GLOBAL JOURNAL OF EDUCATIONAL RESEARCH VOL. 1, NO. 1, 2002: 23 - 30
COPYRIGHT(C) BACHUDO SCIENCE CO. LTD. PRINTED IN NIGERIA

REGIONAL DIFFFERENTIAL ITEM FUNCTIONING (DIF) IN MATHEMATICS ACHIEVEMENT TEST USING THE MANTEL-HAENSZEL (M-H) STATISTICS

IMO E. UMOINYANG

(Received 19 October 2001; Revision accepted 28 January 2002).

ABSTRACT

It is difficult to consider that any achievement test designed to be used nationally could be absolutely free from Differential Item Functioning (DIF) irrespective of the comparative groups of the test-taking population. This study examines the presence of regional DIF in the 50 West African Examination Council's (WAEC) General Certificate in Education (GCE) Ordinary Level Objective Mathematics test items using the Mantel-Haenszel (M-H) statistics.

I488 (586 northern and 902 southern) Nigerian candidates who took GCE O/L Mathematics in November/December, 1990 served as samples in the study. Results showed that eleven (11) items exhibited significant DIF at .05 level of significance with four (4) and seven (7) items favouring northerners and southerners, respectively. It is recommended based on the results that, DIF analysis be adopted by WAEC as one of their standardization procedures to reduce DIF in their test and more research should wade into the possibility of DIF in other school subjects.

Key words: DIF, Mantel – Haenszel (M-H) Statistics, Regional bias in mathematics achievement test

INTRODUCTION AND STATEMENT OF PROBLEM

The issue of differential item functioning (DIF) in achievements tests, otherwise known as test bias, has been a major issue in educational and psychological measurements (Doolittle, 1987) for some years now. It is however surprising that Nigerian professionals in psychological and educational measurements cannot in a large measure claim to have joined in the race for this all-important scholarly exploit. Reasons abound which make it imperative for them to identify themselves urgently in the task of spotting achievement test item which are unfair or yield different scores/responses from subgroups of the test-taking population with the same abilities in the characteristics being measured by a set of test items. Firstly, the Federal Government of Nigeria has urged that any existing ambiguities and lack of uniformity in educational practices should be removed (Federal Government of Nigeria, 1981). Thus, there is the need for any national examination/tests that measure the knowledge and skills taught in schools to be equally appropriate for all candidates irrespective of sex, socio-economic background, location, school type, state and ethnic group of the Federal Republic of Nigeria.

Secondly, research reports should be available to improve upon standards of examining bodies in the country especially the West African Examination Council (WAEC), because their tests "should represent the best test one could obtain for each group given the over riding constraint of content validity" (Green, 1980, P.11). If a biased test is that which is unfair to identifiable subgroups of the general population in which it is being used (Umoinyang, 1996) then a national examination used in predicting future achievements in academics as well as serving as a yardstick for securing a job in the country should be equally appropriate to all the candidates.

IMO E. UMOINYANG, Institute of Education, University of Calabar, Calabar, Nigeria

Thirdly, it is convicting that a study of this nature may generate strongly based theories on which Nigerian educational policies and practices could be based. If those who develop Senior Secondary School Certificate Syllabuses are aware of the diversity of the test-taking population, test developers should also attempt to construct tests based on a broad sampling of tasks set out in the syllabus that tend not to favour any subgroup of the test taking population. This could not be effectively done until an empirical study is available to identify such items functioning differentially for some subgroups in Nigeria.

Nigeria as a country is made up of number of subgroups. However, the present research shall be based purely on the subgroups created by region-northern and southern Nigerians taking Mathematics at the Senior Secondary School Certificate level to justify the quota system adopted for admission into tertiary institutions and employment.

Review of Related Literature

With the identification of differential item functioning as a thing that invalidate the meaning of test results for some subgroups of the population (Miller, Doolittle, and Ackerman, 1988), several statistical methods for estimating the index were later developed by different researchers Since, then, many reviews, criticisms and recommendations have been developed or proposed in relation to what are currently available in the field (e.g Jensen, 1980; Holland & Thayer, 1986; Ryan, 1991). The existence of DIF in many types of tests have been confirmed but none on a national examination based on regions in Nigeria.

Schmitt (1988) used the item difficulty approach to study bias on the Scholarstic Aptitude Tests (SAT) for Hispanics, in particular, Mexican-Americans and Puerto-Ricans who reside in the continental United States and who speak English as their best language. Two forms of the SAT were studies. In the first form, 278, 166 city whites were used as the reference group, 2,963 Mexican-American and 3,230 Puerto-Rican candidates were respectively used as the focal group. In the second form, the base group were 285,885 whites with 3,456 Mexican-Americans, and 3,384 Puerto-Rican candidates as the focal groups. Schmitt observed no much bias in the SAT-Mathematics test. For the verbal test, Schmitt observed 6 and 12 items out of 85 biased for Mexican-Americans and Puerto-Ricans, respectively. In the second form, 14 and 16 items were flagged biased for Mexican-Americans and Puerto-Rican groups respectively.

Other studies are available to show regional bias using the item difficulty approach. Rogers and Kulick (1987) in particular, with the item difficulty approach reported 14, 11 and 11(out of 85) SAT-verbal items of the three forms of SAT as performing differential in favour of city candidates. For SAT-Mathematics, Rogers and Kulick reported 7, 7 and 4 (out of 60) items on the three forms of SAT respectively as exhibiting the same bias.

Van der Flier (1980) applied the logit model approach to a Word Exclusion and a Word Anologies test administered to 500 Tanzanian and 500 Kenyan students. The statistical method was able to exclude 8 out of the 29 items in the test. Van der Flier, Mellenbergh, Ader and Wijn (1984) applied the iterative logit procedure to the same group of 500 Tanzanian and 500 Kenyan on 29 items used by Van der Flier (1980). The total scores were split into five categories, and used 15 iterations to test the X^2 statistics at 0.05 and 0.01 levels of significance. Without iteration, no item was identified as exhibiting DIF at ,01 level, but only two items were identified as exhibiting DIF when the significant level was .05. With iteration, 15 out of the 29 items were flagged biased. They thus concluded that the iterative procedure was a more effective means of detecting biased items even though it takes a lot of computer time.

With whatever detection method there is evidence that there exist DIF in achievement and other types of tests due to the region of subjects sampled.

In another study, Nenty (1986) used four detection methods to determine cross-cultural validity of the scale 2, Form A Cattell Culture Fair

Intelligence Test (CCFIT) items on three mutually remote and culturally disperate groups. The sample was made up of 600 Americans, 231 Indians, and 803 Nigerians. The result from the four detection methods revealed that 23 out of 46 items were detected as being biased by the Schueneman's chi-squares technique, 28 items were identified by the transformed item difficulty – major axis (TID –45^0), 35 were identified by the 1-parameter item characteristic curve method (Rasch) and 34 were identified by the Cochran's chi-square method (CTX^2). The agreement between the detection methods was shown by the high correlation indices between the detection methods which ranged from .81 to .96. Also their high agreement rations which ranged from .68 to .83 are indications of significant regional bias in the CCFIT.

In the first indigenous study of DIF, Inyang (1991) studies location bias on 522 rural and 512 urban Akwa Ibom and Cross River States examinees at the 1986 Common Entrance Examination in Mathematics. She used three detection methods: the modified Scheuneman chi-square (SSX^2) procedure, transformed item difficulties (TID-45^0) and item discrimination methods. The SSX^2 method identified 13 items out of the 33 multiple choice test as being biased, 5 items were flagged biased by the transformed item difficulty, while nine (9) out of the 33 items in the test were spotted as exhibiting DIF by the discrimination procedure. In Inyang's (1991) results, 5 items already identified by the TID-45^o were identified by the SSX^2. Six out of the nine (9) identified by the SSX^2 method were also identified by the discrimination method. Her finding supports the hypothesis that there is location bias in Mathematics achievement tests.

In the light of the results of the above studies, it is really necessary that region based – DIF studies be carried out on some Nigerian mathematics tests used for certification purposes at the Secondary School Level since no such bold attempt so far, have been made to see whether such test items could discriminate between northern and southern Nigerian candidates. The result of this study shall assist policy makers in taking educational decisions.

METHODOLOGY

DATA SOURCE:

The data for this research were derived from a random sample of the November/December 1990 GCE examination. The sample was drawn from seven (7) states (3 northern and 4 southern) of the 21 Nigeria State structure. It consisted of 467 males and 119 females from Northern Nigeria: and 353 males and 549 females from southern Nigeria giving a total of 1488 subjects. The states were first stratified into North and South before simple random sampling of states and candidates.

THE INSTRUMENT:

The WAEC/GCE mathematics examination Paper 1 used for the analysis have seven (7) content areas: number and numerations, algebraic processes, menstruation, plane geometry, trignometry, statistics, and probability with item distributed as 16, 11, 6, 9, 4, 3 and 1, respectively. The 50 – multiple choice items in the test are expected within 11/2 'hours, to measure mathematical 'reasoning' ability.

DIF INDEX

The index of DIF utilized in this study is the Mantel-Haenszel Delta statistic (M-H D-DIF) (Holland & Thayer,1986). The statistic is based on the odds-ratio procedure. A M-H D-DIF value of zero indicates no difference in differential difficulty. A positive M-H D-DIF value indicates that the item is differentially easier for the focal group, while a negative value indicates that the item is differentially more difficult for the focal group. After comparison with other detection methods, Holland and Thayer (1986) concluded that the M-H statistic is the optimal chi-square statistics because it matches subjects that most precisely provide a powerful test of significance, provide a single summary measure of the magnitude of the departure of the null hypothesis, and has standard error formulas.

It is robust to item context effects (Ryan, 1991). Besides, it is relatively inexpensive and

TABLE I: Cumulative frequency distribution for the two regions (north and south)

Class	Mid-	Frequency		Cumulative Percentage		Percentile Rank	
Interval	Point	North	South	North	South	North	South
45 –47	46	1	3	100	100	100	100
42 – 44	43	3	9	100	100	100	99
39 – 41	40	6	25	99	99	99	97
36 – 38	37	16	42	98	96	97	94
33 – 35	34	12	66	96	91	95	86
30 – 32	31	20	92	94	84	92	79
27 – 29	28	44	105	90	74	90	68
24 – 26	25	33	135	83	62	80	55
21 – 23	22	32	136	77	47	74	40
18 – 20	19	42	104	72	32	68	26
15 – 17	16	87	78	64	21	57	16
12 – 14	13	118	58	49	12	39	9
9 – 11	10	115	33		5	20	4
6 – 8	9	46	12		2	6	1
3 – 5	4	11	4	2	0	1	0

In Table 2, the statistics for the M-H analysis, comparing the two groups are reported.

yet a statistically powerful technique for identifying DIF. It has small sample applicability and is manually computable (Umoinyang, 2000).

DATA ANALYSIS

The means and standard deviations from the total raw score for each comparable groups were computed, as well as the Kuder-Richardson 21 reliability estimates. The M-H statistic was employed to spot items exhibiting regional DIF. Computations were based on 5 –2 x 2 contingency tables for each item as against the 32 different observed distinct number-correct scores observed in the 50-item objective test. This was to ease the manual computation of the DIF indices.

RESULTS:

The means, standard deviation for the two groups were 17.15 and 8.44 respectively for candidates from northern Nigeria, and 24.33 and 7.86 for those from southern Nigeria. The mean raw scores were thus significantly (t = 16.47; P = 0.0001) higher for the southern Nigerian candidates. The Kuder-Richardson 21 reliability coefficients were 0.87, 0.84 and 0.87 for northern, southern and all the candidates, respectively. The cumulative frequency distribution and polygons for the two groups are shown in Table 1 and Figure 1, respectively.

From Table 2, eleven (11) items were flagged as exhibiting significant DIF at .05 level of

TABLE 2 : Summary of Mantel-Haenszel DIF Statistics for regional DIF in WAEC GCE November/December 1990 Mathematics.

Item	M-H Alpha	M-H Delta	Item	M-H Alpha	M-H Delta
1	0.55	1.41	26	1.64	-1.17
2	0.41	2.10*	27	1.09	-.21
3	.49	1.68*	28	.65	1.03
4	1.95	-1.57*	29	1.06	-.09
5	1.52	-.98	30	.86	.35
6	1.02	-.05	31	.97	.06
7	.38	2.29*	32	.80	.53
8	1.25	-.52	33	.48	1.71*
9	47	1.78*	34	1.00	.00
10	1.50	-.95	35	.71	.79
11	.72	.76	36	.69	.88
12	.60	1.20	37	1.72	-1.27
13	1.08	-.17	38	1.16	-.35
14	1.69	-1.23	39	.57	1.32
15	1.65	-1.17	40	1.56	-1.05
16	1.54	-1.02	41	2.55	-2.20*
17	.27	3.03*	42	1.34	-.69
18	.53	1.48	43	.83	.44
19	1.75	-1.32	44	2.72	-2.35*
20	.56	1.38	45	1.39	-.77
21	1.18	-.39	46	.75	.68
22	.69	.88	47	.45	1.88*
23	.99	.07	48	1.18	-.39
24	.72	.79	49	1.51	-.97
25	.57	1.32	50	.51	1.58*

* Items flagged as showing significance DIF at 0.05 level.

significance. Seven (7) out of the eleven (11) items were in favour of the southerners while four (4) were in favour of the northerners. The M-H statistic showed that 23 and 26 items were relatively easier for the northern and southern Nigerian candidates, respectively. When items in the test were classified in terms of content area and cognitive level, there appeared to be no systematic differences in the items favouring candidates from north and south.

DISCUSSION, CONCLUSION AND RECOMMENDATIONS.

With the reported index of reliability, there is obviously little evidence of DIF by the criterion under the measure. The indices indicate that the test variance attributed to random error was not substantial for either of the groups under comparison. However, these results do not preclude the possibility that other kinds or reliability estimates might make DIF more evident (Green & Draper, 1972).

The percentile rank and cumulative percentage curve in Table 1 and Figure 1 shows that the southern Nigerian candidates obtained about 9 scores point higher than the Northern candidates. These are pointers to the fact that there may exist significant differential item functioning in the test due to region of the examinee, the arrangement of items in the test from simple to complex which was violated by WAEC might have introduced somewhat higher random guessing probability across the groups under comparison.

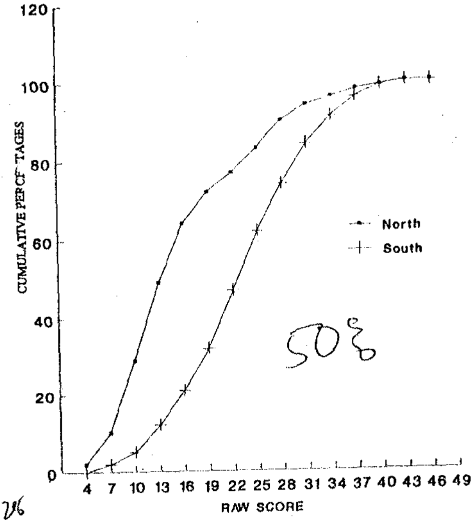

Figure 1: Cumulative Percentage Curve of the Total Raw Score

Comparing of the content classification and cognitive level of items significantly functioning for the candidates from the two regions gave no grounds for suspecting them as a source of DIF. It is however confounding as low and high cognitive items functioned significantly in favour of the southern candidates. Two items in the test with comparable difficulty and content area designed to measure candidate's spatial reasoning abilities were found. One favouring northerners while the other favoured the southerners. One would have expected an item relating a cliff to favour the southerners since they are closer to the Atlantic ocean, and thus could perceive the expected diagram needed than candidates from the north. Rather, it exhibited significant DIF in favour of the northerners who are farther away from the sea.

The characteristics of examinee to be a probable source of the DIF in the test is ruled out as the comparable groups were equated in terms of ability for the computation. However, the DIF was not due to systematic errors in scoring because the items analysed were objectively scored (multiple-choice items). It could however not result from timing of test, examiners instructions and modification of testing procedures, since the test items were printed and administered according to specifications outlined in the test paper. The examiners / supervisors attended the same briefing on the conduct of the examination. As a group administered examination, the influence of motivation was also non-existent, except the motivation were intrinsic. Then, it is hard to consider intrinsic motivation to be differentially distributed across region.

Another source which may come to reasoning is the unpleasant experience of performaning poorly on tests in the past, which may for that reason induce the examinee to perform poorly at future test situation which may be anxiety provoking. But if this is a significant source of DIF in the test, there is no evidence available to support the position that northern and southern Nigeria candidates receive varying treatment from WAEC. Also, all candidates serving in the sample for this analysis are assumed to have taken WAEC examination before and did not perform well, thereby warranting them to seek for all possible means of obtaining the certificate.

Interest, however, may be responsible for the observed regional bias because Northern and Southern Nigerian candidates may differ in their level of interest in Mathematics. They may be exposed to and compelled to take Mathematics achievement test since the subject is compulsory in schools, and a major conditionality for gaining admissions into University courses in the Sciences and Social Sciences. Another source of regional DIF in cognitive test items is individual or group differences in feelings of self-esteem or self-confidence. With the classification of most of the Northern states as educationally less developed, they may develop somewhat low self-esteem which significantly influenced their performance in the 50 – item objective Mathematics test administered by the West African Examination Council. These external factors may be responsible, in part, for the 9 score point difference between Northern and Southern candidates as shown in the cumulative percentage curve in Figure 1.

Based on the findings of the present study, it is concluded that the WAEC Mathematics achievement test designed and used for certification in 1990 for GCE was not absolutely free from regional DIF. Since WAEC examinations are meant for certification, and Nigerians are supposed to received equal treatment on such measures, the examination council should perform DIF analysis as one of the item analysis procedures utilized for their pilot studies prior to the actual testing with the target population to reduce the number of biased items in their test.

Future DIF researchers on a national examination in Nigeria in addition to using other DIF indices to study regional bias in Mathematics and other school subjects, should use large samples on any other groupings.

REFERENCES

Doolittle, A., 1987. Gender differences in performance on mathematics achievement items. Paper presented at the annual meeting of the American Psychological Association, New York.

Green, D., 1980. Procedure used for reducing bias in tests at CTB/McGraw-Hill, Paper presented at the Third Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C.

Green, D. & Draper, J., 1972. Exploratory studies of bias in achievement tests. A Paper presented at the Annual Meeting of the American Psychological Association, Honolulu.

Holland, P. W. & Thayer, D. T., 1986. Differential item performance and the Mantel-Haenszel procedure (Research Report No. *83 – 31).* Princeton: Educational Testing Service.

Inyang, M., 1991. Location bias analysis of students item/test performance Paper presented at the 7th Annual Conference of the National Association of Educational Psychologists, Zaria.

Miller, S., Doolittle, A; & Ackerman, T., 1988. Differential item performance for Mexican-American ESL students and white Non-ESL students on Mathematics and English achievement tests. Paper presented at NCME Annual Meeting, New Orleans, LA.

Federal Government of Nigeria, 1981. National Policy on Education. Lagos: Federal Ministry of Information

Nenty, H. J., 1986. Cross-cultural bias analysis of Cattell Culture-Fair Intelligence, Test. Paper at the Annual Meeting of AERA, San Francisco.

Rogers, H. J. & Kulick, E., 1987. An investigation of unexpected differences in item performance between Blacks and Whites taking the SAT, In A P. Schmitt & N. J. Dorans (Eds). Differential item functioning on the Scholastic Aptitude Test (Rm-87-1). Princeton, N. J. Educational Testing Service.

Ryan, K., 1991. The performance of the Mantel-Haenszel

Procedure Across Samples and Matching Criteria Journal of Educational Measurement, 28 (4): 325 – 337.

Schmitt, A . 1988. Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test Journal of Educational Measurement, 25 (1): 1 – 13.

Van der Flier, H., 1980. Vergelijkbaarheil van individual test prestaties. Lisse, The Nether Lands: Swets & Zeitlinger.

Van der Flier, H.; Mellenbergh, G. Ader, A. & Wijn, M., 1984. An iterative item bias detection method Journal of Educational Measurement, 21(2)· 131 – 145.

Jensen, A., 1980. Bias in mental testing. New York: The Free Press.

Umoinyang, I. E., 1996. A case of differential item functioning in achievement tests. In G. A. Badmus, & P. I. Odor, (Eds) Challenges of Managing Educational Assessment in Nigeria. Kaduna: Atman Publishers.

Umoinyang, I. E., 2000. Educational Development and differential item functioning (DIF) of Mathematics achievement test items in Nigeria. African Journal of Research in Education, 1: 23 – 31.