Randwick International of Education and Linguistics Science (RIELS) Journal
Vol. 1, No. 1, June 2020 | Page 1-13 |
http://www.randwickresearch.com/index.php/rielsj

The Applicability of Item Response Theory Based Statistics to Detect Differential Item Functioning in Polytomous Tests

DOI: https://doi.org/10.47175/rielsj.v1i1.23

| Abdul Wahab Ibrahim |

Department of Education, Faculty of Education, Sule Lamido University, Nigeria

wahabpsychodata2017@gmail.com

ABSTRACT

The study used statistical procedures based on Item Response Theory to detect Differential Item Functioning (DIF) in polytomous tests. These were with a view to improving the quality of test items construction. The sample consisted of an intact class of 513 Part 3 undergraduate students who registered for the course EDU 304: Tests and Measurement at Sule Lamido University during 2017/2018 Second Semester. A self-developed polytomous research instrument was used to collect data. Data collected were analysed using Generalized Mantel Haenszel, Simultaneous Item Bias Test, and Logistic Discriminant Function Analysis. The results showed that there was no significant relationship between the proportions of test items that function differentially in the polytomous test when the different statistical methods are used. Further, the three parametric and non-parametric methods complement each other in their ability to detect DIF in the polytomous test format as all of them have capacity to detect DIF but perform differently. The study concluded that there was a high degree of correspondence between the three procedures in their ability to detect DIF in polytomous tests. It was recommended that test experts and developers should consider using procedure based on Item Response Theory in DIF detection.

KEYWORD

Item Response Theory (IRT); Differential Item Functioning (DIF); DIF Magnitude; Dichotomous Test; Ordinal Test; Polytomous Tests.

INTRODUCTION

The presence of Differential Item Functioning (DIF) jeopardizes the ideal of a correct measurement procedure. Once identified, DIF may be attributed to item impact or to item bias. Item impact is evident when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item because there are true differences between the groups in the underlying ability being measured by the item. Item bias occurs when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. DIF is required, but not enough, for item bias (Ibrahim, 2016).

DIF is said to be present in a test item when, despite controls for overall test performance, examinees from different groups have a different probability of answering an item correctly or when examinees from two subpopulations with the same trait level have different expected scores on the same item (Osterlind & Everson, 2009). One of the pioneering methods used to detect DIF is known as the Generalized Mantel-Haenszel

procedure (GMH). This method is based on contingency table analysis and was first used to detect DIF by Holland and Thayer (1988). The GMH procedure compares the item performance of the reference and focal groups, which were previously matched on the trait measured by the test; the observed total test score is normally used as the matching criterion. In the standard GMH procedure, an item shows DIF if the odd of correctly answering the item is different for the two groups at a given level of the matching variable (Ibrahim, 2018).

According to Kristjansson, Aylesworth & Zumbo (2005), the Generalized Mantel-Haenszel (GMH) is a generalized statistic for nominal response data based on group differences in the entire response distribution. Because the GMH tests differences across the entire response scale, it should be sensitive to both uniform and non-uniform DIF. Formulae for calculating the GMHχ^2 statistic are given by Zwick, Donoghue & Grima (2010). The data for the studied item for the examinees in the reference and focal groups are arranged into a series of *2 x 2* contingency tables, one for each level of the matching variable. In the notation of GMHχ^2 statistic,

$$A^1_m = (N_{Rm1}, N_{Rm2}, \ldots N_{Rm(k-1)}),$$

$$E(A^1_m) = N_{Rm} N_{1t} 1 N_m,$$

$$N^1_T = (N_{ml}, N_{m2}, \ldots, N_{m(k-l)}),$$

$$V(A_m) = N_{Rm} N_{Fm} (N_m \, dia(N_\tau) - N_\tau N^1_\tau),$$

$$N^2_m (N_m - 1)$$

and $dia(N_T)$ is a $(k-1)$ diagonal matrix with elements N_T. Whereas A_m, $E(A_m)$, and $V(A_m)$ are scalars in the dichotomous case, A_m and $E(A_m)$ are now vectors of length k -1, corresponding to k -1 of the k response categories, and $V(A_m)$ is a (k-1) by (k – 1) covariance matrix. Then the GMH test statistic is given by

$$GMH\chi^2 = [\sum A_m - \sum E(A_m)] \; [\sum V(A_m)]^1 \; [\sum A_m - \sum E(A_m)] \qquad \text{(Equation 1)}$$

This statistic has a large sample chi-square distribution with k-1 degrees of freedom under the null hypothesis of conditional independence between group membership and item response. A significant test statistic implies that uniform DIF is present in the item. The GMH statistic does not explicitly take into account the possible ordering of response categories; instead, it provides for the comparison of the two groups in terms of their entire response distributions, rather than their means alone. The odds that focal group members will be assigned a particular score category can be compared to the odds for the reference group, conditional on the matching variable (Holland & Wainer, 2009). Using a log-odds transformation, Holland & Thayer (2006) converted α_{MH} *into* a difference on a delta (Δ) scale, called $MH_{D\text{-}DIF}$. $MH_{D\text{-}DIF}$ has been frequently used as a measure of DIF. Two generalizations of the dichotomous MH procedure can be applied to assess DIF in polytomous item responses, one for ordinal response data, the other for nominal response data.

Like the GMH procedure, Simultaneous Item Bias Test (SIBTEST) proposed by Shealy & Stout (1993) is a conceptually simple method and involves a test of significance based on the ratio of the weighted difference in proportion correct (for reference and focal group members) to its standard error. SIBTEST was originally intended for use with dichotomous test items but has since been extended to handle ordered items. Like the GMH procedure, SIBTEST yields an overall statistical test as well as a measure of the effect size for each item (β is an estimate of the amount of DIF). SIBTEST is the designation given to the

statistical methodology for detecting uniform DIF and is based on the comparison of the probability of a correct response on the target item for the reference group at a given value of the latent ability (θ), with the probability of a correct response on the target item for the focal group at the same ability level (Ibrahim, 2018). The null DIF definition for SIBTEST is that an item exhibits DIF if the expected scores are identical for the reference and focal groups matched on θ. The amount of DIF at θ is measured by:

$$B_0(\theta) = E_R[Y/\theta] - E_F[Y/\theta] \quad \text{(Equation 2)}$$

At a given ability (θ), this difference is expressed as

$$B(\theta) = P_R(\theta) - P_F(\theta) \quad \text{(Equation 3)}$$

The SIBTEST statistic, β is the average difference in the probability of a correct response for the two groups, so when uniform DIF is not present this value is 0. Because the true distribution of θ is unknown, examinees are matched on their observed scores from a subset of the items. To correct for ability differences in the two groups, which are known to influence comparison of conditional probabilities, these observed subtest scores are taken separately for each group and adjusted using a regression equation based in classical test theory to estimate true scores, $T_R(s)$ and $T_F(s)$, for members of the reference and focal groups, respectively. The proportion correct for each group is then conditioned on a common true score, which is estimated as the average of $T_R(s)$ and $T_F(s)$ (Ibrahim, 2018).

As with SIBTEST and the Generalized Mantel-Haenszel procedures, Logistic Discriminant Function Analysis (LDFA), proposed by Miller & Spray (2003), is a parametric DIF detection approach which provides both a significance test and a measure of effect size. LDFA is closely related to logistic regression, and it is also model- based. However, there is one major difference in the LDFA method namely that group membership is the dependent variable rather than item score. Thus, in LDFA, the probability of group membership is estimated from total score and item score. This is a logistic form of the probability used in discriminant function analysis. LDFA is a DIF identification of items that are polytomously scored (items with multiple correct responses such as a Likert scale or a constructed-response item). In LDFA, three equations are derived: an equation predicting group membership from total score only; an equation predicting group membership from total score and item score; and an equation predicting group membership from total score, item score, and item by total score. A likelihood ratio goodness-of-fit statistic, G^2, is computed for each model. As with the other two DIF techniques described here, its Type 1 error is generally near or below the normal rate of 0.05 but may be problematic when group ability differences are present. In the logistic regression model, the item response variable, *U*, is treated as a random variable and X and G are assumed to be fixed explanatory variables. However, it has been shown that it is reasonable to use the logistic regression procedure to estimate Prob (G\X,U) even though G is fixed and U is random (Hosmer & Lemeshow, 2000).

In this form, Prob (G\X, U) is simple a logistic form of the posterior probability used in discriminant analysis. This procedure is called logistic discriminant function analysis (LDFA). When applying the logistic discriminant function analysis to assess DIF in ordered item responses, the discriminant function (without item notation) can be written as:

$$\text{Prob}(G\backslash X, U) = \frac{e^{(1-(i)(-a_o - a_1 x - a_2 U - a_3 xU)}}{1 + e^{(1-G)(-a_0 - a_1 x - a_2 U - a_3 xU)}}, \quad \text{(Equation 4)}$$

With these methods, however, there is not yet a consensus about how to test DIF when item responses are polytomously scored, even though, the most widely used DIF detection methods are procedures based on Item Response Theory (IRT). These methods have been useful in detecting DIF over time. Several extensions of the DIF procedures have been proposed for use with polytomous item responses, such as the Ordinal Logistic Regression procedure, the Mantel procedure for ordered response categories, the Generalized Mantel Haenszel procedure for nominal data, the polytomous extension of SIBTEST, the polytomous extension of the standardization approach, and Logistic Discriminant Function Analysis. However, their utilities in assessing DIF in ordinal items have not received the thorough and rigorous study accorded to the dichotomous DIF, thus necessitating further research to investigate their performance before they are ready for routine operational use (Ibrahim, 2017). As a corollary to the above, this study empirically compares the relative ability of the three statistical methods for detecting Differential Item Functioning in polytomous test items. Towards this end, the specific objective of the study appears germane namely to determine the relationship between the proportions of test items that function differently in the polytomous tests when the different methods are used. To achieve the objective of the study, a null research hypothesis was postulated:

Research Hypothesis

There is no significant relationship between the proportions of test items that function differentially in the polytomous tests when the different methods are used.

METHOD

This study employed the descriptive-comparative research design. According to Upadhya and Singh (2008), descriptive-comparative research design study compares two or more groups on one variable with a view to discovering something about one or all the things being compared. Hence, two groups: reference and focal groups' combination were used in the Differential Item Functioning analysis. In carrying out this study, therefore, the researcher collected data from subset of the population (undergraduate students in 300 level) in such a way that the knowledge to be gained is representative of the total population under study. Essentially, the researcher used the data collected to explore the three statistical DIF detection methods being studied in this study. In this study, the three methods SIBTEST, GMH, and LDFA were based on a contingency table framework; within this framework, total test score was used as the measure of trait level. Hence, DIF was held to exist if group differences occur in item score after matching on total score. The nature of this research, the sample and data collected determined the relevance/appropriateness of this design.

Participants

All undergraduate students who registered for a compulsory course in Tests and Measurement during the Second of 2017/2018 Session in the Faculty of Education of the Sule Lamido University, Kafin Hausa, Jigawa State, Nigeria, constituted the target population for the study. There were 513 undergraduate students who registered for the course during the session. The sample consisted of an intact class of 513 part 3 undergraduate students who registered for EDU 304 in Second Semester of 2017/2018 session. Thus, the entire population was therefore used, and no sampling was carried out as sampling procedure was a convenience type.

Research Instrument

A self-developed polytomous instrument was used in the study namely: "Undergraduate Students Achievement and Efficacy Scale (USAES)". The instrument contained 74 items divided into dichotomous and ordinal scales and rated on a five-point Likert-scale. First, the dichotomous section of the instrument consists of a 50, 4-option multiple-choice test that was developed using the course (EDU 304: Tests and Measurement) content. Second, the ordinal section of the instrument consists of 24-item which is made up of six subscales. The response format for the scale was the Likert type with five options of Strongly Agree (SA), Agree (A), Undecided (U), Disagree (D), and Strongly Disagree (SD).

The content and construct validity of the instrument was established using expert judgments. Experts in Tests and Measurement, Statistics, Psychology for scrutiny and modification established the content validity of the instrument. The experts were able to review the items in the instrument in terms of relevance to the subject-matter, coverage of the content areas, appropriateness of the language usage and clarity of purpose. The experts' judgments revealed that the instrument had adequate content, construct and face validity. Thereafter, a reliability process was done to establish how reliable the instrument is. Hence, reliability test was conducted on the whole data collected for pilot testing using the reliability analysis tool on the Statistical Package for Social Sciences (SPSS), version 24.0. The instrument was pilot tested using 60 part three students in the Faculty of Education, Bayero University, Kano, Kano State, Nigeria, who were also offering the same course with similar course content. The reliability of the scores obtained in the pilot study was estimated using Cronbach's Alpha, Spearman Brown Split-Half Coefficient, and Guttman Split-Half Coefficient. The Coefficients obtained were 0.76, 0.89, and 0.89 respectively. Its mean ($\bar{x}$) difficulty index is 0.70 with a standard deviation of 0.28. The item discrimination indices have a mean ($\bar{x}$) value of 0.23 and a standard deviation of 0.17, with minimum and maximum scores of 10.0 and 35.0 respectively, and a variance of 67.7. Noteworthy, the Split-Half reliability method was preferred because of the desire to determine the internal consistency of the instrument for data collection. The Split-Half method was preferred because it was not feasible to repeat the same test. Also, it was considered a better reliability method in the sense that all the data required for computing reliability are obtained on one occasion, and therefore, variations arising out of the testing situations do not interfere with the results and outcomes of this study. According to Afolabi (2012), Split-Half reliability provided a measure of consistency with respect to content sampling; hence its preference in this study. All these values were acceptable as appropriately high for study of human behaviour due to its complexity. Consequently, the instrument was accepted being stable over time, hence its usage in this study.

Procedure for Data Collection

The instrument was administered by the researcher. The hard copies of the instrument were administered on the students with the assistance of the course Lecturers of EDU 304, as well as a handful of some Assistant Lecturers in the Department of Education of the Sule Lamido University, Kafin Hausa, Jigawa State, Nigeria. The instrument administration was conducted under strict but friendly condition. However, adequate time was provided for respondents to respond to all the items. Furthermore, the respondents were instructed not to omit any item as it is mandatory to answer all items in the instrument as they marked on the instrument that response which they have decided is most correct. Such a procedure provided a uniform response set thereby minimizing individual differences in responding. Consequently, the administered instrument copies were collected immediately. A total of 513 copies of the instrument were administered, while 502 copies were finally collected on return, as being properly completed and were used for analysis.

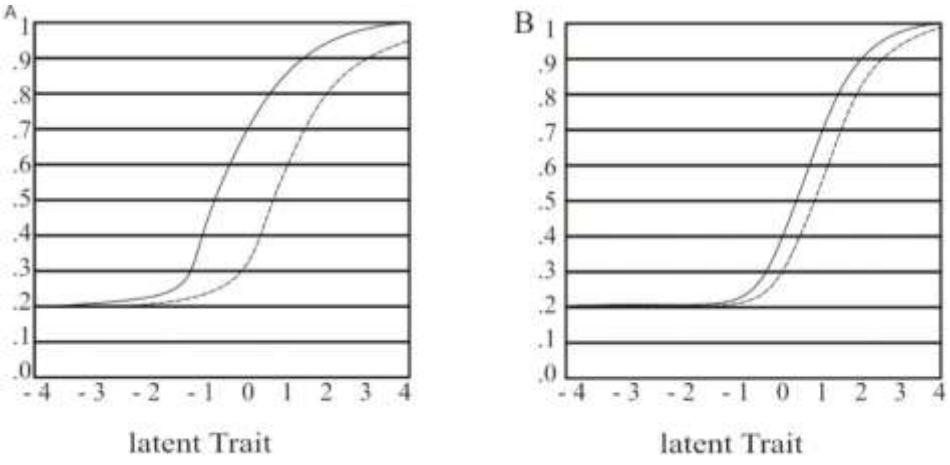
Method of Data Analysis

DIF statistical analyses were conducted for each item using GMH, SIBTEST, and LDFA statistical methods. These test statistics were interpreted at an alpha-level of 0.05. The software package DIF OpenStat developed by Miller (2011); and DIF LazStats developed by Pezzulo (2010) were used to run the three statistical procedures. Updated SPSS version 24.0 and Microsoft Excel version 12.0 were used to manage and organize the datasets.

RESULTS

In an effort to better understand the direction of DIF magnitude, we investigated the Item Characteristic Curves (ICCs) for items displaying the highest and lowest error rates. Figure 1 contains a two ICCs for the two groups, one for a high-difficulty, medium discrimination item with DIF of 1.0 (Figure 1A) and the other with DIF of .4 (Figure 1B). The DIF error rate for SIBTEST, GMH and LDFA were .971, .658, and .427, respectively, in the dichotomous test items and .622, .436 and .239, respectively, in the ordinal test items. For examinees with ability values from approximately -1 through 4, it appears that the separation between the two groups is clear in the DIF = 1.0 case. Conversely, at the low end of the ability scale, the ICCs come very close to one another so that they are difficult to disentangle visually. Indeed, at θ = - 1.0 the probability of a correct response for the reference group was .243 and for focal group, .206. In contrast, at θ = -3.0 this gap had closed considerably, with a correct response probability for the reference group of .202 and for the focal group, .200. In short, for individuals at lower ability levels, the probability of a correct response approached the lower asymptote of .2, regardless of group membership, thus quite possibly leading to the detection of DIF. The DIF = .4 condition resulted in a much lower error rate for SIBTEST, GMH, and LDFA.

In this case, the gap between curves for the two groups was much smaller across the ability levels as compared to the ordinal test item. At θ = -1.0 the reference and focal groups' probabilities of a correct response were .224 and .211, respectively, while at θ = -3.0 both groups had a probability of a correct response of .201. For both dichotomous and ordinal test items there was very little difference in the ability of a correct response for the two groups at low abilities, with both approaching the lower asymptote of 0.2. However, the item containing greater *b* DIF had a larger gap in the probability of a correct response between the two groups for abilities greater than -1 than did the item containing less *b* DIF. Thus, for items where the ICCs experience a greater change in the gap between the two across ability scale, SIBTEST and LDFA detect an interaction and signal the presence of DIF, although the *a*-parameter values for the two groups were equal.



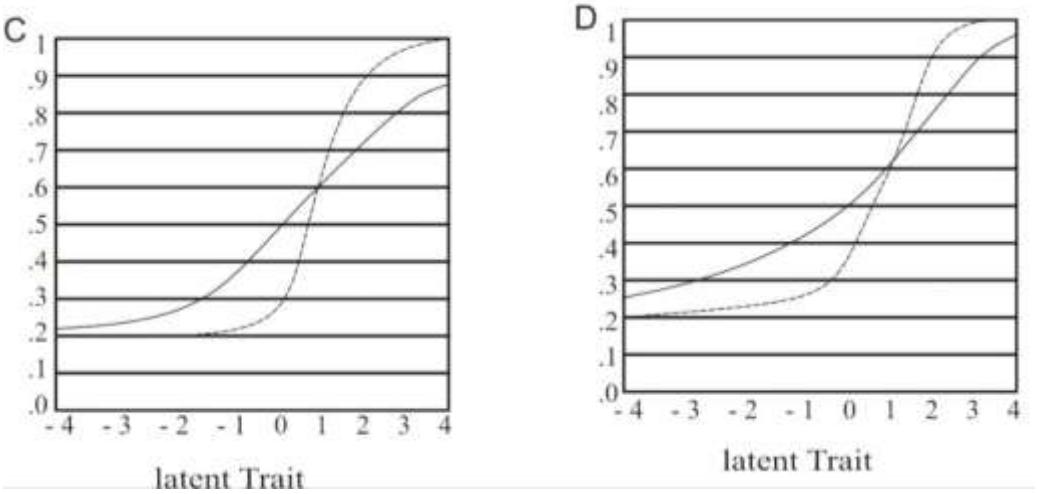


Figure 1. Item Characteristic Curves (ICCs) for Reference and Focal Groups for Dichotomous and Ordinal Items

Further, upon completion of the analysis, the proportion of correctly identified DIF items was used as a power estimate for combining the Δp_j (*p- values*) across the levels of total scores, the results are presented in Table 1 for dichotomous test and Table 2 for ordinal test.

Table 1. Proportion of Test Items that Function Differentially in the Dichotomous Test Using GMH, SIBTEST, and LDFA Methods

	Reference Group			**Focal Group**			***p-values difference***		
Items	GMH *p-values*	SIBTEST *p-values*	LDFA *p-values*	GMH *p-values*	SIBTEST *p-values*	LDFA *p-values*	GMH *b*	SIBTEST *b*	LDFA *B*
1	.24(-0.04)	.28(0.02)*	.26(-0.02)	.37(0.04)	.33(-0.05)	.38(-0.01)	-0.13	0.03	- 0.12
2	.11 (-0.22)	.33(-0.14)*	.47(-0.36)	.35(-0.13)	.48(0.01)	.47(-0.12)	-0.24	-0.15	0.00
3	.20(0.02)	.18(-0.17)*	.35(-0.15)	.48(0.12)	.36(-0.08)	.44(-0.04)*	-0.28	-0.18	0.09
4	.24(-0.12)	.36(0.03)	.33(-0.09)	.30(0.08)	.22(-0.06)	.28(0.02)	-0.06	0.14	0.05
5	.11(-0.19)	.30(0.05)	.25(-0.14)	.15(-0.13)	.28(0.02)	.26(-0.11)	-0.04	0.02	-0.01
6	.33(-0.13)	.46(0.04)	.42(-0.09)	.52(-0.04)	.56(0.23)	.33(0.19)	-0.19	-0.01	0.09
7	.66(0.08)	.58(-0.09)	.67(-0.01)	.70(0.05)	.65(0.07)	.58(0.12)	-0.04	-0.07	0.09
8	.63(-0.10)	.73(-0.03)	.76(-0.13)	.79(-0.08)	.87(0.05)	.82(-0.03)	-0.16	-0.14	-0.06
9	.84(0.08)	.76(-0.07)	.83(0.01)	.83(-0.04)	.87(-0.06)	.93(-0.10)	0.01	-0.11	-0.10
10	.92(0.04)	.94(-0.04)	.98(-0.06)	.95(-0.02)	.97((-0.03)	1.00(-0.05)	-0.03	-0.03	-0.02
11	1.00(0.00)	1.00(0.50)	.50(0.50)	.90(0.03)	.57(0.26)*	.31(0.59)	0.10	0.43	0.19
12	.35(0.10)	.25(0.05)	.20(0.15)	.43(-0.32)	.75(0.47)	.28(0.15)	-0.08	-0.50	-0.08

13	.66(0.14)	.52(0.19)	.33(0.33)	.76(0.36)	.40(-0.35)	.75(0.01)	-0.01	0.12	-0.42
14	.75(0.25)	.50(-0.40)	.90(-0.15)	.69(-0.31)	1.00(0.00)	1.00(-0.31)	0.06	0.50	-0.01
15	.83(0.08)	.75(0.25)	.50(0.33)	1.00(0.00)	1.00(0.00)	1.00(0.00)	-0.17	-0.25	-0.50
16	1.00(0.03)	.97(0.01)	.96(0.04)	.98 (0.04)	.94(0.01)	.93(0.05)	0.02	0.03	0.03
17	.66(0.13)	.53(0.20)	.33(0.33)	.77(0.00)	.77(0.37)	.40(0.37)	-0.11	0.24	-0.07
18	.67(0.34)	.33(-0.27)	.60(0.07)	.53(0.42)	.11(0.73)	.84(-0.31)	0.14	0.22	0.24
19	.76(0.01)	.77(0.44)	.33(0.43)	.82(0.17)	.65(0.01)	.64(0.18)	-0.06	0.12	-0.31
20	.32(0.13)	.19(-0.12)	.31(0.01)*	.23(0.09)	.14(-0.17)	.21(0.02)	0.09	0.15	0.10
21	.92(0.20)	.72(0.05)	.67(0.25)	.23((0.01)	.22(-0.64)	.86(-0.63)	0.69	0.50	-0.19
22	.12(0.06)	.18(-0.38)	.56(-0.44)	.78(0.25)	.53(0.15)	.38(0.40)	-0.66	-0.35	0.18
23	.59((0.09)	.50(-0.05)	.55(0.04)	.64(0.02)	.66(0.07)	.59(.0.05)	-0.05	- 0.16	- 0.04
24	.29(-0.04)	.33(-0.14)	.47(-0.18)	.50(0.27)	.23(-0.08)	.31(0.19)	-0.21	0.10	0.16
25	.50(-0.08)	.58(0.03)	.55(-0.05)	.49(-0.08)	.57(0.02)	.55(-0.06)	0.01	- 0.01	-0.00
26	.33(-0.14)	.47(-0.22)	.69(-0.36)	.69(0.08)	.61(0.04)	.57(0.12)	-0.36	-0.14	0.12
27	.98(0.10)	.88(0.31)	.57(0.41)	.63(0.08)	.55(-0.44)	.99(-0.36)	0.35	0.33	0.42
28	.83(0.26)	.57(0.25)	.32 (0.51)	.49 (-0.20)	.69(-0.31)	1.00(-0.51)	0.43	-0.12	-0.68
29	.53(-0.23)	.76(0.05)	.71(-0.18)	.32 (-0.40)	.72(-0.15)	.85(-0.53)	0.21	0.04	-0.14
30	.51(0.19)	.32 (0.09)	.23(0.28)	.40 (-0.12)	.52 (0.25)	.27(0.13)	0.11	-0.32	-0.04
31	.74(-0.21)	.95(0.13)	.82 (-0.08)	.66 (-0.09)	.75 (0.10)	.65(0.01)	0.08	0.20	0.20
32	.81(0.16)	.65(-0.26)	.91(-0.10)	.73(0.11)	.62 (-0.04)	.66(-0.07)	0.08	0.03	0.25
33	.68(-0.06)	.74(0.28)	.46(0.20)	.57(0.06)	.51(-0.18)	.69 (-0.12)	0.11	0.23	-0.23
34	.85(0.63)	.22(-0.64)	.86(-0.01)	.51(-0.10)	.61(0.07)	.54 (-0.03)	0.34	-0.39	0.25
35	.55(0.02)	.53(-0.03)	.56(-0.01)	.60(-0.30)	.90(0.12)	.78(-0.18)	-0.05	-0.37	-0.22
36	.59(-0.28)	.87(-0.05)	.92(-0.33)	.45(-0.28)	.73(0.00)	.73(-0.28)	0.14	0.14	0.19
37	.58(-0.16)	.74(0.33)	.41(0.17)	.72 (0.08)	.64 (0.18)	.46(0.26)	0.14	0.10	-0.05
38	.63(0.02)	.61(-0.28)	.89 (-0.26)	.52(-0.17)	.69(-0.06)	.75(-0.23)	0.11	-0.08	0.14
39	.66(-0.05)	.71(0.39)	.32(0.34)	.77(0.26)	.51(-0.27)	.78(-0.01)	-0.11	-0.06	-0.46
40	.18(-0.05)	.23(-0.62)	.85(-0.67)	.78(0.53)	.25(-0.07)	.32(0.46)	-0.60	-0.02	0.53
41	.73(0.26)	.47(0.15)	.32((0.31)	.22 (-0.22)	.44(0.08)	.36(-0.14)	0.51	0.03	-0.04
42	.30 (0.56)	.86(-0.14)	1.00(-0.70)	.60(-0.30)	.90(0.59)	.31(0.29)	-0.30	-0.04	0.69

43	.80(0.08)	.72(-0.11)	.83(-0.03)	.77(0.09)	.68(-0.04)	.72(0.05)	0.03	0.04	0.11
44	.78 (0.09)	.69 (0.05)	.64(0.14)	.89(0.27)	.62(0.05)	.67(0.22))	-0.11	0.07	0.03
45	.62(0.31)	.31(0.13)	.18(0.44)	.61(-0.34)	.95(0.04)	.91(0.30)	0.01	-0.64	-0.73
46	.98(0.05)	.93(0.58)	.35((0.63)	.14(-0.38)	.52(-0.14)	.66(-0.52)	0.84	0.41	-0.31
47	1.00(0.00)	1.00(0.10)	.90(0.10)	.35(-0.65)	1.00(0.20)	.80(-0.45)	-0.65	0.00	0.10
48	1.00(0.65)	.35(-0.65)	1.00(0.00)	.15(0.02)	.13(-0.67)	.80(-0.65)	0.85	0.22	0.20
49	.38(-0.12)	.50(-0.06)	.56(-0.18)	.52(0.00)	.52(0.02)	.50(0.02)	-0.14	-0.02	0.06
50	.46(-0.05)	.51(0.00)	.51(-0.05)	.68(0.22)	.46(-0.06)	.52(0.16)	-0.22	0.05	-0.01

Significant, p < .05

Table 1 presents the proportions of test items (*p-values*) that function differentially in the dichotomous test when the GMH, SIBTEST, and LDFA methods are used. In the 50-item test, both GMH and LDFA identified seventeen items each as proportionately functioning differently in the dichotomous test for the reference and focal groups respectively. GMH found such items as being items 2, 3, 21, 22, 24, 26, 27, 28, 29, 34, 40, 41, 42, 46, 47, 48, and 50. Also, LDFA flagged items 13, 15, 18, 19, 27, 28, 31, 32, 33, 34, 35, 39, 40, 42, 45, 46, and 48. Further, SIBTEST identified such items in at least fourteen items, being items 11, 12, 14, 15, 17, 18, 21, 22, 27, 30, 31, 33, 34, and 35. Hence, both GMH and LDFA are very promising procedures for detecting proportion of DIF items in dichotomous test but the SIBTEST is less effective.

Similarly, Table 2 displays the results of the proportions of test items that function differentially in the ordinal test when the different methods are used. As seen in the table, the p- value difference for item 1, when the three methods are used, appears to be out of line with the results for the same item for both reference and focal groups. For instance, when SIBTEST was used, the p-value difference of .25, $p < .05$, was obtained, as compared with the GMH result 0.03, $p < .05$, and LDFA result 0.05, $p < .05$, which shows that item 1, was identified as proportionately functioning differently when the three methods are used. Hence, ten such items were found, being items 1,3, 7, 8, 11, 14, 17, 19, 20, and 22 identified by SIBTEST as proportionately functioning differently for both reference and focal groups. Further examination of the 24 items indicated that only items 5, 9, 11, 13, 14, and 22 were flagged by GMH and items 3, 16, 18, and 20 were identified by LDFA respectively as proportionately functioning differently for both reference and focal groups. Also, GMH flagged item 13 as outlier (.25, $p < .05$) amongst the remaining two methods (SIBTEST = .08, $p < .05$, and LDFA = .01 $p < .05$).

Table 2. Proportion of Test Items that Function Differentially in the Ordinal Test Using GMH, SIBTEST, and LDFA Methods

	Reference Group			**Focal Group**			***p-values difference***		
Items	GMH *p-values*	SIBTEST *p-values*	LDFA *p-values*	GMH *p-values*	SIBTEST *p-values*	LDFA *p-values*	GMH *b*	SIBTEST *b*	LDFA *b*
1	.67(-0.32)	.98(0.29)*	.69(-0.03)	.64(-0.09)	.73(0.05)	.68(0.01)	0.03	0.25	0.05

2	.33 (-0.62)	.96(-0.04)	1.00(-0.67)	.35(-0.63)	.98(0.01)	.97(-0.62)	-0.02	-0.02	0.03
3	.60(-0.38)	.98(-0.02)	1.00(-0.4)	.68(0.01)	.67(0.03)	.64((0.04)*	-0.08	0.31	0.36
4	.24(-0.77)	1.00(0.17)	.83(-0.59)	.70(-0.22)	.92(0.14)	.78(-0.08)	0.17	0.08	0.05
5	.11(-0.89)	1.00(0.25)	.75(-0.64)	1.00(0.12)	.88(0.20)	.68(0.32)	-0.89	0.12	0.07
6	.37(-0.13)	.50(0.00)	.50(-0.13)	.22(-0.14)	.36(-0.17)	.53(-0.31)	0.15	0.14	-0.03
7	.47(0.38)	.09(-0.91)	1.00(-0.53)	.30(-0.70)	1.00(0.17)	.83(-0.53)	0.17	-0.91	0.17
8	.58(0.28)	.31(-0.69)	1.00(-0.42)	.46(0.38)	.80(-0.18)	.98(-0.52)	0.12	0.23	0.02
9	.59(0.24)	.36(-0.65)	1.00(-0.41)	.66(0.36)	.30(-0.60)	.90(-0.24)	0.23	0.06	0.10
10	.67(0.43)	.25(-0.75)	1.00(-0.32)	.68(0.42)	.26((-0.74)	1.00(-0.32)	-0.01	-0.01	0.00
11	.64(0.44)	.20(-0.23)	.43(0.21)	1.00((0.43)	.57(0.15)*	.42(0.58)	-0.36	-0.37	0.01
12	.73(0.29)	.44(-0.21)	.64(0.09)	.71(0.28)	.43(-0.15)	.58(0.13)	0.02	0.01	0.06
13	.76(0.01)	.75(0.21)	.54(0.22)	.51(-0.16)	.67(0.14)	.53(-0.02)	0.25	0.08	0.01
14	.79(0.51)	.29(-0.44)	.73(0.06)	1.00(0.80)	.20(-0.49)	.69(0.31)	-0.21	0.27	0.04
15	.87(0.21)	.67(-0.02)	.69(0.19)	.83(0.32)	.51(-0.06)	.57(0.26)	0.04	0.16	0.12
16	.82(0.21)	.53(-0.15)	.67(0.15)	.98 (0.51)	.47(-0.53)	1.00(-0.02)	-0.16	0.06	-0.33
17	.84(0.51)	.33(-0.29)	.62(0.22)	.90(0.20)	.70(0.00)	.70(0.20)	-0.06	-0.37	-0.08
18	.77(0.00)	.77(0.00)	.77(0.00)	.70(-0.05)	.75(0.22)	.53(0.17)	0.07	0.02	0.24
19	.84(0.07)	.77(0.27)	.50(0.34)	.65(0.10)	.55(-0.09)	.64(0.01)	0.19	0.22	-0.14
20	.84(0.44)	.40(-0.05)	.45(0.39)*	.87(0.11)	.76(0.11)	.65(0.22)	-0.03	-0.36	-0.20
21	.87(0.12)	.75(0.09)	.66(0.21)	.90((0.10)	.80(0.25)	.55(0.35)	-0.03	-0.05	0.11
22	.93(0.18)	.75(0.26)	.49(0.44)	.45(-0.08)	.53(0.15)	.38(0.07)	0.48	0.22	0.11
23	.93((0.43)	.50(-0.08)	.58(0.35)	.87(0.41)	.46(-0.07)	.53(.0.34)	0.06	0.04	0.05
24	.94(0.04)	.90(0.25)	.65(0.29)	1.00(0.00)	1.00(0.53)	.47(0.53)	-0.06	-0.10	0.18

Significant, p < .05

Further, Table 3 presents the results of the Chi-square (χ^2) analysis. From Table 3, 8% of the proportion of items functioning differentially in ordinal test flagged DIF when GMH was used as compared with 23% of the items flagged as containing DIF in dichotomous test. Also, SIBTEST flagged 14% of the items in the ordinal test as proportion of items functioning differentially and 22% of the items in the dichotomous test flagged as proportion of items functioning differentially. Similarly, LDFA flagged 11% of the ordinal items as proportion of items functioning differentially and 23% of the items flagged as proportion of items functioning differentially in the dichotomous test.

Table 3. Relationship Between the Proportion of Test Items That Function Differentially in the Dichotomous and Ordinal Tests

Methods	Dichotomous	Ordinal	Total	χ^2	P
GMH	17(22.9%)	6 (8.1%)	23		
SIBTEST	16 (21.6%)	10 (13.5%)	26	0.98	> 0.05
LDFA	17 (22.9%)	8 (10.8%)	25		
Total	50	24	74		

Not significant, p>0.05

Further, the Chi-square (χ^2) analysis of the results yielded 0.98, which is not significant at p>0.05. Thus, the null hypothesis is confirmed; that is, there is no significant relationship between the proportion of test items that function differentially in the dichotomous and ordinal tests when the different methods are used.

DISCUSSION

The finding of this study indicated that there was no significant relationship between the proportion of test items that function differentially in the dichotomous and ordinal tests when the different methods are used. The results of this study are in consonance with the earlier findings of DeAyala (2012) which concluded that the LR procedure was as powerful as the MH procedure in detecting uniform DIF, and more powerful than the MH in detecting. In addition, as Dorans and Schmitt (2009) stated, If LR DIF can detect non-uniform DIF better than the MH DIF method, and is as powerful at detecting uniform DIF as the MH DIF method, then the inclusion of an effect size would make LR DIF a very attractive choice as a DIF detection method. The researcher believes that the results of this study can bear more significance by taking one point into account. LDFA is a parametric DIF detection approach which is a response to the previous DIF techniques which could only screen uniform DIF such as Standardization, GMH or SIBTEST. This, implicitly, can be considered as a reassuring point for the developers of the dichotomous and ordinal tests. This finding is similar to Gierl, Khaliq and Boughton (2003) who reported that while LDFA has comparable power to GMH and SIBTEST in detecting uniform DIF, it is superior in power for detecting non-uniform DIF. Hosmer and Lemeshow (2000) found that effect size measures (for GMH and SIBTEST) were highly correlated across DIF procedures except the measure for non-uniform DIF, which could only be assessed by GMH.

These results are in line with Miller and Spray (2003) who used LDFA and SIBTEST for DIF identification in polytomously scored items, confirmed that for both item 4 and item 17, the power to detect DIF increased as the DIF magnitude increased. This trend occurred when there were no missing data as well as when missing data were present. Conditions with a DIF magnitude of .25 had the poorest power, while conditions with a DIF magnitude of .75 had the highest power. Conditions with a DIF magnitude of .25 typically had power below 70% which is generally considered adequate power. On average, item 17 had slightly higher power values than item 4. This difference may be due to the varying degree of difficulty of item 4 and item 17. Altogether, these findings provide tacit confirmation as to the superiority of GMH over SIBTEST.

CONCLUSION

On the strength of the findings obtained from the study, it can be concluded, therefore, that the three methods complement each other in their ability to detect DIF in the dichotomous and ordinal test formats as all of them have capacity to detect DIF but perform differently. From the findings of this study, the following recommendations were made: (i) statistical methods for detecting Differential Item Functioning should be an essential part of test development and test evaluation efforts; (ii) moreover, quantitative and qualitative (expert judgment) analyses that can inform the test development process should be conducted after the administration of a test; and (iii) test experts and developers should consider using contingency table approaches, preferably the GMH and LDFA approaches in DIF detection. DIF testing must be conducted especially for very important tests like psychological instruments used by various researchers in all Nigerian Universities.

REFERENCES

Afolabi, E. R. I. (2012). Validity and reliability. In E.R.I. Afolabi & O. O. Dibu-Ojerinde (Eds.), *Educational tests & measurement* (pp. 190-200). Ile-Ife, Nigeria: Obafemi Awolowo University Press.

DeAyala, R. J. (2012). *The theory and practice of item response theory*. New York: The Guilford Press.

Dorans, J., & Schmitt, K. (2009). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine*, 19, 165 – 183.

Gierl, M. J., Khaliq, S. N., & Boughton, K. A. (2003). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*, 26–36.

Holland, P. W., & Wainer, H. (2009). *Differential item functioning.* New York: Routledge Taylor & Francis Group.

Holland, P. W., & Thayer, D. T. (2006). Differential item performance and the Mantel

Haenszel procedure. In H. Wainer, H. I. Braun (Eds.), *Test validity* (pp. 23-25).Erlbaum, Hillsdale: NJ Educational Testing Service.

Hosmer, D., & Lemeshow, S. (2000). *Applied logistic regression*. New York: John Wiley.

Ibrahim, A. (2018). Using statistical power analysis for identifying differential item functioning in two test formats. *FUDMA Journal of Educational Foundations*, 1(2), 123-131.

Ibrahim, A. (2017). Empirical comparison of three methods for detecting differential item functioning in dichotomous test items. *Journal of Teaching and Teacher Education*, 5(1), 1-18. Retrieved from http://journals.uob.edu.bh.

Ibrahim, A. (2016). Measuring differential frequency of option response patterns in four-five

options multiple-choice test item among undergraduate students in Nigeria. *IOSR-Journal of Research and Method in Education*, 6 (2), 42 - 48. Retrieved from *www.iosrjournals.org.*

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, 65(6), 935-953.

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

Miller, T. R., & Spray, J. A. (2003). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, 30,107-122.

Miller, B. (2011). *OpenStat.* Available at http://statpages.org/miller/openstat.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Los Angeles: Sage Publications, Inc.

Pezzulo, J. (2010). *LazStats.* Available at http://statpages.org.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 58, 159–194.

Sidhu, K.S. (2012). *New approaches to measurement and evaluation*. New Delhi,India: Sterling Publishers Private Limited.

Upadhya, B., & Singhlk, Y.K. (2008). *Advanced educational psychology*. New Delhi: APH Publishing Corporation.

Zwick, R., Donoghue, J. R., & Grima, A. (2010). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233-251.