Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol 3, No.8, 2012

Analysis of Gender-Related Differential Item Functioning In Mathematics Multiple Choice Items Administered By West African Examination Council (WAEC)

Barnabas C. Madu
Department of Science Education, Faculty of Education, University of Nigeria,Nsukka
Enugu State, South –East, Nigeria.
E-mail: bcmad4owa@yahoo.com barnabas.madu@unn.edu.ng

Abstract

The purpose of the study was to investigate which items show differential item functioning (DIF) for male and female students in mathematics examination conducted by West African Examination Council (WAEC) in 2011 in Nigeria. The study was carried out in Nsukka Local Government Area using the responses of secondary school students who sat for June/July 2009 examination in Mathematics conducted by WAEC. Data were obtained from responses of 1671 students in 50 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 ($SS\chi^2$). The results of the analysis indicated that male and female examinees function differential in 39 items and no difference in ii 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.

Key words: Differential Item Functioning, bias, focal group, reference group Scheuneuman modified chi square statistics,

1. Introduction

In our educational system, the National examination tests are important partly to calibrate grades for certification and to give indications of the quality of education, as well as for admission into higher institutions. The National examination tests are administered in English, Mathematics, Science subjects, commercial subjects and technical subjects. One of the aims of this National tests is to make the grounds for assessment across the country as uniform as possible.

When entering the Senior Secondary school the students are required to choose their subjects for National Examination Council (NECO) examination and West African Senior School Certificate Examination (WASSCE). All the subjects include two core subjects namely English and Mathematics. Every subject is divided into many topics specific to that subject.

In mathematics it is possible to study all the topics. The major topics are Algebra, Geometry and Trigonometry, problem-solving, Number and Numeration and Statistics and Probability. The Mathematics subject is compulsory in our both primary and secondary schools, as such Mathematics test is compulsory for all the students in the secondary schools in both internal and external examinations.

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). A valid test should not consist of biased items. Bias is said to exist when a test or an item cause systematic errors in the measurement (Ramstad, 1996; Schumacker, 2005). For instance, if test scores indicate that males perform better on a certain test or item than females do, when in fact both groups have equal ability, the test or item provides a biased measure. This means that something else than what was intended is measured in biased items.

Many aspects of a test and its use have to be considered when discussing test fairness; the way in which tests are used, the participants and the whole testing process (Willingham & Cole, 1997). Willingham and Cole defined a fair test as a test that is comparably valid for all individuals and groups. Fair test design should, according to them, 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.

There are many studies that focus on differences between male and female in tests (for instance, Wang and Lane, 1996; Willingham and Cole, 1997; Gallagher, De Lisi, HoIst, McGillicuddy-De Lisi, Morely, and Cahalan, 2000). The above studies indicate that males have better spatial ability than females (for instance Geary, 1994). This suggests that males use this spatial ability often than females when solving problems which can give them advantages when solving certain kinds of problems in geometry (Geary 1994). Some studies also indicate that females are better than males in verbal skills (Willingham and Cole, 1997) which can give them advantages in items where communication is important. Females also score relatively higher in tests in mathematics that better match course work (Willingham and Cole, 1997). Males tend to outperform females in geometry and number and numeration and algebraic reasoning abilities. Independent of the interest all the students have studied the same topics in mathematics. The results from national subject tests are comparable for different interests, since the students have all taken the same mathematics topics. However, there are cases of differential in performance of students in some form of examination, test or assessment. In some cases the differential favour one group of examinees. For instance, females may function significantly better than males or vice-versa or examinees from urban schools may function better than examinees from rural schools or vice-versa. Test with differentially functioning items cannot be used as a tool for taking a decision or for certification. This offers an opportunity to study difference between male and female students and also study differences between topics (units).

The purpose of this paper therefore is to study which items that show DIF, for male and female students as well as for different topics of mathematics using Schuneuman modified chi-square method (SSχ^2) which is based on Item Response Theory (IRT). Item response theory the study of test and item scores based on assumption concerning the mathematical relationship between abilities (or other hypothesized trait) and item responses. Modeling the relationships between ability and a set of items provides the basis for numerous practical applications, Most of which have advantages over their classical measurement theory counterparts. IRT provides a framework for evaluating how well assessments work, and how well individual items on assessments work. The most common of IRT is in education where psychometricians use it to achieve tasks such as developing and refining examinations, maintaining banks for examinations and equating for the difficulties of successive versions of examinations, for example, to allow comparison between results of subgroups in a population. The comparison between results of subgroups gives indication of items that are functioning differently for different groups of students. This is regarded as differential item functioning (DIF). Differential item functioning (DIF) is a collection of statistical methods that gives indications of items that are functioning differently for different groups of students. Hamilton, Swaminathan and Rogers (1991) defined DIF as: an item shows DIF if individuals having the same ability but from different groups, do not have probability of getting the item right. But it can also be added, that in order to be able to determine whether an item that shows DIF is biased or not further analysis have to be done (camellia and shepherd, 1994). It is then of interest to determine whether the differences depend on differences of ability of the compared groups (not biases) or on the item measuring something else than intended (biases).

There has been variety of methods proposed for detecting DIF. Some of these methods are Mantel-Hansel (M-H) Procedure, Schenuman's modified chi-squared method, Distractor Analsis method, item characteristic curve and Transform item difficulty method. For this study, Schenuman's modified chi-squared method was used because many researchers and authors have classified this method as a major method of detecting DIF.

2. Concept of Scheuneman Modified Chi-Square Method (SSχ2)

With this method, an item is unbiased if for all persons of equal ability, the probability of a correct response is the same regardless of group membership (Scheunuman 1987). With the method each major comparative group is divided into various groups based on the ability level on the basis of observed total test scores. The P-values for each score group are then computed for comparison using the chi-square (χ^2) statistics. In this procedure each item is separately tested for bias. Ability is measured by the total score in a homogeneous test item that measures only mathematics ability. All candidates in the reference group (males) should have equal probability of correct response with all candidates in the focal group (females). Where such probability is different for an item, it is described as differential functioning. The first step in this method involves grouping the testes into score intervals (Table 1). According to Scheuneman, four or five intervals can be created. The factors that determine number of intervals are: difficulty of items, length of the test and size of the sample. The number of testes that fall within each score interval for the two groups (reference and focal testes) is determined with a total across groups for each interval. The number with item correct (observed frequencies) for the two groups in each interval is determined. This is followed by computation of the proportion correct by dividing observed frequency for each score interval by the total number of testes in each score interval. Expected frequency for each group within a given score interval is obtained by multiplying proportion correct (p) for the score interval by the number in each group (reference or focal) who scored within that range. Having determined the observed and expected frequencies, the chi-square is calculated. The degree of freedom for this procedure is (K-1) (r-1) where k is the number of the groups (male and female groups), and r is the number of score groups formed. This is symbolically represented as

$$\chi^2 = \frac{\sum(Mo-Me)^2}{Me} + \frac{\sum(Fo - Fe)^2}{Fe}$$

Where,
Mo = observed frequency of males with item correct
Me = expected frequency of males with item correct
Fo = Observed frequency of females w[th item correct = expected frequency of females with item correct.

3. Method

The purpose of this study is to analyze gender-related differential item functioning in mathematics multiple choice items administered by WAEC in 2011 in Nsukka Local Government Area of Nigeria. Specifically, this study was to identify if WAEC test items in mathematics multiple choice of WASSCE function differentially in terms of gender based on Item Reponses Theory (IRT).

3.1. *Research Question*

What are the items in the multiple-choice test items administered by WAEC function differentially in terms of gender?

3.2. *Hypothesis*

Items in mathematics multiple - choice test items administered by WAEC do not function differentially between male and female examinees in 2011 WASSCE.

A sample of 12 schools was randomly selected from 20 government co-education senior secondary schools with a population of all Senior Secondary Three (3) students who took West African senior school certificate Examination (WASSCE) in 2011 in Nsukka local government Area. In each of the sampled schools, all the students who wrote the WASSCE were studied. In all there were 825 males and 846 females giving a total of 1671. The data for this study were gathered from responses of candidates in 50 multiple-choice questions set and administered by West African Examination Council (WAEC), for 2011 Senior Secondary School Certificate Examination (SSSCE) in mathematics. Person-by-item response matrix obtained form WAEC office was used to map out the ability groups for each of the subgroups for the analysis of DIF. All the candidates from both the reference (male) and focal (female) groups were grouped into five score intervals with respect to the observed total test scores and gender. The multiple-choice items were scored I for correct option and 0 for wrong option with maximum score of 50 and minimum of 0. This is shown in table 1

4. Results

The results presented below are used to answer the research question and test the hypothesis. Items in the mathematics multiple-choice questions set by WAEC do not significantly function differentially between male and female students at .05 level of significance. The frequency of male and female examinees that got the item right in each of the 50 items was subjected to SS. The results are shown in table 2. The results in table 2 indicate that there are items that differentially function for the male and female examinees at $p<.05$; df. = 4; x2 9.488. The data in table 2 indicates that thirty nine (39) items in the mathematics test (stared) were identified as significantly exhibiting differential item functioning between male and female examinees at .05 level of significant while 11 items do not differential function between male and female examinees. The results also indicate that out of 50 test items set by WAEC male and female examinees perform differently in 39 items and none in 11 items.

5. Discussion

The result of analysis of examinees' response to mathematics multiple choice test items set by WAEC for June/July 2011 examination indicated that the mathematics test contained items with significant gender differential functioning. This means that the test contained items that measured different things for male and female examinees with the same mathematics ability. This result agrees with the findings of Miller, Doo little and Acherman (1980), that there is an incidence of gender differential item functioning in mathematics. Literature also revealed that this tendency is not specific to questions used by WAEC, since other public examinations like NECO contained items with similar characteristics. Similarly study conducted by Abiam (1996) demonstrated that mathematics test in public examination like first school leaving certificate examination (FSLCE) showed evidence of gender differential item functioning. This difference could be as a result of the abstract nature of mathematics which demands for perseverance constant practice and a lot of thinking both critically and analytically from the learners. Unfortunately, most of the secondary school students lack the patience and required time to think properly in solving mathematics problems (lheanyi, 2005).

In another thought, mathematics being a compulsory subject in our secondary schools makes it difficult for our students who are good in Arts and social science subjects to drop it for external examinations. This is likely to introduce differences in the performance of mare and female students in mathematics as some of our student read mathematics not because of interest or its importance in their future career but because it is compulsory. Therefore,. differences in the social background of the two groups are likely to contribute to disparities in the performance in the national examination test such as the one of WAEC.

6. Conclusion and Recommendations

A complete evaluation of test quality must include an evaluation of each question. Therefore, questions should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors and examination items should be fair among examinees from all possible subgroup of the population of the examinees. The result of WAEC examination tests are used as indicator of the quality of education in our country. For this reason, it is necessary that he WAEC examination tests should be set in such a way that all the members of the groups will be in a position to answer the questions. If not, the results can show an incorrect picture of the quality of education vis-à-vis certificate for different groups and can lead to the resources for education being distributed in an unfair manner. Similarly, it is necessary for item writers to develop test items and subject them to pilot study so as to select items that do not function differentially that are free from differential functioning.

Reference:

Abiam, P.O (1990). An Analysis of differential item functioning of 1992 first school leaving certificate examination (FSLCE) in cross River state. In G.A Badnus and P.1. Odor (eds) *Challenges of Managing Educational Assessment in Nigeria JAMB, NABTEB, NBEM.*

Camilla, G; & Shepherd, L.A. (1994). *Methods for identifying biased test items*.
London: Sage Publications Ltd. •

Gallagher, A.M. De lesi, R; Holst, P.C; McGill Cuddy Delsi, A.V, Morley, M. & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75, 165-190*

Geary, D.C. (1994). *Children's mathematical* development.
U.S.A: American Psychological Association

Hamilton, R.K; Swami Nathan, H.& Roger, H.J. (1991). *Fundamentals of Item Response Theory.* U.S.A: SAGE publications, INC

Iheanyi, S.J. (2005). *Effect of Peer assessment or academic achievement and interest of students in* geometry. Unpublished M.Ed project, U.N.N.

Doolittle, A.E & Cleary, T.A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement 24(2), 157-166*

Federal Republic of Nigeria (FRN) (2004). National Policy on Education (4th ed).

Lagos: NERDC Press.

Ramstedt, K. (1996). *Electrical girls and mechanical boys on groups* differences *in* tests *a* method development *and a study of differences between girls and boys in national tests in physics.* Doctoral Dissertation, University of Umed.

Scheuneman, J.D. (1979). A Method of assessing bias in test items. *Journal of Educational Measurement, 16(3); 143-152*

Scheuneman, J.D. (1987). An experimental explorations study of causes of bias in test items. *Journal of Educational Measurement*, 24(2); 97 — 118.

Schumacher, R. (2005). Test bias and differential item functioning. www.appliedmeasurementassocaition.com Retrieved July 2010.

Wang, N; & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. *Applied Measurement in Education, 9(2), 175-199*

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

Acknowledgements
I appreciate the assistance of the principals of the study schools who provided me with information on their candidates scores in West African School Certificate Examination.

B.C. Madu Ph.D
Born in Umuowa, Orlu Imo State South East Nigeria on 20th November 1959.
He holds the following qualifications:
B.Sc Physics Edu. Univ. of Ilorin, Ilorin Nig. 1981
M.Ed Science Edu. Univ. of Jos, Jos Nig. 1986.
Ph.D Science Education (Measurement & Evaluation) 2004
Area of Specialization – Science Education and Measurement & Evaluation

Table 1: *Distribution of the observed test scores by distant groups and by gender*

Group	41-50	31-40	21-30	11-20	1-10	Total
Male	40	130	300	320	35	825
Female	36	90	270	400	50	846
Total	76	220	570	720	85	1671

Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol 3, No.8, 2012

Table 2: *Summary of Analysis of Gender DIF in WAEC 2011 June/July Multiple Choice Mathematics Test items using the Scheuneman Modified Chi-Square Statistics (*

S/N	Group	1-10	11-20	21-30	31-40	41-50	SSX^2
1	Male Female	4 17	162 145	252 198	109 76	41 34	36.64*
2	Male , Female	6 9	66 106	113 133	95 60	49 28	4.49
3	Male Female	9 9	139 138	241 179	104 80	42 32	35.53*
4	Male Female	8 4	103 112	228 127	50 62	37 30	21.36*
5	Male Female	9 7	116 127	188 150	97 74	44 35	13.25*
6	Male Female	6 4	53 56	124 84	42 65	42 28	38.25*
7	Male Female	7 5	46 55	25 20	04 44	00 00	25.73*
8	Male Female	7 5	49 83	131 83	*90* 63	35 31	29.74*
9	Male Female	4 7	101 130	121 95	88 56	45 31	13.99*
10	Male Female	5 10	79 115	193 140	101 69	29 32	28.96*
11 *	Male Female	6 9 9	67 101	170 153	120 72'	41 31	12.81*
12	Male Female	11 28	191 213	216 188	98 68	44? 29	20.72*
13	Male Female	3 1 1	63 45	23 82	52 52	35 25	34.38*
14	Male Female	9 12	84 65	100 49	57 40	34 16	41.84*
15	Male Female	13 18	140 214	176 169	84 75	39 35	12.03*
16	Male Female	3 7	81 96	110 159	97 72	39 31	21.68*
17	Male Female	1 3	54 50	151 120	65 39	29 11	91.41*
18	Male Female	7 5	67 62	79 71	34 §1	15 03	15.46*
19	Male Female	5 9	124 117	219 164	93 74	44 29	34.52*
20	Male Female	4 5	53 ,66	70 79	47 52	39 29	4.83

21	Male Female	11 7	89 129	177 135	39* 77	44 29	53.24*
22	Male Female	2 5	31 65	31 39	37 27	33 27	15.81*
23	Male Female	10 20	82 183	141 172	90 73	38 30	19.30*
24	Male Female	4 8 ;%	87 134	160 168	71 61	sn 34	23.44*
25	Male Female	1 8	50 120	60 85	61 51	34 30	23.12*
26	Male Female	10 7	250 124	250 191	112 81	41 34	39.30*
27	Male Female	4 5	226 97	226 181	112 81	44 29	35.61* '
28	Male Female	3 6	125 126	125 130	87 62	35 32	3.53
29	Male Female	8 3	110 99	110 92	85 65	41 29	14.86*
30	Male Female	10 8	127 107	127 101	92 ,69	41 30	15.30*
31	Male Female	3 2	69 40	69 37	16 15	08 05	20.67*
32	Male Female	3 7	107 126	107 85	79 57	39 31	15.17*
33	Male Female	5 9	101 134	138 76	69 61	31 37	9.02
34	Male Female	4 9	68 80	76 66	61 56	37 31	4.21
35	Male Female	10 5	95 123	109 111	75 48	38 28	7.54
36	Male Female	7 1 U	51 1 75	57 I 59	33 I 24	16 106	4.65
37	Male Female	3 5	42 59	39 37	16 11	07 04	2.26
38	Male Female	3 5	44 51	42 50	65 28	34 30	6.41
39	Male Female	5 5	61 75	129 102	37 23	07 05	13.92*
40	Male Female	0 7	90 115	65 60	26 31	08 03	15.68*
41	Male Female	3 5	85 106	100 98	59 39	35 30	5.07
42	Male Female	6 6	68 86	113 82	60 52	44 29	17.23*
43	Male Female	17 4	147 163	195 182	82 73	41 31	33.27*

44	Male Female	7 5	64 98	81 62	37 12	07 09	19.26*
45	Male Female	6 2	56 53	67 53	27 11	08 08	16.05*
46	Male Female	7 5	121 136	201 177	102 68	44 31	16.15*
47	Male Female	2 6	41 69	50 34	29 25	16 09	9.39
48	Male Female	9 3	70 87	194 116	62 27	20 22	55.28*
49	Male Female	2 7	32 68	61 49	27 16	16 14	17.37*
50	Male Female	8 15	30 195	239 216	115 74	44 24	102.36*