**Journal of Teaching and Teacher Education**
ISSN (2210-1578)
J. Tea. Tea. Edu. 5, No. 1 (Jan-2017)

# Empirical Comparison of Three Methods for Detecting Differential Item Functioning in Dichotomous Test Items

**Abdul-Wahab Ibrahim**

*Department of Education, Faculty of Education, Sule Lamido University, Kafin Hausa, Jigawa State, Nigeria*

*Received 29Apr.2016, Revised 07Nov. 2016, Accepted 17 Nov.2016, Published 01Jan. 2017*

**Abstract:** The study classified individual items that function differentially according to the magnitude of DIF in dichotomous test and compared the power of Generalized Mantel Haenszel (GMH), Simultaneous Item Bias Test (SIBTEST), and Logistic Discriminant Function Analysis (LDFA) methods in detecting DIF in dichotomous test items. It also determined the relationship between the proportions of test items that function differentially in dichotomous test items when the different methods were used. These were with a view to improving the quality of dichotomous test items construction. The study adopted a survey design. The population consisted of all undergraduate students who registered for EFC 303 (Tests and Measurement) at Obafemi Awolowo University during 2011/2012 Harmattan Semester. The sample consisted of an intact class of 457 Part 3 undergraduate students who registered for the course. A 50-item multiple -choice achievement test designated "Undergraduate Students' Achievement Test (USAT)" was developed by the researcher for the study. A total of 445 scripts were found properly completed. Data collected were analysed using Generalized Mantel Haenszel, Simultaneous Item Bias Test, and Logistic Discriminant Function Analysis. The results showed that there was a significant difference in the performance of the GMH, SIBTEST, and LDFA methods in detecting DIF in dichotomous test items. Further, a non significant relationship existed between the proportions of test items that functioned differentially in the dichotomous test items when the different methods were used. The study concluded that GMH, SIBTEST and LDFA were effective in detecting DIF across dichotomous test items but complement one other in their ability to detect DIF in dichotomous test items.

**Keywords:** Differential item functioning, dichotomous test, magnitude, proportions of test items, psychometrics and test-takers

## 1. INTRODUCTION

In the last two decades, the study of Differential Item Functioning (DIF) has attracted much attention in psychometrics, partly because of its ability to ensure the metric equivalence of measurement instruments, and improve the validity of tests and questionnaire. DIF statistical techniques are based on the principle that if different groups of test-takers (e.g., males and females) have roughly the same level of knowledge, then they should perform similarly on individual test items regardless of group membership. In essence, all DIF techniques match test -takers from different groups according to their total test scores and then investigate how the different groups performed on individual test items to determine whether the test items are creating problems for a particular group. Thus, as a test validation process, DIF analyses do not only ensure that item content is appropriate but help test developers identify items that function differentially between two groups of examinees since the ultimate criterion of item equivalence must come from an analysis of the examinee responses (Ibrahim, 2013).

As a psychometric technique, DIF was developed to counteract test item bias, and DIF tells whether a particular test item functions differently to different groups. Based on Item Response Theory (IRT), DIF equips the instrument developer to become aware of situations where examinees of the same ability but from different groups have different probabilities of success on an item. It is expected under equivalent testing conditions, that individuals from different groups (but with comparable ability levels) exhibit similar probability of responding correctly to a given item. Therefore, DIF represents a modern psychometric approach to the investigation of between-group score discrepancies. An advantage of DIF over Classical Test Theory (CTT) is that reliability is not constrained to a single co-efficient, but instead can be measured continuously over the entire ability spectrum- continuum of variation. Therefore, total

*E-mail address: edutestsexpert2013@gmail.com*

score is used as a reference to classify testees into high and low ability groups (Ibrahim, 2013).

One of the pioneering methods used to detect DIF is known as the Generalized Mantel-Haenszel procedure (GMH) (Mantel &Haenszel, 1959). 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 (Stephen-Bounty, 2005).

According to Guilera, Gomez and Hildago (2009), the GMH procedure has been widely used to detect DIF because it is conceptually simple, relatively easy to apply, offers a test of statistical significance and provides an estimate of the effect size used on the common odds ratio. Furthermore, the GMH statistic can be calculated using easily accessible statistical software, including Statistical Package for Social Sciences, OpenStat, and LazStats. However, it is not efficient for non-uniform DIF (Osterlind and Everson, 2009), but Magis, Beland, Tuerlinckx&DeBoeck (2010), proposed a variation that reduces this limitation. Uniform DIF occurs when an item is more difficult at all ability levels for one group than for the other. Non-uniform DIF occurs when there is an interaction between ability level and group, so that the item is more difficult, for example, for one group at lower levels of ability but more difficult for the other at higher levels of ability. Also, large sample sizes are not required for GMH statistical procedure and Hollabdm and Thayer (1998) reported high power and good control to the Type 1 Error in samples of 100 subjects per group. Probably, these are the reasons why the GMH is top-ranking in the detection of items with differential functioning.

Similar to the GMH procedure, Simultaneous Item Bias Test (SIBTEST) proposed by Shealy and 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. The matching criterion is a latent measure rather than the observed total score used by the GMH procedure. Estimation of the matching latent trait includes a regression-based correction that has been shown to be useful in controlling Type 1 error – the proportion of times that a DIF item was falsely rejected at the 0.05 level. SIBTEST was originally intended for use with dichotomous test items, but has since been extended to handle ordered items (Gierl, Jodoin, & Ackerman, 2000). 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).

As with SIBTEST and the Generalized Mantel-Haenszel procedures, Logistic Discriminant Function Analysis (LDFA), proposed by Miller & Spray (1993), 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 (Finch & French, 2008). 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).

According to Kristjansson, Aylesworth, McDowell &Zumbo (2005), 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.

Traditionally, dichotomous tests have been the most widely used item format in educational achievement tests. For the vast majority of dichotomous tests, items are scored dichotomously (i.e. correct or incorrect). In the same vein, many current psychometric procedures were developed primarily for analysis with dichotomously scored test items, and they do not easily accommodate ordered scored item responses. As a corollary to the above, this study will empirically compare the relative ability of the three statistical methods for detecting Differential Item Functioning in dichotomous test items.

The failure of judgmental methods to provide a satisfactory means of screening test items for differential difficulty gave impetus to the development of statistical DIF procedures. According to Osterlind and Everson (2009), efforts to use judges only to identify differential difficulty are virtually at a dead end. Given what is generally known about the subtle and complex nature of mismeasurement, it is likely that statistical indices and logical analyses are needed jointly to make inferences about DIF. Because some DIF studies have produced highly interpretable bias results, it is assumed that in

some situations, judges could do much better than chance in detecting Differential Item Functioning; but given a negative track record for expert judgments, it is unlikely that one can rely entirely on judgmental review. Statistical indices produce a high dross rate – studies in which DIF flags are unreliable or uninterpretable, or in which they signal a valid secondary trait. Therefore, both statistical and judgmental approaches are needed as checks on each other. Against this backdrop, the study will provide information on suitability and easy of use of the three methods. It will also provide information on basis for the comparison of the effectiveness of methods in detecting DIF in dichotomous test items. Researchers and item developers will be provided with information on how and when to use each of the methods. Furthermore, the findings of this study will enable the examination bodies and test constructors to make better decisions with regard to the presence or absence of DIF. Not only this, the combination of a test with an effect size measure will reduce false identification rates of Differential Item Functioning.

According to Ibrahim (2013), several methods have been proposed for the detection of Differential Item Functioning. These methods include Analysis of Variance (ANOVA), Correlational Methods, Lord's Chi-square Test, Logistic Regression, Exact Test, Simultaneous Item Bias Test, Mantel-Haenszel Statistic, Standardization Statistic, Unconstrained Cumulative Logits Ordinal Logistic Regression, Logistic Discriminant Function Analysis, Likelihood Ratio Test, and a host of others. However, identifying the method that is most effective is vital in the current testing climate. Based on theoretical strengths and the results of numerous empirical comparisons, a relatively small number of these methods have emerged as preferred. These are Standardization, Logistic Regression, Exact Test, Generalized Mantel-Haenszel, Simultaneous Item Bias Test, and Logistic Discriminant Function Analysis. All of these approaches provide for comparison of performance in a studied item after matching examinees on the ability of interest. With these methods, however, there is not yet a consensus about how to test DIF when item responses are dichotomously scored. With dichotomous item responses, the most widely used DIF detection methods are procedures based on Item Response Theory (IRT). These methods have been useful in detecting DIF in dichotomous item responses. Several extensions of the dichotomous DIF procedures have been proposed for use with ordinal 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 dichotomous items have not received the thorough and rigorous study accorded to other tests DIF, thus necessitating further research to investigate their performance before they are ready for routine operational use (Ibrahim, 2013).

In a bid to ensure that tests are fair for all examinees, examination bodies in Nigeria such as West African Examinations Council (WAECC), National Examinations Council (NECO), and the Joint Admission and Matriculation Board (JAMB), have a formal review, which is part of the test development process, where items are screened by content specialists for tests that might be inappropriate or unfair to relevant examinees in the test-taking population. This judgmental procedure of identifying items that function differentially for individuals with comparable ability but from different groups is employed by these examination bodies rather than applying statistical analyses to detect the items in a test that function differentially for individuals with the same ability from different groups. Thus, apart from the traditional professional judgment used by examination bodies in Nigeria, 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. Also, in most Nigerian Universities, little attention is given to the presence of DIF in dichotomous test items used for academic and research purposes. However, there is little agreement on which DIF statistical procedure is most accurate for dichotomous tests (Ibrahim, 2013). Hence, this study compares the relative performance of three statistical methods for detecting DIF in dichotomous items. Towards this end, the following specific objectives of the study appear germane namely to: (i) classify individual items that function differentially according to the magnitude of DIF in dichotomous test item; (ii) compare the performance of GMH, SIBTEST, and LDFA methods for detecting DIF in dichotomous test items; (iii) determine the relationship between the proportions of test items that function differently in the dichotomous test when the different methods are used. To achieve the objectives of the study, the following research question and research hypotheses were raised:

**Research Question**

1. What is the classification level for each item that functions differentially based on the magnitude of DIF detected in dichotomous test items?

**Research Hypotheses**

1. There is no significant difference in the performance of the GMH, SIBTEST, and LDFA methods for detecting DIF in dichotomous test items.

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

## 2. METHODOLOGY

The descriptive survey research design was used in this study. According to Upadhya and Singh (2008), descriptive survey research design is a type of research design that is explaining phenomena by collecting numerical data that are analyzed using mathematically based methods. 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 (Zumbo and Hubley, 2003); 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.

All undergraduate students who registered for a compulsory course in Tests and Measurement during the Harmattan Semester of 2011/2012 Session in the Faculty of Education of the Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria, constituted the target population for the study. There were 457 undergraduate students who registered for the course during the session. The sample consisted of an intact class of 457 part 3 undergraduate students who registered for EFC 303 in Harmattan Semester of 2011/2012 session. Thus, the entire population was therefore used and no sampling was carried out. As a DIF procedure, the sex of the subjects was used to stream the subjects into two distinct groups namely reference and focal groups without gender bias. The male undergraduate students were considered as the reference group, and the female undergraduate students were taken as the focal group of the study. Also, the choice of the EFC 303 is appropriate in this study due to its Faculty status as a compulsory course aimed at introducing the basic procedures involved in educational testing as well as the procedures involved in the development, administration, scoring of test items and reporting of test scores of acceptable standard. Also, as a Faculty course, all undergraduate students of different majors must register for the course to be allowed to earn a Bachelor Degree in Education.

A 50-item multiple -choice achievement test designated "Undergraduate Students' Achievement Test (USAT)" (Ibrahim, 2013), was developed by the researcher for the study and it contains dichotomous items. The instrument is a 50-4 option multiple-choice test that was developed using the course (EFC 303: Tests and Measurement) content. In multiple-choice test, using more options make a test item proportionately more reliable, as the chance factor in an item is reduced as the number of options is increased. Hence, the chance of answering a 4-choice item correctly just by guessing is one-fourth or 25 per cent, while for a 3-choice item, it is one-third or 33 per cent, and for a 5-chance is one-fifth or 20 per cent respectively (Sidhu, 2005; Ivowi, 2009). In order to resolve this issue, the number of options was maintained at four (4) to give the probability of selection of a key (correct answer) by mere guess as one-fourth or 25 per cent. By so doing, this has reduced the burden (difficulties) on the researcher in providing options for items in the study. Test items are all multiple-choice items and consisted of an incomplete statement which the examinee could complete correctly by selecting one of the four phrases following it. Below is the content of the EFC 303: Tests and Measurement course outline: (a) definition of basic concepts such as test, measurement, evaluation, and assessment; (b) types of tests; (c) basic qualities of a good test(validity & reliability); (d) general principles of test construction; (e) test item analysis; (f) test administration; (g) test scoring and test reporting; (h) measure of dispersion and derived scores ; and (i) use of bivariate analysis in the measurement relationships and statistical differences.

The content and construct validity of the instrument (Undergraduate Students' Achievement Test - USAT) 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 validation 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 17.0. The instrument was pilot tested using 60 part three students in the Faculty of Education, University of Lagos, Akoka-Yaba, Lagos, 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(x) difficulty index is 0.70 with a standard deviation of 0.28. The item discrimination indices have a mean(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 Sidhu (2005), 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.

The instrument was administered by the researcher. The hard copies of the test were administered on the students with the assistance of the course lecturers of EFC 303, as well as a handful of some graduate students in the Department of Educational Foundations and Counselling of the ObafemiAwolowo University, Ile-Ife, Osun State, Nigeria. The test administration was conducted under strict examination condition. However, adequate time was provided for testees to respond to all the items. Furthermore, the testees were instructed not to omit any item as it is mandatory to answer all items in the test as they marked on their answer sheets that alternative which they have decided is most correct. Such a procedure provided a uniform response set thereby minimizing individual differences in responding. At the end of testing time, test copies were collected immediately.

A total of 457 copies of the instrument were administered, while 445 copies were finally collected on return, as being properly completed and were used for 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 17.0 and Microsoft Excel version 12.0 were used to manage and organize the datasets.

## 3. RESULTS

**Research Question:** What is the classification level for each item that functions differentially based on the magnitude of DIF detected in dichotomous test items? To answer this research question, item parameter analysis was carried out on each of the items, following the procedure by Narayanan and Swaminathan (1994). Hence, the three-parameter logistic item response model (3PLM IRT) was used for the generation of examinee response data which necessitated the stipulation of item parameters for the dichotomous test items. According to IRT theory, the one-parameter logistic item response model (1PLM IRT) is used to test only the test items discrimination parameters (*a*-parameters), the two-parameter logistic item response model (2PLM IRT) is used to test both the test items' discrimination and difficulty parameters (*a*-parameters, *b*-parameters), while the three-parameter logistic item response model (3PLM IRT) includes item-level parameters of the items' discrimination (*a*-parameters), difficulty (*b*-parameters), and susceptibility to guessing (*c*-parameters). The same item parameters were used for both the reference and focal groups resulting in unbiased items that were expected to reflect realistic items that were free of DIF. These item parameters are shown in Table 1 for dichotomous test items.

**Table 1. Item Parameters for Dichotomous Test Items**

| Items | Parameters | | | Items | Parameters | | |
|---|---|---|---|---|---|---|---|
| | *a* | *b* | *c* | | *a* | *b* | *c* |
| 1. | 0.44 | 0.30 | 0.30 | 26. | 0.62 | 0.64 | 2.00 |
| 2. | 0.55 | 0.76 | 0.10 | 27. | 0.48 | 2.12 | 2.25 |
| 3. | 0.82 | 0.32 | 0.20 | 28. | 0.55 | 0.91 | 1.29 |
| 4. | 0.52 | 0.60 | 0.60 | 29. | 0.53 | 0.87 | 1.22 |
| 5. | 0.62 | 0.28 | 2.40 | 30. | 0.36 | 0.36 | 0.60 |
| 6. | 0.82 | 0.61 | 1.28 | 31. | 0.32 | 0.21 | 0.10 |
| 7. | 0.92 | 0.42 | 0.50 | 32. | 0.86 | 0.57 | 2.27 |
| 8. | 0.65 | 0.68 | 2.20 | 33. | 0.59 | 0.29 | 1.11 |
| 9. | 0.56 | 2.70 | 1.20 | 34. | 0.56 | 0.40 | 0.24 |
| 10. | 0.29 | 0.39 | 1.10 | 35. | 0.69 | 0.41 | 2.26 |
| 11. | 0.35 | 0.32 | 2.30 | 36. | 0.88 | 0.96 | 0.04 |
| 12 | 0.31 | 0.37 | 0.42 | 37. | 0.95 | 0.68 | 3.10 |
| 13. | 0.55 | 0.30 | 0.28 | 38. | 0.48 | 0.67 | 0.21 |
| 14. | 0.51 | 0.69 | 0.31 | 39. | 0.43 | 0.64 | 0.26 |
| 15. | 0.73 | 0.61 | 0.62 | 40. | 0.73 | 0.77 | 0.24 |
| 16. | 0.88 | 0.95 | 0.24 | 41. | 0.47 | 0.63 | 0.19 |
| 17. | 0.31 | 0.35 | 2.00 | 42. | 0.51 | 0.69 | 0.29 |
| 18. | 0.32 | 0.57 | 1.32 | 43. | 0.29 | 0.92 | 0.13 |
| 19. | 0.55 | 0.19 | 0.80 | 44. | 0.61 | 0.93 | 0.12 |
| 20. | 0.40 | 0.64 | 0.76 | 45. | 0.49 | 0.73 | 0.63 |
| 21. | 0.92 | 0.13 | 1.72 | 46. | 0.45 | 0.81 | 0.22 |
| 22. | 0.64 | 0.55 | 2.07 | 47. | 0.54 | 0.67 | 0.21 |
| 23. | 0.61 | 0.81 | 1.12 | 48. | 0.72 | 0.77 | 0.60 |
| 24. | 0.61 | 0.53 | 0.52 | 49. | 0.75 | 0.82 | 0.34 |
| 25. | 0.70 | 1.05 | 0.26 | 50. | 0.92 | 0.67 | 0.64 |

**Key: *a* = Item Discrimination, *b* = Item Difficulty, and *c* = Pseudo-guessing**

Table 1 reveals item parameters for the reference and focal groups with different guessing levels for the two groups. As shown in Table 1, item 37 exhibited highest discrimination (0.95) level for both the reference and focal groups, with difficulty index of 0.68; the same item suggests a highest probability guessing level (3.10) for the two groups. Similarly, item 36 shows a highest difficulty parameter value (0.96) with a higher discrimination level (0.88) and the least guessing level (0.04) for the two groups. This is followed by items 12 and 17 with the same discrimination power (0.31 each) but different difficulty values (0.37; 0.35) and different guessing levels (0.42; 2.00) respectively. Further, upon completion of the analysis, items that manifested DIF were classified into three categories representing different magnitudes of DIF guidelines proposed by Roussos and Stout (1996) adopted by Gierl et al., (2000):

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

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

3. 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 2 for direction and magnitude of DIF items in dichotomous test that favour the reference ($b_R$) and focal ($b_F$) groups.

**Table 2. Classification of Magnitude of DIF Items in Dichotomous Test Items that Favour the Reference ($b_R$) and Focal ($b_F$) Groups**

| DIF Magnitude | Methods | | | | | |
|---|---|---|---|---|---|---|
| | Items | GMH *B* | Items | LDFA *B* | Items | SIBTEST *B* |
| Negligible (A- level DIF) | 2 | 0.027 | 1 | 0.029 | 3 | 0.042 |
| | 5 | -0.042 | 4 | 0.026 | 7 | 0.009 |
| | 6 | -0.005 | 9 | -0.008 | 13 | 0.045 |
| | 11 | 0.001 | 10 | 0.032 | 15 | -0.037 |
| | 21 | 0.032 | 12 | -0.028 | 18 | -0.010 |
| | | | 22 | 0.002 | | |
| | | | 24 | -0.018 | | |
| | | | 25 | 0.005 | | |
| Moderate (B-level DIF) | 29 | 0.072 | 31 | -0.075 | 26 | 0.069 |
| | 30 | 0.074 | 33 | 0.078 | 27 | 0.068 |
| | 40 | -0.070 | 35 | 0.072 | 28 | 0.071 |
| | 41 | -0.073 | 36 | -0.069 | 32 | -0.063 |
| | 42 | -0.064 | 37 | -0.074 | | |
| | | | 39 | 0.077 | | |
| Large (C-level DIF) | 44 | 0.096 | | | 43 | -0.097 |
| | 48 | 0.110 | | | 45 | -0.119 |
| | 49 | 0.094 | | | 46 | 0.163 |

Table 2 displays the magnitude of DIF detected in dichotomous test and classification level for each item. Using a critical *p-value* of 0.05, 18 items exhibit negligible (A- level) DIF. These are items 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 15, 18, 21, 22, 24, and 25. Further, 15 items exhibit moderate (B-level) DIF, these are items 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 39, 40, 41, and 42. Also, 6 items exhibit large (C-level) DIF, these are items 43, 44, 45, 46, 48, and 49.

Using GMH, items 2, 5, 6, 11, and 21 were classified as having negligible DIF (A-level), with items 2, 11, and 21 favouring reference group, and items 5, and 6 favouring focal group respectively. Also, GMH classified items 29, 30, 40, 41, and 42 as having moderate (B-level) DIF, with items 40, 41, and 42 favouring focal group, and both items 29 and 30 favouring reference group. Similarly, GMH classified items 44, 48, and 49 as large (C-level) DIF, all favouring the reference group only. On the other hand, when LDFA was used, items 1, 4, 9, 10, 12, 22, 24, and 25 were classified as having negligible (A-level) DIF, with items 9, 12, and 24 favouring focal group, while items 1, 4, 10, 22, and 25 favour reference group. In the same vein, LDFA classified items 31, 33, 35, 36, 37, and 39 as having moderate (B-level) DIF, with items 31, 36, and 37 favouring focal group, but the trio of items 33, 35 and 39 favouring reference group respectively. Noteworthy, LDFA did not classified any items as having large (C-level) DIF. Further, SIBTEST classified items 3, 7, 13, 15, and 18 as having negligible (A- level) DIF, with both items 15 and 18 favouring focal group, while items 3, 7, and 13 favouring reference group. Also, with SIBTEST, items 26, 27, 28, and 32 were classified as having moderate (B-level) DIF, while item 32 only favouring focal group, items 26, 27, and 28 favouring reference group. More so, SIBTEST classified

items 43, 45, and 46 as having large (C-level) DIF, with both items 43 and 45 favouring focal group, and item 46 only favouring the reference group respectively. Consequently, both GMH and SIBTEST displayed good but efficient power in classifying item that functions differentially based on the magnitude of DIF detected in dichotomous test items than LDFA does.

**Hypothesis One:** There is no significant difference in the performance of the GMH, SIBTEST, and LDFA methods for detecting DIF in dichotomous test items. To test this hypothesis, the items flagged as having statistically significant DIF were then categorized as having either negligible (category A), intermediate (category B) or large (category C) DIF. The proportions of these statistically significant DIF items that fell in the three categories are presented in Table 3 for each of the three DIF detection procedures.

**Table 3. Difference in the Performance of the GMH, SIBTEST, and LDFA Methods for Detecting DIF in Dichotomous Test Items**

| Variables | Methods | | |
|---|---|---|---|
| | GMH | LDFA | SIBTEST |
| Group ability differences | | | |
| No differences both = (N (0, 1)) | 0.998 | 0.991 | 0.970* |
| Unequal: reference group = (N(-5, 1)) | | | |
| Focal = (N(-0, 5, 1)) | 0.978 | 0.987 | 0.959 |
| Studied item discrimination | | | |
| Low (0.8) | 0.951 | 0.970* | 0.051 |
| Moderate (1.2) | 0.968 | 0.999 | 0.048 |
| High (1.6) | 0.999* | 1.000 | 0.046 |
| Sample size ratio | | | |
| 1:1 | 0.994 | 0.997* | 0.985 |
| 4:1 | 0.970 | 0.983 | 0.940 |
| Skewness | | | |
| No skewness | 0.980 | 0.988 | 0.960 |
| Moderate skewness | 0.985 | 0.991 | 0.963 |
| Ability differences x item discrimination | | | |
| Equal x Low | 0.056 | 0.049 | 0.050 |
| Unequal x Low | 0.051 | 0.046 | 1.053 |
| Equal x Moderate | 0.053 | 0.054 | 0.049 |
| Unequal x Moderate | 0.047 | 0.033 | 0.048 |
| Equal x High | 0.051 | 0.047 | 0.040 |
| Unequal x High | 0.061 | 0.069 | 0.052 |
| Average | 0.983 | 0.989* | 0.963* |
| Wald$\chi^2$ | 87.80 $\alpha = 0.05$ | | |
| GMH$\chi^2$ | 103.7 $\alpha = 0.05$ | | |
| SIBTEST | 170.9 $\alpha = 0.05$ | | |
| LDFA | 186.2 $\alpha = 0.05$ | | |

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

From Table 3, all three procedures had good power for detecting DIF in dichotomous test items; the average power was 0.983, p< .05 and 0.989, p< .05 and 0.963, p< .05 for the GMH, LDFA, and SIBTEST, respectively. Thus, the SIBTEST displayed a slightly poorer performance than the other two procedures for all items (SIBTEST = 0.051, p< .05), with average power of 42.6% compared to approximately 45% for both GMH and LDFA. The power of the SIBTEST to detect DIF decreased markedly as item discrimination increased. Power was 0.051, when item discrimination was low, 0.048, when item discrimination was moderate and 0.046, when item discrimination was high. There was an interaction between ability differences and item discrimination, so that power was slightly higher under the condition of an equal by low ratio for the three methods in combination with high item discrimination (GMH = 0.056, p< .05; LDFA =0.049, p < .05; SIBTEST = 0.050, p< .05) than it was when item discrimination was high and ability difference was equal by high ratio ( GMH = 0.051, p< .05; LDFA = 0.047, p < .05; SIBTEST = 0.040, p< .05). This scenario was reversed when item discrimination was moderate (GMH = 0.053, p< .05;

LDFA = 0.054, p< .05; SIBTEST = 0.049, p< .05), which is significant at p< .05.

To determine whether significant difference exist in the performance of GMH, SIBTEST, and LDFA methods for detecting DIF in dichotomous test items, the three procedures yielded a value of GMH$\chi^2$ = 103.7, p<.05, SIBTEST = 170.9, p<.05, LDFA = 186.2, p<.05, and Wald$\chi^2$ = 87.80, p<.05, which are significant values. These results suggested there was a significant difference in the performance of GMH, SIBTEST, and LDFA methods for detecting DIF in dichotomous test items. Thus, the null hypothesis is disconfirmed.

**Hypothesis Two**: There is no significant relationship between the proportions of test items that function differentially in the dichotomous test when the different methods are used. To test this hypothesis, the proportion of correctly identified DIF items was used as a power estimate for combining the $\Delta p_j$(*p-values*)across the levels of total score, following the procedure used by Dorans and Kulick (1986). The result is presented in Table 4 for dichotomous test.

**Table 4. 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 |

***s*Significant, p<0.05**

Table 4 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.

Further, Table 5 presents the results of the Chi-square ($\chi^2$) analysis. From Table 5, 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 5. Relationship Between the Proportions of Test Items That Function Differentially in the Dichotomous Test**

| Groups | Methods GMH | SIBTEST | LDFA | Total | $\chi^2$ | p |
|---|---|---|---|---|---|---|
| Reference | 12 (24%) | 5 (10%) | 10 (20%) | 27 | 6.269 (ns) | > 0.05 |
| Focal | 6 (12%) | 12 (24%) | 5 (10%) | 23 | | |
| Total | 18 | 17 | 15 | 50 | | |

**ns = not significant, p>0.05**

Further, the Chi-square ($\chi$2) analysis of the results yielded 6.269, which is not significant at p>0.05. Thus, the null hypothesis is confirmed; that is, there is no significant relationship between the proportions of test items that function differentially in the dichotomous test items when the different methods are used.

## 4. DISCUSSION

From the above, the results of the analysis of the research question indicated that for dichotomous test, 18 items exhibit negligible (A- level) DIF. These are items 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 15, 18, 21, 22, 24, and 25. Further, 15 items exhibit moderate (B-level) DIF, these are items 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 39, 40, 41, and 42. Also, 6 items exhibit large (C-level) DIF, these are items 43, 44, 45, 46, 48, and 49. In dichotomous test, GMH classified items 2, 5, 6, 11, and 21 as having negligible DIF (A-level), with items 2, 11, and 21 favouring reference group, and items 5, and 6 favouring focal group respectively. Also, GMH classified items 29, 30, 40, 41, and 42 as having moderate (B-level) DIF, with items 40, 41, and 42 favouring focal group, and both items 29 and 30 favouring reference group. Similarly, GMH classified items 44, 48, and 49 as large (C-level) DIF, all favouring the reference group only. On the other hand, LDFA classified items 1, 4, 9, 10, 12, 22, 24, and 25 as having negligible (A-level) DIF, with items 9, 12, and 24 favouring focal group, but items 1, 4, 10, 22, and 25 favouring reference group. In the same vein, LDFA classified items 31, 33, 35, 36, 37, and 39 as having moderate (B-level) DIF, with items 31, 36, and 37 favouring focal group, but the trio of items 33, 35 and 39 favouring reference group respectively. Noteworthy, LDFA did not classified any items as having large (C-level) DIF. Further, SIBTEST classified items 3, 7, 13, 15, and 18 as having negligible (A- level) DIF, with both items 15 and 18 favouring focal group, while items 3, 7, and 13 favouring reference group. Also, with SIBTEST, items 26, 27, 28, and 32 were classified as having moderate (B-level) DIF, while item32 only favouring focal group and items 26, 27, and 28 favouring reference group. More so, SIBTEST classified items 43, 45, and 46 as having large (C-level) DIF, with both items 43 and 45 favouring focal group, and item 46 favouring the reference group respectively.

The results of this study are in consonance with the earlier findings of Wang and Su (2004) 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 Wang and Yeh (2003) 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.

The result of the analysis of the first hypothesis revealed that there was a significant difference in the performance of the GMH, SIBTEST, and LDFA methods for detecting DIF in dichotomous test items. That all three procedures had excellent power for detecting DIF in dichotomous test items; the average power was 0.983, $p < .05$, and 0.989, $p < .05$, and 0.963, $p < .05$, for the GMH, LDFA, and SIBTEST, respectively. Thus, the SIBTEST displayed a slightly poorer performance than the other two procedures for all items (SIBTEST = 0.051, $p< .05$), with average power of 42.6% compared to approximately 45% for both GMH and LDFA. This result is consistent with the findings of Gierl et al., (2001) who reported that in comparison with MH, which they referred to as "the conservative procedure", LR could flag a large number of items as exhibiting DIF. Also, as Gierl et al., (2003) have found, LR has excellent Type I error rates which is a reassuring point for the researchers who choose LR as their DIF detection method.

On the whole, what can be inferred from this comparative discussion is that, generally, LDFA is more likely to flag an item with moderate or large DIF than the other two DIF detection methods which have been generally used at different times by scholars/researchers. In other words, by utilizing LDFA, the researchers can be sure that they obtain a list of DIF items which might not be flagged as displaying DIF by either the GMH or SIBTEST procedures. Metaphorically, LDFA feels free to accuse an item of displaying DIF. This, implicitly, can be considered as a reassuring point for the developers of the dichotomous test items (Ibrahim, 2013).

These findings compliment the work previously done by Teresi et al., (2000) as well as the work done by Zumbo & Hubley (2003) regarding the strength of MH when compared to the Exact Test and Logistic Regression when compared to MH. Jodoin and Gierl (2001) also emphasized that while LR has comparable power to MH and SIBTEST in detecting uniform DIF, it is superior in power for detecting non-uniform DIF. In addition, Gierl et al., (2001) found that effect size measures (for MH, SIBTEST and LR) were highly correlated across DIF procedures except the measure for non-uniform DIF, which could only be assessed by LR. Altogether, these findings can provide tacit confirmation as to the superiority of GMH and LDFA over SIBTEST. There is another point which may add to the value of the findings of the study. In an attempt to examine the reliability of different DIF detection methods, Swanson et al., (2002) made a comparison between standardization, MH and LR methods and found that more items could be identified as exhibiting DIF by LR than the MH and standardization methods. That is, items which were labelled as 'exhibiting DIF' by the MH and STD (i.e., standardization) methods could be identified as either uniform DIF or non-uniform DIF in the LR method. In other words, all the items labelled as "exhibiting DIF" by both the MH and standardization methods, were also detected to exhibit DIF by the LR method.

It can be inferred, therefore, that in detecting DIF items the LDFA method is more sensitive than the GMH and SIBTEST methods, and that in comparison with GMH and SIBTEST, the LDFA method tends to label the most items as exhibiting DIF as regards dichotomous test items.

Another finding of this study indicated that there was no significant relationship between the proportions of test items that function differentially in the dichotomous test when the different methods are used. This finding is similar to Gierl et al., (2003) and Su & Wang (2005) respectively 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. Altogether, these findings provide tacit confirmation as to the superiority of GMH over SIBTEST.

## 5. PEDAGOGICAL IMPLICATIONS OF THE RESULTS OF STUDY TO EXAMINATION BODIES AND UNIVERSITY TEACHERS

The findings of this study have provided supporting evidence that use of DIF statistical procedures are plausible alternatives to the use of expert judgement in determining the psychometric properties of both achievement and psychological tests. Not only this, but also this study has demonstrated the process of developing fairness in testing and provides statistical perspectives on the ways that scores from tests or items are interpreted in the process of evaluating test takers for a selection or classification decision. Fairness in testing is closely related to test validity, and the evaluation of fairness requires a broad range of evidence that includes empirical data such as the statistical detection of a number of items within a test (be it dichotomous or ordinal) for DIF investigation. Also, this study gave credence to the premise that when investigating DIF in achievement or academic and research measures (psychological and educational tests), analyzing test items is an exercise requiring not only psychometric and statistical skill but also good judgment. Such skill and judgment are needed even when the item analysis is

relatively straightforward, as, for instance, when examining the difficulty and the discriminating power of an item.

As educators, testing bodies are called to provide fair and equitable tests regardless of the population size of the students, testing bodies will have to address these concerns. This is significant because large testing bodies (WAEC, NECO, & JAMB) typically rely primarily on statistical and judgemental expertise. The magnitude of DIF and or the population size cannot easily be controlled in real life scenarios. Using a powerful DIF identification procedure is therefore important.

## 6. CONCLUSION AND RECOMMENDATIONS

Based on the findings obtained from the study, it can be concluded, therefore, that in detecting DIF in dichotomous, the LDFA method is more sensitive than the GMH and SIBTEST methods and that in comparison with GMH and SIBTEST, the LDFA method tends to label the most items as exhibiting DIF. Thus, a statistically significant relationship existed between individual items that function differentially due to the magnitude of DIF detected. The three methods complement each other in their ability to detect DIF in the dichotomous test format 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 judgement) analyses that can inform the test development process should be conducted after the administration of a test; (iii) test reviews at WAEC, NECO, and JAMB should be conducted using DIF statistical analyses to flag the DIF items and to sensitize developers and item writers to the sources of DIF, and to reduce the number of DIF items on a test; and (iv) also, DIF analysis should be employed in such tests used for promotion examinations for civil servants in all government ministries in the country (Nigeria) in order to ensure good psychometric procedure(s) and test fairness and equity.

## REFERENCES


Gierl, M. J. Jodoin, M. G. & Ackerman, T. A. (2000).*Performance of mantel-haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans: LA.

Gierl, M. J. Bisanz, J. Bisanz, G. L. Boughton, K. A. &Khaliq, S. N. (2001). 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.

Gierl, M. Khaliq, S. N. and Boughton, K. (2003). *Gender differential item functioning in Mathematics and Science: Prevalence and policy implications.* Paper presented at the symposium entitled "improving Large-scale Assessment in Education" at the annual meeting of the Canadian Society for the Study of Education, Quebec: Canada.

Guilera, G. Gomez, J. and Hildago, M.D. (2009). Scientific production on the Mantel-Haenszel procedure as a way of detecting DIF. *Psicothema*, 21(3), 492-498.

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

Ibrahim, A. (2013). *Comparative study of generalizedMantel-Haenszel, simultaneous item bias test and logistic discriminant function analysis in detecting differential item functioning in dichotomous and ordinal test items*. Unpublished Ph.D. Thesis, Faculty of Education, ObafemiAwolowo University, Ile-Ife, Nigeria.

Ivowi, U. M. O. (2009). A critique on protocol in the assessment of science questions at the senior school certificate examination. In Nigerian Academy of Education (Eds.), *Educational Measurement and Evaluation*, pp. 261-280, Ibadan, Nigeria: Ark Publishers.

Jodoin, M. G. &Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *AppliedMeasurement in Education, 14*, 329 – 349.

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.

Magis, D. Beland, S.Tuerlinckx, F. &DeBoeck, P. (2001). A general framework and an R package for the detection of dichotomous differential itemfunctioning. *BehaviourResearch Methods*, 42, (3), 847-862.

Mantel, N. and 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. and Spray, J. A. (1993). 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 itemfunctioning*. Los Angeles: Sage Publications, Inc.

Pezzulo, J. (2010). *LazStats.* Available at http://statpages.org .Roussos, L. A. & Stout, W. (1996). Simulation studies of the effects of small sample and studied tem parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, 33, 215-230.

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

Sidhu, K.S. (2005). *New approaches toeasurement and evaluation*. New Delhi,India:Sterling Publishers Private Limited.

Stephens-Bonty, T. A. (2008). *Using three different categorical data analysis techniques to detect differential item functioning*. Unpublished Ph.D. dissertation, Georgia State University,Georgia.

Su, Y. H. & Wang, W. C. (2005). Efficiency of theMantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement inEducation*, *18*(4), 313-350.

Swanson, D. B. Clauser, B. E. Case, S. M. Nungester, R. J. &Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics, 27*, 53–76.

Teresi, J. A. Kleinman, M. and Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine*, 19, 165 – 183.

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

Wang, W.C. and Yeh, L.Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*,27, 479-498.

Wang, W.C. & Su, Y.H. (2004). Factors influencing the mantel and generalized mantel-haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, 28,450-481.

Zumbo, B. D. &Hubley, A. M. (2003). Item bias. In R. Fernandez-Ballesteros(Ed.), *Encyclopedia of psychological assessment*. Thousand Oaks, CA: Sage

## APPENDIX: UNDERGRADUATE STUDENTS' ACHIEVEMENT TEST (USAT)

### INTRODUCTION

This instrument is purely for research purpose. It has nothing to do with your academic standing in the university as well as your integrity. Please, give your response as candidly as possible and do not leave out any item(s) unattended.
Thank you very much.

SECTION A: PERSONAL INFORMATION

Name (optional)……
Sex: Male…… Female …Course of Study: …
Level: …… Department: …… ……

SECTION B: INSTRUCTIONS:

Answer All Questions. Each item below has four alternative answers (A-D). Circle the correct or most appropriate option. Attempt all items on the question paper given to you by encircling appropriate letter-alternative per item.

1.Theoretically, reliability may be defined as X = T + E. In this equation, X is called ……..
(a) True score
(b) Error score
(c) Observed score
(d) Appropriate score

2. Test development involves all but one of the following procedures:
(a) Item analysis
(b) Test validity
(c) Test scoring
(d) Table of specifications

3. Obtaining a dependable ranking of students is of major concern when using:
(a) Vocational test
(b) Diagnostic test
(c) Mastery tests
(d) General achievement tests

4. The parts that provide possible solution to the problem in multiple-choice test are:
(a) Stems
(b) Distracters
(c) Options
(d) Alternatives

5. Which one of the following is not a category in thetaxonomy for cognitive domain?
(a) Comprehensiveness
(b) Application
(c) Analysis
(d) Interpretation

6. Among these factors, which typically testwiseness would be expected to influence performance on a standardized cognitive test?
(a) Practice on a parallel form of the test
(b) The gambling response style
(c) The positional preference response set
(d) Testwiseness

7. A norm-referenced test refers to:
(a) Measure of performance of an individual relative to a defined test domain.
(b) Measure of performance of a person in terms of his/her relative standing in a known group.
(c) A test that is ascertained to be valid and reliable
(d) A test manipulated to conform to the normal curve having a known mean and standard deviation.

8. The thrust of classroom testing is:
(a) Evaluation of pupils
(b) Measurement of pupil behaviour
(c) Assessment of students' potential
(d) Determination of curriculum effectiveness

9. Universe is to population as sample is to:
(a ) Randomization
(b) Inference
(c) Representatives
(d) Research work

10. Rank is to ordinal as category is to:
(a) Interval
(b) Ratio
(c) Variance
(d) Percentile

11. The single most important attribute of test is ……
(a) Affordability
(b) Validity
(c) Reliability
(d) Usability

12. What is really "objective" in an objective test is:
(a) The answer
(b) The score
(c) The format
(d) The marking

13. An example of a supply-test item is the ……
(a) Multiple-choice item
(b) True-false item
(c) Short-answer item
(d) Completion item

14. If the length of a test is increased, this will increase the test's ……..
(a) Difficulty
(b) Validity
(c) Discrimination
(d) Reliability

15. A test that consist of items of equal or approximately equal difficulty could be labeled a …….
(a) Multiple-choice test
(b) Power test
(c) Short-answer item
(d) Speed test

16. …… are tests given with ample time for all examinees to demonstrate how well they can perform.
(a) Speed tests
(b) Power tests
(c) Personality tests
(d) Intelligence tests

17. A test answer that gives different results when graded by different examiners is most likely to result from a/an …….. test.
(a) Verbal
(b) Objective
(c) Short-answer
(d) Essay

18. Despite disagreement among educators about what intelligence is, intelligence tests are still used as measures of ……
(a) Personality traits
(b) Vocational aptitudes
(c) Intellectual capacities
(d) Artistic aptitudes

19. A pupil's performance in an achievement test could be affected by his……
(a) Innate ability
(b) Intelligent quotient
(c) Acquired ability
(d) Testwiseness

20. Extraneous factors that influence performance of students on cognitive tests are among the alternatives EXCEPT
(a) Cheating
(b) Coaching
(c) Answer changing
(d) Interest

21. Which of the following forms of reliability require more than one administration?
(a) Split-half reliability
(b) Parallel-form reliability
(c) Test-retest method of reliability
(d) Kuder-Richardson reliability

22. Methods of estimating a test's reliability that require only one administration are more commonly employed especially for ……
(a) Classroom tests
(b) Vocational tests
(c) Psychological tests
(d) Standardized tests

23. All but one of the following is important features of a test-blue print. Which is the exception?
(a) The topics that were treated
(b) The number of questions in each topic
(c) The test format to be used
(d) The marks assigned to each question

24. Which of the following test formats is likely to be most vulnerable to guessing?
(a) Matching
(b) Multiple-choice
(c) Short answer
(d) True-false

25. Some objective items are of the supply type. Which pair below suits that description?
(a) Completion/multiple-choice tests
(b) Close test/ matching tests
(c) Matching/completion tests
(d) Short answer/true-false tests

26. One advantage of multiple-choice items over essay questions is that they..……..
(a) Provide for the measurement of more complex learning outcomes
(b) Place greater emphasis on the recall of factual information
(c) Require less time for test preparation andscoring
(d) Provide for a more extensive sampling of course content

27. Instructional objectives are most useful for test construction purposes when they are stated in terms of ……
(a) Course content
(b) Learning activities
(c) Behavioural terms
(d) Entry behavior

28. Teacher-made tests suffer some deficiencies.Which of the following is not weakness of teacher-made tests?
(a) Excessive wording
(b) Extraneous clues to the answer
(c) Ambiguous questions
(d) Inappropriate test formats

29. Which of the following does not describe a standardized test?
(a) Equivalent forms of the test are provided
(b) Norms are used in interpreting scores
(c) The items are of a high technical quality
(d) Its psychometric properties are well stated

30. The essential difference between tests and measurement is:
(a) Tests involve questions while measurement does not
(b) Measurement may not involve numericalscores
(c) Tests yield numerical scores
(d) Tests give more precise scores

31. The standard error of measurement (Sem) may be expressed as SEm = $\sigma$ /1-r. If $\sigma$ = 4.5; r = 0.61. Then SEm =
(a) 1.5
(b) 2.0
(c) 2.8
(d) 3.6

32. The standard error of measurement is an especially useful way of describing test
(a) Validity
(b) Reliability
(c) Objectivity
(d) Usability

33. In treating test scores, the standard error of measurement indicates:
(a) The amount of error allow for when interpreting individual test scores
(b) How much confidence to place in the results
(c) Reliability for each test
(d) The discriminating power of the items

34. The reason a standard score is preferred to raw score in test reporting is:
(a) Standard scores are difficult to comprehend
(b) Standard scores allow for comparability
(c) Standard score are easier to understand
(d) Standard scores tend to reduce errors due to raw scores

35. The National Policy on Education (NPE) recommends the use of standard scores in test reporting. Which standard score in particular does the policy recommended?
(a) Z-score
(b) T-score
(c) S-score
(d) Y-score

36. Practical considerations aside, which method of estimating reliability is generally preferable?
(a) Parallel- form
(b) Test- retest
(c) Split-half
(d) Both (a) and (c)

37. Which formula is used to estimate the reliability of a test if its length is increased ordecreased?
(a) Flanagan formula
(b) Spearman-Brown formula
(c) r KR20
(d) r KR21

38. Which of these tends to yield the lowest reliability co-efficient for power tests?
(a) Corrected split-half
(b) Test-retest
(c) r KR20
(d) r KR21

39. Which of these methods of estimating reliability requires the copulation of a correlation co-efficient?
(a) Kuder-Richardson formula
(b) Flanagan formula
(c) Co-efficient alpha
(d) All of the above

40. Which of these factors that influence test performance is not classified as a response set?
(a) Acquiescence
(b) Practice
(c) Guessing
(d) Both (c) and (a)

41. Which of these is not an example of testwisenessbehaviour?
(a) Using the process of elimination in selecting correct answer
(b) Working carefully, double-checking each item before going on to the next item
(c) Guessing at ball times even when the standard correction for chance formula is used
(d) Using specific determiners as clues to correct answers

42. Test-retest practice effects on cognitive test are:
(a) Greater for inexperienced examinees
(b) Greater on power tests than on speeded tests
(c) Less when the time interval between the tests is short
(d) Usually greater than .25

43. On a five-option multiple-choice test, what is the appropriate form for the correction-for-chance formula?
(a) S = R – W / 2
(b) S = R – W / 3
(c) S = R – W / 4
(d) S = R – W / 5

44. A test consists of 10 five-option multiple-choice items. If you answered all 10 correctly, what is your corrected-for-chance score?
(a) 5
(b) 10
(c) 15
(d) 20

45. On multiple-choice tests, the standard deviation for corrected-for-chance scores, S, compared to incorrect scores, R, is……
(a) Greater
(b) The same
(c) Less
(d) Both (a) and (c)

46. If no examinee omitted any items and if your score was at the $90^{th}$ percentile in the uncorrected ( R ) distribution, in the corrected-for-chance ( S ) distribution:
(a) Your percentile rank score would not change
(b) Your percentile rank score would decrease
(c) Your percentile rank score would increase
(d) Your percentile rank score would neither increase nor decrease

47. Which of these is not true regarding effects of response styles on test performance?
(a) They tend to decrease the validity of tests
(b) They tend to decrease test reliability
(c) They have greater influence on difficult test
(d) They have greater influence on speeded test

48. A test that possesses both intra-rater and inter-rater's reliability is:
(a) Reliable
(b) Objective
(c) Formal
(d) Subjective

49. Of the three basic attributes of a good test, the most important of them is:
(a) Usability
(b) Reliability
(c) Objectivity
(d) Validity

50. Basic attributes of a good test are:
(a) Validity, reliability and usability
(b) Validity, reliability and objectivity
(c) Reliability, objectivity and usability
(d) Validity, reliability and subjectivity.