1,751 SS2 students (883 males & 868 females) in 30 public secondary schools in Cross River State through stratified sampling technique. The data generated were analyzed using the maximum likelihood estimation technique of BILOG-MG V3 of a Three-Parameter Logistic Model of IRT to assess and detect Differential Item Functioning of the male and female students.

RESULTS

In the recent study, Table 1 shows the adjusted threshold values for group differential item functioning of the test items of the Multiple-Choice Achievement Test in Mathematics. From the data, the result indicated that differential item functioning effects were not observed on 47 items representing 94% of the total items namely: items 1 to 24, 26 to 39 and 42 to 50. Three items representing 6% of the total items namely: items 25, 40 and 41 were identified as exhibiting differential functioning among male and female students.

Table 1
Model for Group Differential Item Functioning of the Test Items of the Mathematics Achievement Test

Item	Group	P	Chi-square
1	Male	0.00	25.3*
	Female	0.00	14.0*
2	Male	0.21	16.0*
	Female	0.00	12.1*
3	Male	0.05	2.9*
	Female	0.11	3.8*
4	Male	0.42	14.0*
	Female	0.41	11.1*
5	Male	0.41	19.6*
	Female	0.38	28.6*
6	Male	0.00	37.1*
	Female	0.05	46.0*
7	Male	0.21	73.1*
	Female	0.40	59.2*
8	Male	0.00	12.0*
	Female	0.00	18.4*
9	Male	0.00	26.0*
	Female	0.00	17.1*
10	Male	0.32	27.1*
	Female	0.14	31.1*
11	Male	0.07	18.0*
	Female	0.19	21.1*
12	Male	0.20	47.0*
	Female	0.18	39.2*
13	Male	0.07	80.1*
	Female	0.12	14.2*
14	Male	0.50	20.7*
	Female	0.43	27.1*
15	Male	0.00	44.2*
	Female	0.00	28.0*
16	Male	0.05	8.1*
	Female	0.08	11.0*
17	Male	0.00	19.3*
	Female	0.00	28.0*
18	Male	0.00	74.0*
	Female	0.00	63.1*
19	Male	0.26	22.1*
	Female	0.00	11.1*
20	Male	0.00	4.31*
	Female	0.21	10.3
21	Male	0.42	13.8*
	Female	0.41	12.8*
22	Male	0.11	28.3*
	Female	0.09	84.0*
23	Male	0.04	74.1*
	Female	0.03	23.0*
24	Male	0.00	11.7*
	Female	0.12	25.1*
25	Male	0.25	16.2
	Female	0.23	16.2
26	Male	0.02	16.2*
	Female	0.02	19.1*
27	Male	0.24	28.0*
	Female	0.21	13.5*
28	Male	0.00	52.1*
	Female	0.00	65.0*
29	Male	0.21	80.3*
	Female	0.18	62.1*
30	Male	0.16	77.2*
	Female	0.15	57.1*
31	Male	0.00	95.8*
	Female	0.02	49.1*
32	Male	0.07	47.0*
	Female	0.06	70.1*
33	Male	0.00	41.2*
	Female	0.00	53.3*
34	Male	0.41	68.4*
	Female	0.38	34.2*
35	Male	0.00	19.0*
	Female	0.00	26.0*
36	Male	0.00	47.2*
	Female	0.00	18.1*
37	Male	0.13	92.3*
	Female	0.11	73.3*
38	Male	0.14	84.4*
	Female	0.06	91.2*
39	Male	0.06	58.1*
	Female	0.13	47.4*
40	Male	0.00	7.8
	Female	0.00	7.8
41	Male	0.11	5.3
	Female	0.10	5.3
42	Male	0.00	96.1*
	Female	0.00	27.1*
43	Male	0.09	83.1*
	Female	0.00	79.0*
44	Male	0.21	43.1*
	Female	0.16	50.1*
45	Male	0.05	86.1*
	Female	0.11	99.0*
46	Male	0.24	16.0*
	Female	0.15	27.1*
47	Male	0.13	16.1*
	Female	0.04	11.5*
48	Male	0.17	13.1*
	Female	0.11	18.5*
49	Male	0.14	67.1*
	Female	0.12	83.1*
50	Male	0.00	19.0
	Female	0.00	12.1*

Asterisks indicate test items without DIF on sex

DISCUSSION

The Differential Item Functioning is generally an undesirable characteristic of the test because; the test is measuring the construct it is not designed to measure that is, some other additional characteristics of performance on classification of the group (Penfield & Lam, 2000). DIF analysis helps to create a better understanding of the difficulty of an item and the characteristics of the group participating in the assessment, indicating the group's relevant strengths and weaknesses. Respectively, Bond and Fox (2001) explained that differential item functioning is also known as 'bias' which refers to differential validity of a given interpretation of a test score for any definable, relevant sub-group of test takers.

In the recent study, the data from Table 1 indicated the adjusted threshold values for group differential item functioning of the test items of the Mathematics Achievement Test of senior secondary students. The result showed that three items namely: items 25, 40 and 41 representing 6% of the total items were identified as exhibiting Differential Item Functioning. The three items were bias for male and female students that is, the Chi-Square values for male and female students were same in items: 25, 40 and 41. This finding corroborates Ani (2014) who detected 11 items out of 50 in a Multiple-Choice Economics Achievement Test for exhibiting Differential Item Functioning on gender.

In a related study, Schumacker (2005) suggested a borderline difference between item bias and Differential Item Functioning, establishing that an item flagged DIF may not be a biased item rather, biased items is an indication of differential item functioning. This corroborates Ibrahim (2018) assertion that there are occasions when examinees from different demographic groups may be expected to differ in ability especially, when learning opportunities are not evenly distributed. In these instances, the result is often termed item impact rather than item bias. However, Zumbo (2007) suggested DIF operational policy by using the Cramer's phi coefficient to find levels of significance for large, medium and small size effects.

The recent result also supports the assertion by Pedrajita (2009) who detected bias test items with the logistic regression method. In the study, 22 biased items were identified between the public and the private examinees. Seven (7) items indicated differential item functioning between male and female students. Pedrajita concluded that "the two groups had not had equal opportunity to learning experience related to the content of the biased items'' (p.67). This finding supports his assertion that the focal group might not have had equal opportunity to learning experience related to the content of the biased item or may have been influenced by other factors like language, poor calculation ability, omission and wrong use of units.

Furthermore, the recent result in addition indicated that 47 items namely: items 1 to 24, 26 to 39 and 42 to 50 representing 94% which did not exhibit differential item functioning for the male and female students implies that 47 of the total test items in the test were fair items and therefore should be accepted and others should be reviewed. The finding corroborates earlier finding by Ani (2014) who retained 39 items out of 50 in a Multiple-Choice Achievement Test in Economics because differential item functioning was not detected on the gender of the respondents. It is the opinion of the investigators that the principle of test fairness requires that examinations undergo scrutiny to detect and remove items that that behave in significantly different ways for the different groups writing a test. That is, identifying bias in assessments across dissimilar groups would improve the test items analysis in educational testing in Nigeria.

CONCLUSION

Issues of test fairness and item bias of achievement tests are very significant in our educational system because test scores are used in making most decisions in our schools. DIF raises concern because its presence suggests that examinees from different demographic groups' example, sex have differing probabilities of success on a test item, after they have been matched on the psychological characteristics of interest. As other investigators have observed, the requirement that performance differences exist, even after matching on the ability of interest, is a central idea, implying, for example, that observed differences in test scores are not, in themselves, evidence of test bias. Indeed, there are occasions when examinees from different demographic groups may be expected to differ in ability especially, when learning opportunities are not evenly distributed. In these instances, the result is often termed item impact rather than item bias. DIF analysis attempt at eliminating irrelevant factors and sources of bias from any kind for a test to yield valid results is among the best methods of recent.

RECOMMENDATIONS

Based on the findings, the following recommendations were made:

1. The Differential Item Functioning effect of items in a given instrument for measurement of latent trait should be checked to ensure that sub-group of examinees are fairly treated.
2. That test developers and examination bodies should adopt the Item Response Theory in their effort to develop a fair credible test items, valid and reliable test.

REFERENCES

Adediwura, A. A., 2006. A comparative Study of Two Methods of Detecting Test Item Bias in Senior School Certificate (SSC) Mathematics Examination. Unpublished Ph.D. Thesis, Faculty of Education, Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria.

Adedoyin, O. O., 2010. Using IRT approach to detect gender biased items in public examinations: A case study from the Botswana junior certificate examination in Mathematics. Educational Research and Review, 5(7), 385-399.

Ani, E. N., 2014. Application of item response theory in the development and validation of multiple-choice test in economics for senior secondary school students in Nsukka education zone of Enugu state. Unpublished M.Ed Thesis, University of Nigeria, Nsukka.

Blank, R. K., Alas, N., and Smith, C., 2007. The power of professional development and self-efficacy in the teaching of mathematics. Washington DC: McGraw Hill Education.

Bond, T. G., and Fox, C. M., 2001. Applying the Rasch model (2nd ed.). Mahwah NJ: Lawrence Erlbaum.

Clauser, B. E., and Mazor, K. M., 1998. Using Statistical Procedures to Identify Differentially Functioning Test Items. Educational Measurement: Issues and Practice, 17(1), 31-44.

Effiom, A. P., 2016. Teachers' Characteristics, School Environment and Mathematics Teachers' Effectiveness in Cross River State, Nigeria. Unpublished M.Ed Thesis. University of Calabar.

Githua, B. N., and Mwangi, J. G., 2003. Students' mathematics self-concept and motivation to learn mathematics relationship and gender differences among Kenya's secondary school students in Nairobi and Rift Valley Provinces. International Journal of Educational Development, 2(23), 487-499.

Ibrahim, A., 2018. Differential Item Functioning: The state of the art Jigawa. Journal of Multidisciplinary Studies (JJMS), 1(1), 37-50.

Inameti, P. U., 2018. Evaluation of Instructional Effectiveness of Mathematics Teacher in Calabar Education Zone, Cross River State. Unpublished M.Ed Thesis. University of Calabar.

Pedrajita, J. Q., 2009. Using Logistic Regression to Detect Biased Test Items. The Internationals Journal of Educational and Psychological Assessment 2(5), 54-73.

Penfield, R. D. and Camilli, G., 2007. Differential Item Functioning and Item Bias. In C. R. Rao & S. Sinharay (Eds.), Handbook of statistics: Elsevier.

Penfield, R. D., and Lam, T. C. M., 2000. Assessing Differential Item Functioning in Performance Assessment: Review and recommendations. Educational Measurement, Issues and Practices, 19(3), 5-15.

Shepard, L. A., Camilli, G., and Williams, D. M., 1985. Validity of Approximation Techniques for Detecting Item Bias. Journal of Educational Measurement, 22, 77–105.

Schumacker, R. E., 2005. Test bias and differential item functioning. Retrieved October 11, 2010, from http://www.appliedmeasurementassociates.com,pdf.

Wiersma, W. and Jurs, S. G., 1990. Educational measurement and testing (2nded.). Massachusetts: Allyn and Bacon

Zumbo, B. D., 2007. A Handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (Ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation.