A COMPARISON OF PERFORMANCE OF TWO STATISTICS FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING IN TWO TEST FORMATS

Abdul-Wahab Ibrahim
Sule Lamido University, Kafin-Hausa, Jigawa State, Nigeria.
Email: *wahabpsychodata2017@gmail.com*
and
Eyitayo Rufus Ifedayo Afolabi
Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria.
Email: eriboyinifedayo99@gmail.com

ABSTRACT
Differential Item Functioning (DIF) is a critical step in the development and validation of educational and psychological instruments used for academic and research purposes in Nigeria. DIF analysis is a means of statistically identifying unexpected differences in performance across matched groups of examinees. The study compared the power of two statistics in detecting DIF in dichotomous and ordinal test items. These were with a view to improving the quality of test items construction. The study employed the descriptive-comparative research design. The population consisted of all undergraduate students who registered for EDU423 (Tests and Measurement) at Sule Lamido University during 2018/2019 Harmattan Semester. The sample consisted of an intact class of 513 Final Year undergraduate students who registered for the course. Thus, the entire population was therefore used, and no sampling was carried out. A null hypothesis was postulated to guide the study. Two self-developed instruments were used to collect data in the study. They were Undergraduate Students Achievement Test (USAT), and Students' Efficacy Scale (SES). Data collected were analysed using Simultaneous Item Bias Test (SIBTEST), and Logistic Discriminant Function Analysis (LDFA) statistical methods. The results showed that there was a significant difference in the proportions of test items that function differently in the dichotomous and ordinal tests when the different methods are used (SIBTEST = 950.8, $p<0.05$; LDFA = 377.5, $p<0.05$). The study concluded that the two methods complement each other in their ability to detect DIF in the two test formats. It was recommended that DIF testing must be conducted especially for very important tests like educational and psychological instruments used by various researchers in all Nigerian Universities.

Keywords: DIF,SIBTEST, LDFA, Dichotomous Test, and Ordinal Test.

INTRODUCTION

As a psychometric technique, Differential Item Functioning (DIF) has generated great interest among psychometricians and test specialists. This is because the presence of DIF in either educational or psychological instrument jeopardizes the ideal of a correct measurement procedure. An important procedure in the development of both educational and psychological instruments is ensuring that no individual or group responding to the instrument is disadvantaged in any way. For instance, DIF has an important impact on the fairness of psychological and educational testing. This is because one of the important factors which should be considered in ensuring the validity of any test is the issue of fairness. A test that shows valid differences is fair; a test that shows invalid differences is not fair. Hence, DIF forms a major threat to the fairness and validity of psychometric measures (Ibrahim, 2017a).

According to Ibrahim (2017a), DIF is a kind of invalidity that arises relative to groups. Validity is an essential requirement of all tests. A valid test produces outcomes that are based only on the trait being measured rather than irrelevant characteristics. When test scores depend on irrelevant characteristics such as group membership (that is, gender, age, social status) then the test is considered as potentially functioning differently. In Item Response Theory (IRT) framework, the DIF is defined as a difference in the conditional probabilities of answering an item correctly in two or more groups. In other words, an item is said to have DIF when its statistical/psychometrical properties vary for the groups that are matched on the attribute measured by the items (Hildago& Lopez-pina, 2010).

One way to investigate potentially malfunctioned items especially at the item level is through DIF analysis. DIF analysis is a means of statistically identifying unexpected differences in performance across matched groups of examinees. It compares the performance of matched majority (or reference) and minority (or focal) group examinees (Ibrahim, 2017b). Logistic Discriminant Function Analysis (LDFA) statistic has been one of the most widely used procedures to evaluate DIF.LDFA 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 (Osterlind & Everson, 2009). 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; Ibrahim, 2017b).

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:

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

Like the LDFA procedure, Simultaneous Item Bias Test (SIBTEST) 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 LDFA 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) (Holland &Wainer, 2009).

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, 2017b). 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:

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

At a given ability (θ), this difference is expressed as: $B(\theta) = P_R(\theta) - P_F(\theta)$ *(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 $T_F(s)$ (Osterlind & Everson, 2009; Ibrahim, 2018).

In this study, the two test formats studied were dichotomous and ordinal tests. Traditionally, dichotomous tests have been the most widely used item format in educational achievement tests. For many dichotomous tests, items are scored dichotomously (that is, correct or incorrect). Whereas, ordinal test items are those that present more than two response categories; a prototypical case would be Likert-type items. Suffice to say that in most Nigerian Universities, little attention is given to the presence of DIF in dichotomous and ordinal test items used for academic and research purposes. In a bid to ensure that tests are fair for all respondents, Lecturers in the Nigerian Universities have a formal review, which is part of the test development process, where items are screened by content specialists for research instruments that might be inappropriate or unfair to relevant examinees in the test-taking population. Thus, apart from the traditional professional judgment used by Lecturers in Nigerian Universities, the use of statistical methods to detect test items that function differentially for individuals from

different groups but with the same ability, will be more accurate and efficient. However, there is little agreement on which DIF statistical procedure is most accurate for dichotomous and ordinal tests. There is no doubt that systematic application of DIF detection methods followed by expert review leads to action to correct biased items, and these form vital steps in test development and validation. This enhances both validity and equity in testing. Against this background, this study empirically compared the relative ability of the two statistical methods (SIBTEST and LDFA) for detecting DIF in both dichotomous and ordinal test items.

Aim and Objective of the Study

The main aim of this study is to empirically compare the relative performance of SIBTEST and LDFA for detecting DIF in both dichotomous and ordinal test items. Towards this end, the specific objective of the study appears germane namely to:

i. determine whether difference exists between the proportions of test items that function differently in the dichotomous and ordinal tests when the different methods are used.

Research Hypothesis

To achieve the objective of the study, a null hypothesis was postulated as follows:

i. There is no significant difference between the proportions of test items that function differentially in the dichotomous and ordinal tests when the different methods (SIBTEST and LDFA) are used.

METHODOLOGY

Research Design

This study used descriptive-comparative research design. According to Upadhya and Singh (2008), descriptive-comparative research design is a type of research design that explains phenomenon by collecting numerical data that are analysed using mathematically based methods. In carrying out this study, therefore, the researcher collected data from subset of the population (that is, Final Year undergraduate students) in such a way the knowledge to be gained is representative of the total under study. Most importantly, the researcher used the data collected to explore the two statistical DIF detection methods being studied in this study.

Population

All undergraduate students who registered for a compulsory course in Tests and Measurement during the Second of 2018/2019 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 513 400 level undergraduate students who registered for the course. The rationale for chosen 400 level undergraduate students was because at their educational level in the University, all undergraduate students of different majors must register for the course to be allowed to earn a bachelor's degree in Education. Thus, the entire population was therefore used, and no sampling was carried out. Also, no sampling technique was used in the study because the sample size in DIF analysis cannot be less than 500 participants (Kristjansson, Aylesworth, McDowell &Zumbo, 2005).

As a DIF procedure, the sex of the subjects was used to stream the subjects into two distinct groups namely reference and focal groups. The male undergraduate students were considered as the reference group, and the female undergraduate students were taken as the focal group of the study. Hence, two groups: reference and focal groups' combination were used in the Differential Item Functioning analysis. Collapsing the participants into two distinct groups was a criterion stipulated in the Item Response Theory (IRT) as a theoretical background for the analysis of DIF (Holland &Wainer, 2009; Ibrahim, 2018).

Research Instruments

Two research instruments were used in the study namely Undergraduate Students Achievement Test (USAT), and Students' Efficacy Scale (SES). The first instrument, USAT, was a self-developed dichotomous instrument which consists of a 20, 4-option multiple-choice test that was developed using the course (EDU 423: Tests and Measurement) content. The second instrument, SES, was developed by the researcher as an ordinal instrument which consists of 24-item which is made up of six subscales namely: Students' Learning; Discipline; Parental Involvement; Testing; Reaching the Poor in the Class; and Overcoming Work Challenges. 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).

Validity of the Research Instruments

The face, construct and content validity of the instruments was established using expert judgments. 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.

Reliability of the Research Instruments

Thereafter, a reliability process was done to establish how reliable the instruments are. Hence, reliability test was conducted on the whole data collected for pilot testing using the Cronbach's Alpha and Split-Half reliability methods. The Cronbach's Alpha and Split-Half reliability methods were preferred because of the desire to determine the internal consistency of the instruments for data collection. 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 and Split-Half Coefficient values obtained were 0.76 and 0.89 respectively. Its mean difficulty index is 0.70 with a standard deviation of 0.28. The item discrimination indices have a mean 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.

Also, the Students' Efficacy Scale (SES) has good reliability: Cronbach's Alpha reliability coefficients for: Students' Learning of 0.68; Discipline was 0.64; Parental Involvement was 0.63; Testing was 0.65; Reaching the Poor in the Class was 0.60; and Overcoming Work Challenges was 0.59. Also, the reliability of the whole test included Cronbach's Alpha coefficient of 0.63; and Split-Half coefficient of 0.67 respectively. Consequently, the instruments were accepted being stable over time, hence their usage in this study.

Procedure for Data Collection

Copies of the instruments were administered on the students with the assistance of the three course Lecturers of EDU 423, as well as two Assistant Lecturers in the Department of Education of Sule Lamido University, Kafin Hausa, Jigawa State, Nigeria. The instrument administration was conducted under strict but friendly condition. Enough 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. This 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 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 two statistical procedures. Updated SPSS version 24.0 and Microsoft Excel version 12.0 were used to manage and organize the datasets.

RESULTS

To better understand the direction of DIF magnitude, items that manifested DIF were classified into three categories representing different magnitudes of DIF guidelines proposed by Hildago& Lopez-pina, 2010):

i. Negligible or A-level DIF: Null hypothesis was rejected and $|\beta| < 0.059$

ii. Moderate or B-level DIF: Null hypothesis was rejected and $0.059 \leq |\beta| < 0.088$

iii. Large or C-level DIF: Null hypothesis was rejected and $|\beta| \geq 0.088$.

The results of the classification analysis of the data are presented in Table 1 for direction and magnitude of DIF items in dichotomous test that favour the reference (b_R) and focal (b_F) groups, and Table 2 for direction and magnitude of DIF items in ordinal test that favour the reference (b_R) and focal (b_F) groups.

Table 1:

Proportion of Test Items that Function Differentially in the Dichotomous Test Using SIBTEST and LDFA Methods

	Reference Group		**Focal Group**		***p-values difference***	
Items	**SIBTEST** ***p-values***	**LDFA** ***p-values***	**SIBTEST** ***p-values***	**LDFA** ***p-values***	**SIBTEST** ***β***	**LDFA** ***β***
1	.28(0.02)*	.26(-0.02)	.33(-0.05)	.38(-0.01)	-0.03	-0.12
2	.33(-0.14)*	.47(-0.36)	.48(0.01)	47(-0.12)	-0.15	0.00
3	.18(-0.17)*	.35(-0.15)	.36(-0.08)	.44(-0.04)*	-0.18	0.09
4.	.36.(0.03)	.33(-0.09)	.22(-0.06)	.28(0.02)	0.14	0.05
5.	.30(0.05)	.25(-0.14)	.28(0.02)	.26(-0.11)	0.02	-0.01
6	.46(0.04)	.42(-0.09)	.56(0.23)	.33(0.19)	-0.01	0.09
7	.58(-0.09)	.67(0.01)	.65(0.07)	.58(0.12)	-0.07	0.09

****Significant, p < .05***

Table 1 presents the proportions of test items (*p-values*) that function differentially in the dichotomous test when the SIBTEST and LDFA methods are used. In the 20-item test, SIBTEST flagged six items as proportionately functioning differently in the dichotomous test for the reference and focal groups respectively. SIBTEST identified such items as items 11, 12, 14, 15, 17, and 18. Also, LDFA flagged four items being items 13, 15, 18, and 19 proportionately functioning differently in the dichotomous test for the reference and focal groups respectively. Comparatively, this implies that SIBTEST is a very promising procedure for detecting proportion of DIF items in dichotomous test but the LDFA 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 two 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 LDFA result 0.05, $p < .05$, which shows that item 1, was identified as proportionately functioning differently when the

two 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 3, 16, 18, and 20 were flagged by LDFA as proportionately functioning differently for both reference and focal groups.

Table 2:

Proportion of Test Items that Function Differentially in the Ordinal Test Using SIBTEST and LDFA Methods

	Reference Group		**Focal Group**		***p-values difference***	
Items	**SIBTEST** ***p-values***	**LDFA** ***p-values***	**SIBTEST** ***p-values***	**LDFA** ***p-values***	**SIBTEST** ***β***	**LDFA** ***β***
1	.98(0.29)*	.69(-0.03)	.73(0.05)	.68(0.01)	0.25	0.05
2	.96(-0.04)	1.00(-0.67)	.98(0.01)	.97(-0.62)	-0.02	0.03
3	.98(-0.02)	1.00(-0.4)	.67(0.03)	.64(0.04)*	0.31	0.36
4	1.00(0.17)	.83(-0.59)	.92(0.14)	.78(-0.08)	0.08	0.05
5	1.00(0.25)	.75(-0.64)	.88(0.20)	.68(0.32)	0.12	0.07
6	.50(0.00)	.50(-0.13)	.36(-0.17)	.53(-0.31)	0.14	-0.03
7	.09(-0.91)	1.00(-0.53)	1.00(0.17)	.83(-0.53)	-0.91	0.17
8	.31(-0.69)	1.00(-0.42)	.80(-0.18)	.98(-0.52)	0.23	0.02
9	.36(-0.65)	1.00(-0.41)	.30(-0.60)	.90(-0.24)	0.06	0.10
10	.25(-0.75)	1.00(-0.32)	.26((-0.74)	1.00(-0.32)	-0.01	0.00
11	.20(-0.23)	.43(0.21)	.57(0.15)*	.42(0.58)	-0.37	0.01
12	.44(-0.21)	.64(0.09)	.43(-0.15)	.58(0.13)	0.01	0.06
13	.75(0.21)	.54(0.22)	.67(0.14)	.53(-0.02)	0.08	0.01
14	.29(-0.44)	.73(0.06)	.20(-0.49)	.69(0.31)	0.27	0.04
15	.67(-0.02)	.69(0.19)	.51(-0.06)	.57(0.26)	0.16	0.12
16	.53(-0.15)	.67(0.15)	.47(-0.53)	1.00(-0.02)	0.06	-0.33
17	.33(-0.29)	.62(0.22)	.70(0.00)	.70(0.20)	-0.37	-0.08

18	.77(0.00)	.77(0.00)	.75(0.22)	.53(0.17)	0.02	0.24
19	.77(0.27)	.50(0.34)	.55(-0.09)	.64(0.01)	0.22	-0.14
20	.40(-0.05)	.45(0.39)*	.76(0.11)	.65(0.22)	-0.36	-0.20
21	.75(0.09)	.66(0.21)	.80(0.25)	.55(0.35)	-0.05	0.11
22	.75(0.26)	.49(0.44)	.53(0.15)	.38(0.07)	0.22	0.11
23	.50(-0.08)	.58(0.35)	.46(-0.07)	.53(.0.34)	0.04	0.05
24	.90(0.25)	.65(0.29)	1.00(0.53)	.47(0.53)	-0.10	0.18

Significant, p < .05

Further, Table 3 presents the results of the SIBTEST and LDFA analysis. From Table 3, 99% of the proportion of items functioning differentially in ordinal test flagged negligible DIF, when SIBTEST was used as compared with 48% of the items flagged as containing negligible DIF in dichotomous test. Also, SIBTEST flagged 28% of items as Moderate DIF in ordinal test and 17% of the items in dichotomous test flagged as moderate DIF. Similarly, LDFA flagged 15% of the ordinal items as large DIF, as well as 79% of the items flagged large DIF in dichotomous test.

Table 3:

Difference between the Proportion of Test Items That Function Differentially in the Dichotomous and Ordinal Tests

	DIF Classification					
	Dichotomous Items			**Ordinal Items**		
Variables	**A**	**B**	**C**	**A**	**B**	**C**
Negligible or A-level DIF	48.64	23.16	28.18	99.21	0.79	-
Moderate or B-level DIF	27.25	17.98	54.72	5276	28.46	18.76
Large or C-level DIF	79.73	15. 26	5.00	72.16	12.69	15.14
Group Ability Differences			**SIBTEST**		**LDFA**	

Ability differences x item discrimination		
Equal x Low	0.049	0.088
Unequal x Low	1.000	0.999
Equal x Moderate	0.854	0.049
Unequal x Moderate	0.533	0.048
Equal x High	0.497	0.040
Unequal x High	0.269	0.052
Average	0.689	0.263
SIBTEST	950.8	$\alpha = 0.05$
LDFA	377.5	$\alpha = 0.05$

Significant, p < .05

Similarly, Table 3 reveals that the two procedures varied greatly in their ability to detect DIF in dichotomous and ordinal test items. The SIBTEST and LDFA had nearly perfect power for detecting DIF at 1.000, $p < .05$, and 0.999, $p < .05$, respectively. However, the LDFA performed rather poorly in detecting DIF at all (0.263, $p < .05$). To determine whether significant difference exists in the performance of SIBTEST, and LDFA methods for detecting DIF in dichotomous and ordinal test items, the two procedures yielded values of: SIBTEST = 950.8, $p<.05$, LDFA = 377.5, $p<.05$, which are significant. These results suggested there was a significant difference in the performance of SIBTEST and LDFA methods for detecting DIF in dichotomous and ordinal test items. Thus, the null hypothesis was disconfirmed.

DISCUSSION

The finding of this study indicated that there was a significant difference between the proportions 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 Miller and Spray (2003) and DeAyala(2012)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.

The researcher believes that the results of this study can bear more significance by taking one point into account. LDFA and SIBTEST are parametric DIF detection approaches, which are a response to the previous DIF techniques which could only screen uniform DIF such as Standardization, GMH, to mention a few. This, implicitly, can be considered as a reassuring point for the developers of the dichotomous and ordinal tests.

CONCLUSION

On the strength of the findings obtained from the study, it can be concluded, therefore, that the two 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.

RECOMMENDATIONS

From the findings of this study, the following recommendations were made:

1. Statistical methods for detecting DIF should be an essential part of test development and test evaluation efforts.
2. Quantitative and qualitative analyses that can inform the test development process should be conducted after the administration of a test.
3. DIF testing must be conducted especially for very important tests like educational and psychological instruments used by various researchers in all Nigerian Universities.

REFERENCES

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

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

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.http://journal.fudutsinma.edu.ng/index.php/fujef.

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

Ibrahim, A. (2017b). Relative Effectiveness of Generalized Mantel-Haenszel, Simultaneous Item Bias Test and Logistic Discriminant Function Analysis for Detecting Differential Item Functioning in Ordinal Test Items. *Ife Psychologia*, 25(1), 104-132. www.ifepsychologia.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.

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.