World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Differential Item Functioning (DIF) of senior school SS II students' mathematics achievement test in ABI local government area cross river state Nigeria

Charles Egbonyi Igiri [1], Ele Edward Bassey [2], Okoli Mark I [1] and Okena Ndum Dominc [1,*]

[1] *Educational foundations, University of Education and Entrepreneurship, Akamkpa.*
[2] *Department of Biology UNICROSS Calabar.*

World Journal of Advanced Research and Reviews, 2025, 27(03), 807–815

Publication history: Received on 13 July 2025; revised on 23 August 2025; accepted on 25 August 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.3.3007

Abstract

This study verified the differential item functioning of secondary school student's achievement in mathematics. One research question and one null hypothesis guided the study. The SS2 students in Abi Local Government Area. Secondary Schools for 2018/2020 session constituted the population. A sample of 70 students selected through stratified sampling technique was used for the study. A 25 items standardized instrument captioned mathematics Achievement Test (MAT) based on SS II mathematics curriculum was used. The research questions was answered using 1RT statistic called DIF contrast, while the hypothesis were tested at .05 levels of significance using Student t-test statistics. All these statistics were generated by the IRT software WINSTEPS 3.92.1. The results of the analysis indicated that male and female examinees function differential in 15 items and no difference in 10 items. The results also revealed that four out of the twenty-five items were significantly biased against gender. On the basis of the analysis, it becomes necessary that the examining bodies should set and administer items that are fair to enhance appropriate interpretation of students' results

Keywords: Differential item functioning; Mathematics achievement test; Secondary school; Abi Local Government

1. Introduction

Mathematics has always been an important subject area for learning and research. It is so vital that the government of the federation of Nigeria made it a compulsory subject offered in both primary and secondary school (NPE 2004). This implies that, Mathematics test is compulsory for all the students in the secondary school and must be developed to provide fair and accurate estimate of the ability of all test takers in the population of the test.

The issue of achieving fairness in educational tests is "the most highly charged issue surrounding testing" (Hambleton, Swaminathan, and Rogers, 1991). The National Policy on Education (2004) has stated that the national examination tests should be as valid as possible and as fair as possible to all students. This statement can also be related to the ambition that the education in the senior secondary school must be equal for all students (NPE, 2004). Willingham & Cole (1997) defined a fair test as a test that is comparably valid for all individuals and groups. Fair test design should, according to them, to provide examinees comparable opportunity, as far as, possible to demonstrate knowledge and skills they have acquired that are relevant to the purpose of the test. An education test that has many items function differently against any subgroups is unfair.

Historically, concerns about test bias have centred around differential performance by groups based on gender or race. If the average test scores for such groups (men and women, Blacks and Whites) were found to be different, then the question arose as to whether the difference reflected bias in the test. Given that a test comprises items, questions soon

* Corresponding author: Okena Ndum Dominc

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

emerged about which specific items might be the source of such bias (Zumbo, 2007). For instance, Mathematics test has boys answering correctly most often than girls of equal ability because the subject of the item is on a topic that is familiar to boys (like sports). Thus, when such an occurrence is found and there is a significant difference in the way items are answered by two or more distinct groups, then differential item functioning is a constant concern.

Differential Item Functioning (DIF) also referred to as measurement bias, occurs when people from different group (like gender or ethnicity) with the same latent traits (abi 1 ity/skill) have a different probability of giving a certain response on a test or questionnaire. An item does not displays DIF if people from different groups have a different probability to give a certain response, it displays DIF if and only if people from different group underlying true ability have a different probability of giving a certain response (Wikipedia.org/wiki). Thus, DIF means one group of examinees performing better another group of examinees on an item when both groups are similar on the trait that is being assessed.

2. Theoretical Framework

The theories related to this work are:

- Classical Test Theory (CTT) and
- Item Response Theory (IRT)

2.1. Classical Test Theory by Charles spearman (1904)

Classical test theory (CTT) was propounded a century ago anchored on the foundation laid by Charles Spearman in his paper of 1904 in which he assumes that each testee has a *true score*, *T,* that would be obtained if there were no errors in measurement. A person's true score is defined as the expected number-correct score over an infinite number of independent administrations of the test. Unfortunately, test users hardly obtain a person's true score, but only an *observed score*, *X.* It is therefore assumed that *observed score is equal to true score* plus some *error*

X = T + E

(Observed score = True score + Error)

This simply indicates that achievement tests and other psychological tests are not error-free. Classical test theory is concerned with the relations between the three variables*X, T,* and *E* in the population. These relations are used to say something about the quality of test scores. In this regard, the most important concept is that of *reliability.*

This theory relates to this work as it helps the researcher to minimize or eliminate errors so that observed scores will be the approximate true abilities. This will in turn improve the reliability of the test.

2.2. Item Response Theory by F. Lord (1953)

In psychometrics, item response theory (IRT) **also known as** latent trait theory, strong true score theory, **or** modern mental test theory, **is a paradigm for the design, analysis, and** scoring of tests, questionnaires, and similar instruments measuring abilities, attitudes, or other variables. The name *item response theory* is due to the focus of the theory on the item, as opposed to the test-level focus of classical test theory. IRT is based on the idea that the *probability of a correct/keyed response to an item is a mathematical function of person and item parameters.* The general aim of item response theory is to understand and improve the reliability of psychological test.

Item Response Theory (IRT) is relevant to this study as it helps the researcher improve in test scoring and developing better test items.

3. Literature review

3.1. Differential Item Functioning

Differential Item Functioning (DIF) is present when people from different groups with the same ability have systematically different responses to specific test items. DIF does not mean simply that an item is harder for candidates in one group than for another group. This could mean:

- One group is performing at its usual "attitude/ability" level on the item, the other is performing better than usual.
- One group is performing at its usual "attitude/ability" level on the item, the other is performing worse than usual.
- The item has its usual difficulty for one group, but is more difficult than usual for the other.
- The item has its usual difficulty for one group, but is easier than usual for the other.

Therefore, If candidates in one group tend to be more capable than candidates in the other

group, they tend to perform better on all the test items, hence a test is never better than the sum of its items. DIF analysis is typically used to identify test items that are differentially difficult for examinees who have the same level of knowledge, skill, or ability but differ in the ways that should be irrelevant to their performance on the test. That is, when a group of examinees score higher than another group on the same item. "The item with the largest DIF is the one with the real (as opposed to artificial) DIF" (Andrich & Hagquist, 2012). However Luppescu (1993) points out situations in which this is not true.

DIF may be attributed to item impact or item bias. Item impact can be described as any group disparity in item performance that reflects actual knowledge and experience difference on the construct of interest (Clauser & Mazor, 1998). Alternatively item bias is defined as invalidity or system error in how a test item measures a construct for members of a particular group (Braimoh 2011; Camili and Adebule, 2009). Item bias can occur when a characteristic of the item that is not relevant to the test purpose differently influences responses of exmainee groups (Ercikan & Lyong-Thomas, 2013). There is an expectation that if an item on a test is not biased, then examinees from two groups who have equal overall ability ought to have the same probability of correctly responding to it. When examinees from different groups that have comparable ability levels have different probabilities of getting on item correct, mfferential item functioning (DIF) is said to occur (Hambleton, Swaminathon & Rogers, HI). Thus, an item is said to be bias if it functions differently for subgroups of test takers of maal ability.

3.2. Differential Item Functioning and Student achievement

Differential item functioning and gender conducted nationally and internationally is a major concern on large-scale achievement tests in mathematics, with it differences between males and females are often found (e.g., Bielinski & Davison, 2001; Boughton, Gierl, & Khaliq, 2000).

In Nigeria, gender-achievement studies include Abiam and Odok (2006) who found no significant relationship between gender and achievement in number and numeration, algbraic processes and statistics. They however found the existence of a weak significant relationship in Geometry and Trigonometry.

Nworgu, (2011), revealed that current national and regional examination is functioning differently with respect to different subgroups. This means that students' score in such examinations are determined largely by the groups to which an examinee belongs and not by ability.

Madu (2012), carries out a study on differential item functioning study was to investigate which items show differential item functionary female students in mathematics examination conducted by West Africans (WAEC) in 2011 in Nigeria. The study was carried out in Nsukka Local v. using the responses of secondary school students who sat for June/July 2009 ex Mathematics conducted by WAEC. Data were obtained from responses of 1671 students multiple-choice test items. The students (examinees) were obtained from 12 senior secondary schools randomly sampled from 20 coeducation schools. DIF was investigated using Scheuneuman Modified Chi square Statistics (SS%2). The results of the analysis indicated that male and female examinees function differential in 39 items and no difference in 11 items. On the basis of the analysis, it becomes necessary that the examining bodies such as WAEC should set and administer items that are fair so that quality education in terms of certification is assured.

Adebule (2013), designed a study to find out if differentially functioning items were used in Ekiti State Unified Mathematics Examination (ESUME) and also to confirm if the test items function in different ways for different groups of test takers. A sample of 400 students selected using the stratified and combined sampling techniques was involved in the study. A 3-20 item multiple choice objective mathematics test items selected from Ekiti State Unified Mathematics Examination for 2008/2009 and 2009/2010 academic sessions were used as instrument of data collection. One research question was raised and one research hypothesis was generated and tested at 0.05 level of significance. The results show closeness in the means and standard deviations of the scores of the groups of testees indicating that the testees are of comparable ability levels. It can be concluded that the items of ESUME did not function differentially

among the testees on the basis of gender, age, parental qualifications and location. It is recommended that differential item functioning procedure should be carried out on all items of the various subject examinations by experts, examination bodies and Ministry of Education.

3.3. Purpose of the study

The purpose of the study was to verify differential item functioning of senior secondary students' Mathematics achievement test in Abi Local Government Area of Cross River State Nigeria. Specifically the study investigate how the items function among groups of students based on gender,

3.4. Research Questions

How do the items in the instrument function among group of students based on gender? Hypothesis

The differential item functioning (DIF) contrast between male and female students measured on the items is not statistically significant.

3.5. Significance of the Study

The findings of the study may be of immense significance to the examination bodies, professional evaluators, teachers, school administrators, parents, counsellors and the general readers. The study presents simple methods for detecting biased items in an achievement test which could be used in overcoming the problem bias in test items.

It may also be of great help to examiners, as well as professionals on the field of testing in the formation of item pool in education based on the fact that from the various test items developed, good test items could be identified and banked for future use.

Immeasurably, the study may assist test developers to adopt new methods of test construction that will encourage the use of Differential Item Functioning (DIF) in measuring examinees' ability in Mathematics.

3.6. Research Design

The study used the survey design. The design is chosen over other research designs because it involves the collection of data at current status to describe "what is" without deliberate effort to control/manipulate variables.

This study was carried out in Abi L.G.A of Cross River State, Nigeria.

3.7. Population of the Study

The population of this study comprised of all the 612 senior Secondary School II (SS2) students from the 17 public secondary schools in Abi Local Government Area of Cross River State for the 2019/2020.

3.8. Sample and Sampling Technique

In selecting the sampled schools, stratified sampling technique was adopted in selecting 7 of the public schools in the local Government area. The stratified random sampling was also used to select a total of 70 students, 5 females and 5 males (10 students) each from the 7 sampled school. The names of students in each class register were written down on pieces of paper and put in a basket. The required number of students were picked randomly from the container. Each time the piece of paper was returned in the basket before the next student was selected.

3.9. Instrumentation

A twenty-five (25) item multiple choice mathematics achievement test (MAT) of four options, A to D, was constructed by the researcher based on the prescribed senior secondary two (SS II) syllabus to cover five topics in the basic areas of Algebraic processes, number and numeration, Mensuration, Geometry, Trigonometry and Statistics/probability. Students were expected to encircle the option bearing the answer. The items were set based on the table of specifications in table 1.

Table I Table of specification for MAT

Content	Knowledge 30%	Comprehension 20%	Application 20%	Thinking 30%	Total
Number/Numeration 20%	2	1	1	2	5
Algebraic process 20%	2	1	1	2	5
Geometry 30%	1	1	1	1	5
Trigonometry 10%					
Statistic/probability 20%	1	1	1	1	5
Total	8	5	5	7	25

The instrument administered to the students through their mathematics class teachers were collected at the expiration of the time for the test by the mathematics teachers who in turn handed them over to the researcher. The scripts were electronically and dichotomously scored. These scores were grouped into female (reference) and male (focal) and into five (5) topics/units labeled A, B, C, D and E for Algebraic process, Number/Numeration, Mensuration, Geometry and Statistics/probability respectively. The reference examinees serves as standard for comparison while the focal groups are the examinees that viewed as being disadvantaged on the test. The attitude surveys and rating data were analyzed using WINSTEP 3.92.1 VERSION. WINSTEPS is Windows-based software which assists with many applications of the Rasch model, particularly in the areas of educational testing analysis.

4. Results

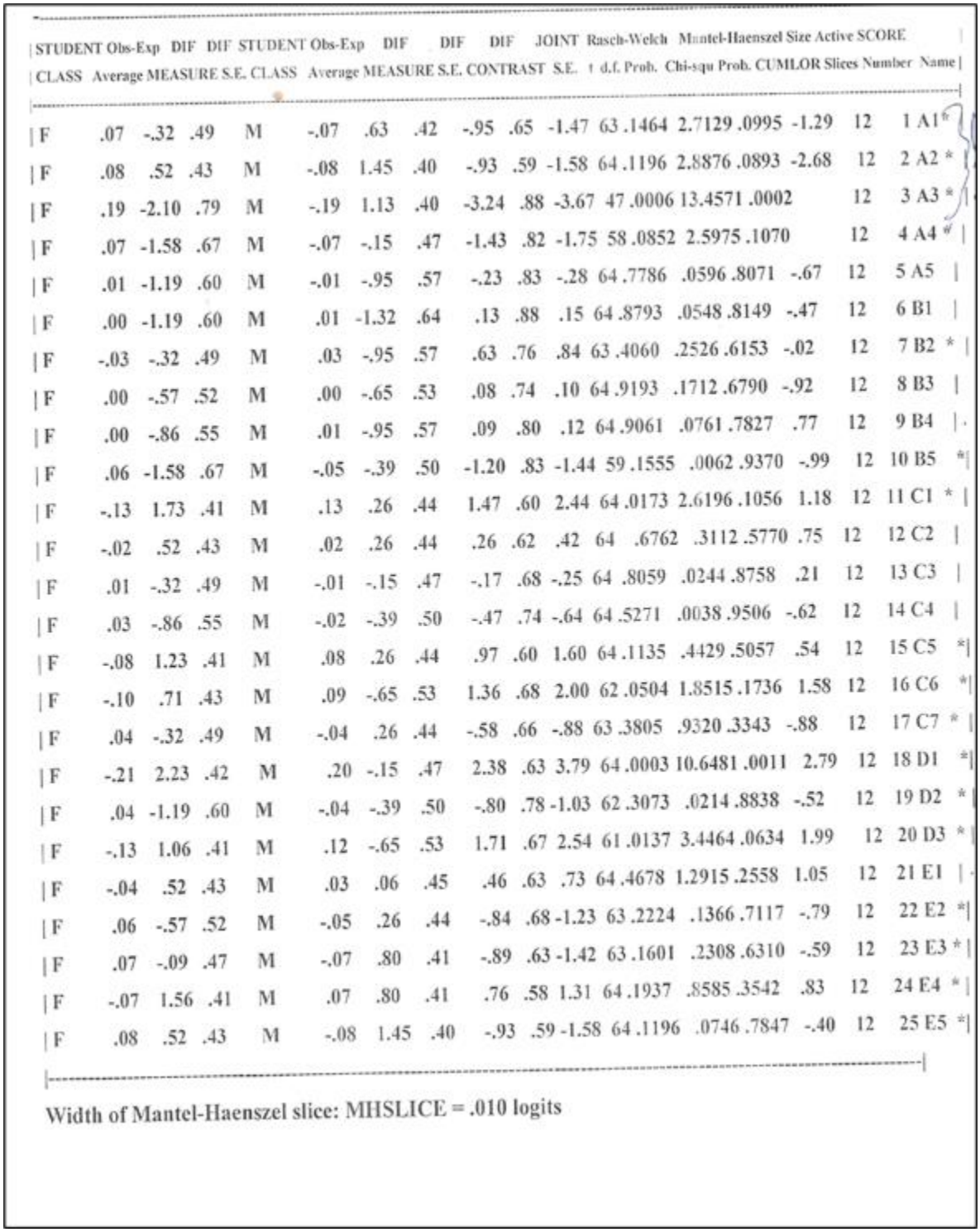
The results of the study are reported according to the research question and the Hypothesis.

4.1. Research Question

How do the items in the instrument function among group of students based on gender? This question was answered using the IRT statistic called DIF. Table 2, figure 1&2 showed the items in relation to Gender (male and female), identified by differential item functioning (DIF) statistical analysis using Winstep version 3.92.1. The analysis showed the values of DIF contrast for the 25items ranging from -3.24 to 2.38. The maximum contrast is 2.38 logits; while the minimum DIF contrast is -3.24. The items with the maximum and minimum DIF contrast values are 18 and 3 respectively. Using the female as reference point; the female students do better in item with positive DIF contrast values than male. The reverse is also correct. Female students are better in items 6, 7, 8, 9, 11, 12, 15, 16, 18, 20, 21 and 24. While male students from are better in items 1, 2, 3, 4, 5, 10, 13, 14, 17, 19, 22, 23 and 25. Analysis flagged item Al, A2, A3, A4, B2, B5, Cl, C5, C6, C7, Dl, D2, D3, E2, E3, E4 and E5 (item number: 1,2, 3, 4, 7, 10, 11, 15, 16, 17, 18, 19, 20, 22, 23, 24 and 25) with DIF contrast not within the -expected range for DIF contrast value of < -0.5 or > 0.5 logits (*item bold in Table.

Figure 1 shows a graphical representation of how each item functioned across the two groups. Two different lines can be seen on the graph. The blue line represents the female group, the red line represents the male group. Clearly seen from the graph is that some items are more difficult for male students (red line) and easier for female students (blue line) and vice versa. The graph revealed gap or distance between both lines (Blue line and Red line). This shows that there is significant difference between male and female students. Thus, it is safe to conclude that the differential item functioning (DIF) contrast between male and female students measured on the items is statistically significant.

Table 2 DIF class specification is DIF- €SS1W



STUDENT Obs-Exp DIF DIF STUDENT Obs-Exp DIF DIF DIF JOINT Rasch-Welch Mantel-Haenszel Size Active SCORE																			
CLASS Average MEASURE S.E. CLASS Average MEASURE S.E. CONTRAST S.E. t d.f. Prob. Chi-squ Prob. CUMLOR Slices Number Name																			
F	.07	-.32	.49	M	-.07	.63	.42	-.95	.65	-1.47	63	.1464	2.7129	.0995	-1.29	12	1 A1	*	
F	.08	.52	.43	M	-.08	1.45	.40	-.93	.59	-1.58	64	.1196	2.8876	.0893	-2.68	12	2 A2	*	
F	.19	-2.10	.79	M	-.19	1.13	.40	-3.24	.88	-3.67	47	.0006	13.4571	.0002		12	3 A3	*	
F	.07	-1.58	.67	M	-.07	-.15	.47	-1.43	.82	-1.75	58	.0852	2.5975	.1070		12	4 A4	*	
F	.01	-1.19	.60	M	-.01	-.95	.57	-.23	.83	-.28	64	.7786	.0596	.8071	-.67	12	5 A5		
F	.00	-1.19	.60	M	.01	-1.32	.64	.13	.88	.15	64	.8793	.0548	.8149	-.47	12	6 B1		
F	-.03	-.32	.49	M	.03	-.95	.57	.63	.76	.84	63	.4060	.2526	.6153	-.02	12	7 B2	*	
F	.00	-.57	.52	M	.00	-.65	.53	.08	.74	.10	64	.9193	.1712	.6790	-.92	12	8 B3		
F	.00	-.86	.55	M	.01	-.95	.57	.09	.80	.12	64	.9061	.0761	.7827	.77	12	9 B4		
F	.06	-1.58	.67	M	-.05	-.39	.50	-1.20	.83	-1.44	59	.1555	.0062	.9370	-.99	12	10 B5	*	
F	-.13	1.73	.41	M	.13	.26	.44	1.47	.60	2.44	64	.0173	2.6196	.1056	1.18	12	11 C1	*	
F	-.02	.52	.43	M	.02	.26	.44	.26	.62	.42	64	.6762	.3112	.5770	.75	12	12 C2		
F	.01	-.32	.49	M	-.01	-.15	.47	-.17	.68	-.25	64	.8059	.0244	.8758	.21	12	13 C3		
F	.03	-.86	.55	M	-.02	-.39	.50	-.47	.74	-.64	64	.5271	.0038	.9506	-.62	12	14 C4		
F	-.08	1.23	.41	M	.08	.26	.44	.97	.60	1.60	64	.1135	.4429	.5057	.54	12	15 C5	*	
F	-.10	.71	.43	M	.09	-.65	.53	1.36	.68	2.00	62	.0504	1.8515	.1736	1.58	12	16 C6	*	
F	.04	-.32	.49	M	-.04	.26	.44	-.58	.66	-.88	63	.3805	.9320	.3343	-.88	12	17 C7	*	
F	-.21	2.23	.42	M	.20	-.15	.47	2.38	.63	3.79	64	.0003	10.6481	.0011	2.79	12	18 D1	*	
F	.04	-1.19	.60	M	-.04	-.39	.50	-.80	.78	-1.03	62	.3073	.0214	.8838	-.52	12	19 D2	*	
F	-.13	1.06	.41	M	.12	-.65	.53	1.71	.67	2.54	61	.0137	3.4464	.0634	1.99	12	20 D3	*	
F	-.04	.52	.43	M	.03	.06	.45	.46	.63	.73	64	.4678	1.2915	.2558	1.05	12	21 E1		
F	.06	-.57	.52	M	-.05	.26	.44	-.84	.68	-1.23	63	.2224	.1366	.7117	-.79	12	22 E2	*	
F	.07	-.09	.47	M	-.07	.80	.41	-.89	.63	-1.42	63	.1601	.2308	.6310	-.59	12	23 E3	*	
F	-.07	1.56	.41	M	.07	.80	.41	.76	.58	1.31	64	.1937	.8585	.3542	.83	12	24 E4	*	
F	.08	.52	.43	M	-.08	1.45	.40	-.93	.59	-1.58	64	.1196	.0746	.7847	-.40	12	25 E5	*	

Width of Mantel-Haenszel slice: MHSLICE = .010 logits

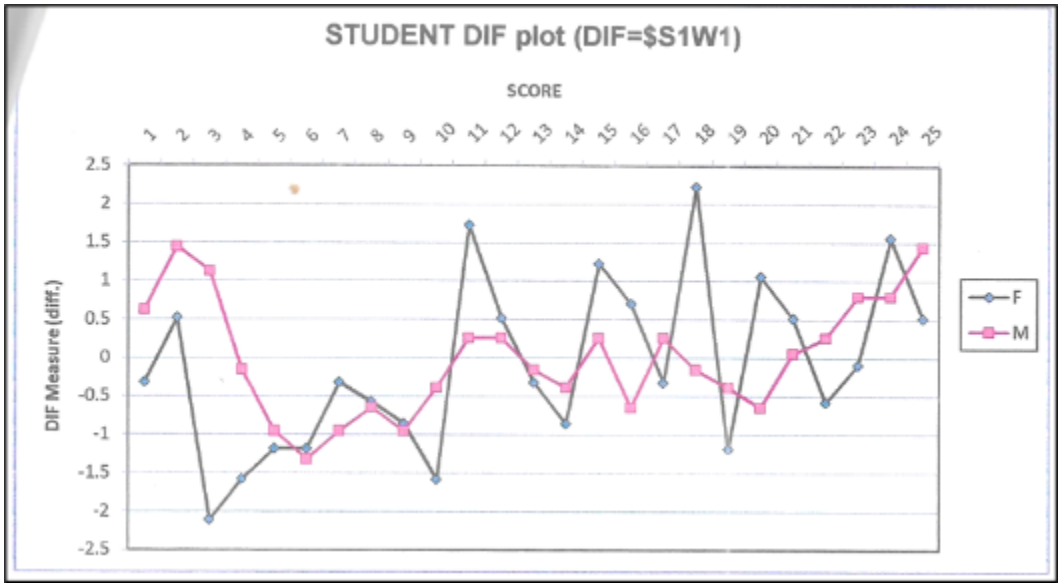


Figure 1 Student DIF mesure Plot

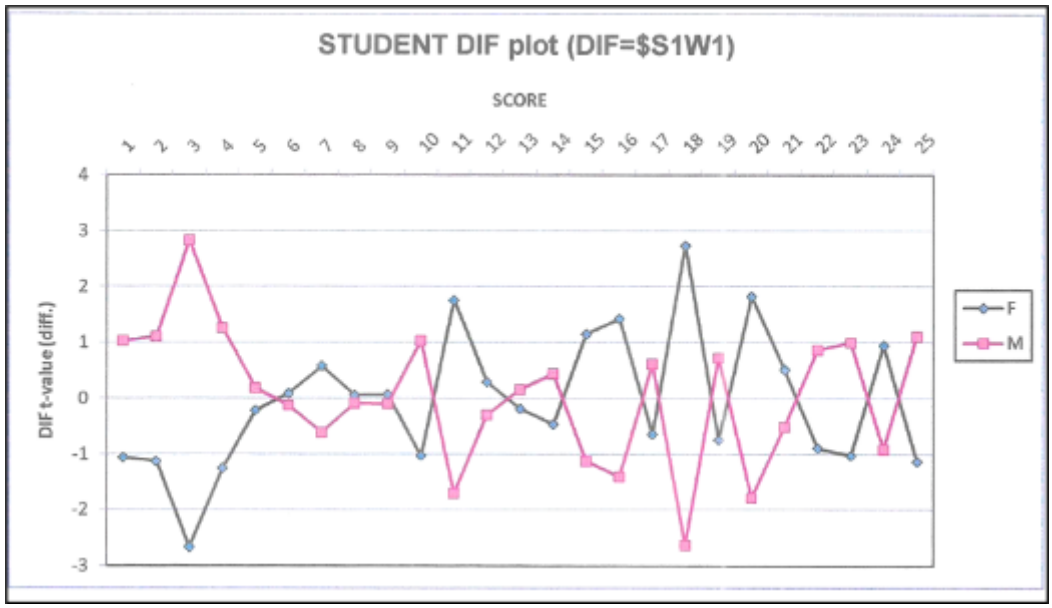


Figure 2 Student DIF t value plot

5. Conclusion and Recommendations

The findings of the study expanded the work of previous researchers in the area of Differential Item Functioning (DIF). This investigation revealed that some items function well with one group, but otherwise with the other. This showed that DIF exists, in the mathematics achievement test (MAI). However, this study did not show that the type of items used in the test is giving extra advantage to one particular sex. Although the findings of this study are reliable, they may not be overgeneralised without further studies. Items exhibiting high and significant DIF indices should be reviewed by content specialists before a decision to either use or discard is made. Future studies are needed to understand why boys and girls perform differently on DIF items, especially when the explanation is not apparent from inspecting the content of an item.

Findings carry implications for both test developers and educators. Test developers must be sensitive to the occurrences of DIF and observe the types of items showing DIF in all subjects tested in examinations. Information on how items 'behave' towards different groups of students can help test developers to enhance test specifications, so that the test is not going to be too lop-sided in terms of design. Test developers who are aware of DIF would be able to control, to a

certain extent, the proportion of item types in a particular test which will be best for the groups taking the test. With DIF analyses results and much experience, it is not impossible that a well-informed test item developer or a trained item writer would be able to anticipate how an item would perform when administered.

Test development work will need to take into account gender differences in test items if equivalent and fair tests are desired. The use of this instrument can be extended to investigate other factors such as ethnic groups, socioeconomic status or other types of schools that may contribute to DIF. DIF analysis can be applied to tests of other subjects. Researchers also recommend that DIF analysis is included in the test construction process in any institution responsible for developing tests and examinations. Educators can use information from DIF analyses to identify the strengths and weaknesses of their students so that more meaningful teaching and learning activities can be planned. DIF analyses provide important quantitative information to the study of fairness in a test item, aimed to reduce, not to totally eliminate unfairness in a test. It is directly relevant to questions of differences in the performance of subgroups of examinees. Although it is undeniably difficult to construct a perfect test that is well balanced and fair to every single group taking a test, DIF analysis is still a critical aspect to consider. If certain items show DIF and judged to be unfair or biased, removing them from the measurement instruments will enhance test validity. If DIF is not conducted, problematic items may not be discovered. An equal proportion of all item types may not be possible after applying DIF in test construction, but the effort would certainly produce the most well thought and fair tests.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Abiarn. P. O. & Odok. J. K. (2006). Factors in Students' achievement in different branches of secondary school Mathematics. *Journal of Education and Technology.* 1(1), 161-168.

[2] AdeteUe. S. O (2013). A study of Differential Item Functioning in Ekiti state unified mathematics Examination for senior secondary school. *Journal of Education and Practice.* Vol. 4. No. 17.

[3] Andrich, D & Hagquist, C (2012). Real and Artificial Differential Item Functioning. *Journal of Educational and Behavioral Statistics, 37, 387-416.*

[4] Bielinski, J., Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple choice mathematics items administered to national probability samples. *Journal of Educational Measurement*, 35(1), 51-77.

[5] Boughton, K. A., Gierl, M. J., & Khaliq, S. N. (2000, May). *Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding differential performance.* Paper presented at the Annual Meeting of the Canadian Society for Studies in Education (CSSE), Edmonton, Alberta, Canada.

[6] Braimoh, M. (20 11). A study of differential item functioning in Mathematics examination on selected secondary school students in Edo State. An unpublished M.Ed. proposal, University of Ado-Ekiti.

[7] Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement,* 77(1), 31-44.

[8] Ercikan, K., & Lyons-Thomas, J. (2013). Adopting Tests for Use in other Languages and Cultures. In K. Geisinger (Ed) *APA Headbooks of Testing and Assessment in Psychology*, Vol. 3. Washington DC: American Psychological Association.

[9] Federal Republic of Nigeria (2004). *National Policy on Education.* Abuja: Nigerian Educational Research and Development Council.

[10] Hambleton, R. K., Swaminathan, FI., & Roger, J. J. (1991). *Fundamentals of Item Response Theory* Clifornia: Sage Publications.

[11] *Linace, J.M.* A user's guide to Winsteps, *2010.* http://www.winsteps.com/winman/index.htm?guide.htm.

[12] Lord, F. M. (1953). The relation of test score to the trait underlying the test. Educational and *Psychological Measurement, 13, 517-548.*

[13] Luppescu, S. (1993) *DIF detection examined.* Rasch Measurement Transactions, 7:2 p.285-6

[14] Madu, B. C. (2012). Analysis of gender-related differential item functioning in Mathematics multiple choice items administered by WAEC. *Journal of Education and Practice. Vol. 3: 222-228.*

[15] Nworgu B. G (2011). Differential item functioning: A critical issue in regional quality assurance. Paper presented in NAERA conference.

[16] Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology 15,* 201-293.

[17] Willingham, W.W. & Cole, N.S; (1997). *Gender and fair assessment.* New Jersey, U.S.A: Lawrence Erbaum associate

[18] Zumbo B.D., (2007) Three generations of DIF analyses: considering Where it has been, Where it is now, and Where it is going, *Language Assessment Quarterly,* Vol.4, No.2, pp. 223.

[19] Nworgu, B.G. (2011) Differential item functioning: A criteria issue in regional quality assurance paper presented in NAERA Conference V.