COMPARISON OF INDEX OF DIFFERENTIAL ITEM FUNCTIONING UNDER THE METHODS OF ITEM RESPONSE THEORY AND CLASSICAL TEST THEORY IN MATHEMATICS

ALORDIAH, CarolineOchuko
PG/10/11/191906

DEPARTMENT OF GUIDANCE AND COUNSELLING
DELTA STATE UNIVERSITY,
ABRAKA

NOVEMBER, 2015

COMPARISON OF INDEX OF DIFFERENTIAL ITEM FUNCTIONING UNDER THE METHODS OF ITEM RESPONSE THEORY AND CLASSICAL TEST THEORY IN MATHEMATICS

ALORDIAH,CarolineOchuko
PG/10/11/191906
B.Sc. (Ed.) 1993 (UNN), M.Ed. 2010 (DELSU).

A Thesis inthe Department of Guidance and CounsellingSubmitted to the Postgraduate School in Partial Fulfillment of the Requirements for the Award of Doctor of Philosophy (PhD) Degree in Measurement and Evaluation of the Delta State University, Abraka.

DEPARTMENT OF GUIDANCE AND COUNSELLING
DELTA STATE UNIVERSITY,
ABRAKA.

NOVEMBER, 2015

CERTIFICATION

We the undersigned certify that this research was carried out by ALORDIAH, CarolineOchukoof the Department of Guidance and Counselling, Delta State University, Abraka, and is adequate in scope and content for the Ph.D. in Measurement and Evaluation.

Prof. C.E. Mordi (Supervisor)	**Date**
Dr. J.N. Odili (Supervisor)	**Date**
Dr. P. U. Osadebe (Ag. Head of Department)	**Date**
Prof. E. P. Oghuvbu (Dean, Faculty of Education)	**Date**

DECLARATION

I declare that this research was carried out by mein the Department of Guidance and Counselling, Faculty of Education.

Alordiah, Caroline.Ochuko. (Student)	**Date**

DEDICATION

This thesis is dedicated to my husband, Pastor MichaelOmoviye Alordiah.

ACKNOWLEDGEMENTS

Firstly, I wish to express my profound gratitude to all those who have contributed immensely to the success of this study. I thank the Almighty God for granting me the grace that sustained me throughout the course of this study.

My appreciation goes to the project supervisors Prof C.EMordi and Dr J. N. Odili for their wonderful guidance and support. I am very grateful to Dr J. N. Odili for his indispensable advice, correction and useful criticisms with which the writing of this thesis is a success. My special thanks go toDrAgustusTrinster,Dr P. U.Osadebe, Prof. R. I. Okorodudu, Dr (Mrs) G. O.Akpochafo and Dr S. D. Clifford who provided me with very useful items of information that have contributed to the success of this work.

I appreciate the scholarship foundation of the instituto de Evaluacion e ingenieriaAvanzada, San Luis Potosi, Mexico, under project number IEIA-140204. The institude did some of the data analyses in this work.

I am indebted to my husband, Pastor Michael Omoviye Alordiah and my children- Vieloze, Ona, Ogaga and Zino for their love, understanding and encouragement throughout the programme. I am grateful to my course mates,DrEbisineSele Sylvester, AliyuTaiwo andOnyehiAbaye Kingsley for the team spirit that existed among us. Finally, my appreciation goes to the principals, teachers and students of the selected secondary schools who have helped to make this project a success.

TABLE OF CONTENTS

Pages

CHAPTER ONE: INTRODUCTION

CHAPTER TWO: REVIEW OF RELATED LITERATURE

CHAPTER THREE: RESEARCH METHOD AND PROCEDURE

CHAPTER FOUR: PRESENTATION OF RESULTS AND DISCUSSION

CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATIONS

APPENDIX

LIST OF TABLE

LIST OF FIGURES

ABSTRACT

Differential item functioning (DIF) in test items has been an issue in testing. It can occur in national examinations conducted in a heterogeneous country like Nigeria. This has generated the proliferation of several methods that can be used to detect DIF items in a test. Whether these DIF methods can detect the same test items as DIF items is of much concern to measurement and evaluation experts. More so that some of these methods of detecting DIF are based on classical test theory (CTT) while others are based on item response theory (IRT). The main purpose of this study is to compare the index of DIF for a given sample under the methods of CTT and IRT for candidates with the same mathematics ability from different socio-economic statuses (SES), location and gender. Four DIF detection methods were used in this study; two of these methods were based on CTT- namely transformed item difficulty (TID) and Mantel-Haenszel (M-H); while the other two were based on IRT- namely item response theory three parameter model (IRT-3P) and Rasch model. The four DIF detection methods were used to determine the index of DIF for gender, location and SES for 2012 WASSCE mathematics objective test. The DIF indexes for these four methods were later compared to find to what extent they were able to detect the same test items as DIF items. An ex-post facto design was adopted. The population of this study consisted of all senior secondary class III students' in public schools in Delta and Edo states. The proportionate stratified random sampling approach was used to sample out one thousand nine hundred (1900) students from the population. Twelve research questions and nine hypotheses testable at 0.05 level of significance were used and data were collected using two instruments, these are the 2012 WASSCE mathematics objective test and the socio-economic status questionnaire whose validity was ensured. The reliability of the 2012 WASSCE mathematics objective test and the socio-economic status questionnaire using test-retest method yielded 0.892 and 0.702 respectively. Data generated were analyzed using SPSS 17, BILOG-MG and WINSTEPS 3.2 packages. Descriptive statistics was used to answer the research questions while chi-square independence test and the contingency coefficient were used to test the hypotheses. The findings of the study revealed that CTT methods of detecting DIF did not agree with IRT methods of detecting DIF in the items flagged as DIF. The methods of detecting DIF within CTT did not agree in the items flagged as DIF. However, there was agreement in the methods of detecting DIF within the IRT in the items flagged as DIF. It was recommended that measurement and evaluation experts should freely use the methods of detecting DIF that are based on IRT and that seminar and workshop should be carried out to aid the proper understanding of DIF detection methods that are based on IRT.

CHAPTER ONE

INTRODUCTION

Background to the Study

The national examinations conducted by West African Examination Council (WAEC), National Business and Technical Education Board (NABTEB) and Joint Admissions and Matriculation Board (JAMB) cater for candidates from various backgrounds all over the country. In some cases, an item in these examinations could be more difficult for particular group of students while very easy for other group of students. When there is something in an item that makes students who are on the same ability level but from different subgroups to perform differently, we say such an item shows differential item functioning. Differential Item Functioning (DIF) in test items has been an issue in testing. It can occur in national examinations conducted in a heterogeneous country like Nigeria. This has generated the proliferation of several methods that can be used to detect DIF items in a test. Whether these DIF methods can detect the same test item as DIF item is of much concern to measurement and evaluation expert. More so that some of these methods of detecting DIF are based on Classical test theory while others are based on Item response theory. Differential item functioning (DIF) implies that even after controlling for ability, an item appears to be more difficult for examinees from one group, as compared to examinees in other groups. Augemberg and Morgan (2008) put it that, Differential item functioning (DIF) is observed when comparable (matched on ability) examinees from different groups have a different probability of answering a given item correctly. Simply put DIF occurs when test takers from different groups (race, ethnicity, language, culture, location, religion, gender, or socio-economic status) that have been matched on ability levels are performing differently in test items. The presence of large number of items with DIF is a serious threat to the validity of a test and any inference made from such scores may not be valid. According to the Federal Government of Nigeria (FGN) (2004) in the national policy on education. Every Nigerian child shall have a right to equal educational opportunities irrespective of any real or imagined disabilities each according to his or her ability and there shall be the provision of equal access to educational opportunities for all citizens of the country at the primary, secondary, and tertiary levels both inside and outside the formal school system. The implication of this to all educationists and most especially to the psychometricians is that items in test should be fair to all subgroups in the population. There should be nothing in the test items that would make the items to favour one group above the other group that are of equal ability level.

There are two main types of DIF, namely uniform DIF and non-uniform DIF. Uniform DIF is said to occur when differences in correct response probability are found across all ability levels for a particular item. On the other hand, non-uniform DIF occurs when there is an interaction between the ability and group membership such that an item may seem difficult for those at the higher level in one group and after a particular point, it becomes more difficult for those at the lower level in the other group.

As a result of the potential danger of DIF in a test, it has generated researchers' interest with so many methods being evident in literature for its detection. These methods could be characterized according to the two measurement theories namely, Item response theory (IRT) and Classical test theory (CTT). Under the CTT an item is said to show DIF if there is a significant difference between the P-value of two groups matched for ability. DIF is also based on group differences in total score. Some DIF detection methods based on CTT are the transformed item difficulty (TID), point biserial correlation, Mantel-Haenszel (M-H), standardization, Scheuneman chi-square etc. CTT based DIF detection methods are easy to use and the standard statistical packages like SPSS and SAS can be used to analyse it with ease. However, the CTT based methods of assessing DIF are limited because they based the assessment of DIF on the presence of group mean differences in total test scores across subgroups. The method based on CTT assumes the same average ability, which is probably false. This is so because classical test theory parameters are test-dependent and examinee-dependent. In other words, item difficulty changes when a shift is made from a sample whose mean ability is high to one whose mean ability is low. Consequently, the same individual tested with two different groups of testees may obtain two different errors of measurement and estimates of true score (Weiss & Davison, 1981).

Warm (1978) lamented that the comparison of p-values across groups assumes that bivariate distribution of the p-value is linearly related but under CTT p-values are not linearly related. Due to the cheapness, availability, and simple to use nature of these methods, people have used it over the years and if these criticisms are anything to go by, they must have been misinforming people

The observed criticisms above gave rise to Item Response Theory (IRT). Under the IRT, the definition of DIF as stated by Odili (2010) is that DIF is the tendency of test takers of the same standing in the latent trait to perform differently in a test item. IRT is a theory that solves the problem of test-dependent and examinee-dependent of CTT item parameters. Some of the IRT based methods of detecting DIF are item characteristic curve (ICC), Item Response Theory likelihood ratio (IRT-LR), parameter comparison using t-test on b-value, Rasch model, IRT-two parameter model, IRT-three parameter model and so on. In testing an individual observe score (X_i) in IRT it is stated as $X_i = \theta_i + \lambda_i + \varepsilon_i$ where θ_i is the true ability component for the examinee, λ_i is the extraneous (systematic) error variable component of the score and ε_i is the random error component. A random error is an error that is not operating in one way and the process of sampling can eliminate it. Systematic errors are those aspects of error which when present in a test give advantage to a group of test takers and disadvantages to another group of test takers of the same ability. Some of the sources of systematic error (λ_i) could be the language of the test item, test wiseness (knowledge of how to respond advantageously to psychological test), culture, race, gender and so on. The recognition of systematic error in IRT is a major point of deviation from Classical Test Theory. It focuses on performance on individual items, rather than only on the intact test. It is equally true that IRT methods of detecting DIF are complex and highly mathematical, the software for it is not readily available in Nigeria, in cases where the software is available, and it is very difficult to operate. IRT needs large samples and relatively large number of items.

Some of the sources of systematic error, which could lead to detection of DIF in a test, could be gender, location- urban/rural and socio-economic status of parents (Odili, 2003). He also, reported that if biology test contains items that are differentially functioning, it might systematically reduce the opportunity of some testees like those from low socio-economic status, rural location and so on, from gaining admissions into such careers like medicine, pharmacy. This is equally true for mathematics more so when mathematics is needed at credit level to gain admission into any course in the university.

Gender is the term that is used to describe any individual due to the behaviour and character that is exhibited for the fact that the individual was born either male or female. In other words, gender is the socio-cultural interpretation of male and female based on their expected role, contributions and assigned duties (Ija, 2009). Gender related DIF is a regular issue in mathematics achievement test (Abedalaziz, 2011). Uwadiae (2008) published that out of about 13.76% of the candidates who had credits and above in mathematics and English language plus three other subjects in the 2008 WAEC/SSCE, 7.33% were males while 6.43% were females. According to Ayodele (2011) this signifies that the males performed slightly better than the females. Viewed from the above perspectives, gender differences in mathematics is inconclusive and widely open to further investigation.

Socio-economic status (SES) is the way people are divided into groups in a society such that they have certain economic or/and social characteristics in common. Socio-economic status of a family is usually linked with the family's income, parent's educational level, parent's occupation, and social status (Okafor, 2007). It is generally well documented that higher family socio-economic status is related to higher educational expectations for youths (Wantzel, 1998). However, focus should be on the integration of the SES classes into our teaching and learning process as well as putting it into consideration during evaluation.

In Nigeria, the lingual Franca is English language, which in most cases is not widely spoken in rural schools. What obtains in most cases is the use of the native language of that setting as means of communication. This can greatly affect students' performance in mathematics since it is with English language that mathematics is taught and assessed in schools. According to Odili (2003) because of an improved language-learning environment, the students in urban area are likely to perform better than those in rural area.

Over the years, DIF has attracted the attention of many researches outside the country. However, there are very few works on it at the local level. Abedalaziz (2010) investigated a gender-related DIF of mathematics test items and found that there are gender differences in performance on test items in mathematics. Abedalaziz (2012) investigated comparison of CTT and IRT methods. He found out that the highest agreement was between chi-square (Scheauneman) and b-parameter, whereas the lowest agreement was between Area index and TID. He also, found that gender difference in mathematics might well be linked to content. Ironson & Subkoviak (1979) carried out a study on a comparison of several methods of assessing item bias; they found that the highest agreement was between chi-square, TID, and the discrimination differences approach (point-biserial) did not relate significantly with any other methods. Baghi and Ferrara (1989) in their study on a comparison of IRT, Delta-plot, and Mantle-Haenszel techniques for detecting DIF across subpopulations in the Maryland test of citizenship skills, found that the proportions of agreement for the Delta-plot and Rasch techniques are stable across sample sizes in white/black and male/female samples. However, an agreement proportion from the M-H technique in black/white samples was not stable but was stable for male/female sample across the different sample sizes. This work focused only on the effects of sample size on the rate at which the three methods can detect DIF. Odili (2003) undertook a study on the effect of language manipulation on DIF of biology multiple-choice test. The result revealed that WAEC/SSCE biology paper 2 for 1999, 2000, and 2001 contains items with significant location, gender, and socio-economic status DIF, with location having more DIF items. This is an indication that DIF occurs in tests used by public examination bodies in Nigeria.

Apart from the work done by Odili (2003), all other works were done outside Nigeria. The works focused on gender except for the work of Baghi and Ferrara (1989) that in addition to gender (male/female) used colour (black/white) which is not relevant in the Nigeria context. Abedalaziz (2011) compared two IRT methods (b-parameter and Area index) with two CTT methods (Scheneman chi-square and TID). Ironson and Subkoviak (1979) compared three CTT methods (TID, Point biserial and Chi-square) with one IRT (ICC). Also, Abedalaziz (2010) compared two CTT (M-H and TID) with one IRT (b-parameter). This shows that the comparison of methods of detecting DIF is not new in research. However, it is worthy to note that much comparison of these methods of detecting DIF that are based on CTT and IRT are not common in Nigeria. Literature revealed that there are several methods that can be used to detect DIF items in a test. Some of these methods are based on CTT while others are based on IRT. Literature also revealed that these theories are different in their approach. The controversy of the supremacy of Item Response Theory (IRT) over Classical Test Theory (CTT) because of sample invariance of test parameter inherent in the former continues to reign in psychometric. This is in spite of the huge cost associated with the application of IRT because of requirement of large sample size and high cost of computer time. Few works has been done to ascertain whether the differences that exist in these theories have effects on the results obtained from the DIF detection methods that are based on these theories. Therefore, it is against such background that the researcher conceived the idea of comparing the index of DIF under the methods of IRT and CTT for gender, socio-economic status and location for 2012 WASSCE mathematics multiple-choice test. The essence of picking the 2012 WASSCE mathematics multiple-choice test is because it covers the national mathematics curriculum; its items are drawn from an item bank from which previous (and likely future) years test items are drawn; as well as to find out if the problem of DIF still exists in such test.

Statement of Problem

Education is for all and every person must have equal access to education at all stages of the educational system. The implication of this to measurement and evaluation experts is that measurement of students learning should be free of bias. There should be nothing in the test items that would make an item to favour one group above the other group that are of equal ability. When individuals of the same ability but from different groups are performing differently in a test item, such item is said to show DIF. Such groups could include race, ethnicity, language, culture, gender, religion, location, socio-economic status and so on. Gender is the socio-cultural interpretation of male and female based on their expected role, contributions and assigned duties. From literature, gender differences in mathematics are inconclusive and widely open to further investigation. Socio-economic status (SES) is the way people are divided into groups in a society such that they have certain economic or/and social characteristics in common. It is generally well documented that higher family SES is related to higher educational expectations for youths. Due to the improved language-learning environment in the urban area, students from such environment are likely to perform better than those from rural environment. Due to the potential danger of DIF in a test, it has generated researchers' interest with so many methods being evident in literature for its detection. These methods can be subdivided based on the two measurement theories namely, item response theory and classical test theories. The parameters of the CTT methods are test-dependent and examinees-dependent. This is a major limitation of CTT. The IRT has been able to overcome this limitation. However, the methods based on CTT are easier to understand and used. It does not require large sample, which is unlike the IRT based methods that required large sample, are difficult to understand, and used which makes it more expensive to use. In addition, the software is not readily available. This implies that if the methods that are based on these theories are able to detect the same items as DIF items, then it will be unwise to use those methods that are based on IRT since they are more expensive to utilize. Seeing that of these advantages and limitation associated with these theories exist, could it be that the methods of detecting DIF based on CTT will detect DIF item differently from those methods of detecting DIF that is based on IRT.

From the above discussions, the problem of the study is that: would the methods of detecting DIF under IRT yield DIF items that significantly agree with DIF items obtained using methods that are based on CTT for candidates with the same mathematics ability from different socio-economic status, location, and gender?

Research Questions

1. What is the index of DIF for gender under the methods of Item Response Theory (Rasch model and Item Response Theory-3 Parameter model (IRT-3P)) and Classical Test Theory(Transformed item difficulty(TID) and Mantel-Haenszel (M-H)) for each item in 2012 WASSCE mathematics multiple-choice test?

2. What is the index of DIF for socio-economic status (SES) under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for each item in 2012 WASSCE mathematics multiple-choice test?

3. What is the index of DIF for location under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for each item in 2012 WASSCE mathematics multiple-choice test?

4. What is the agreement between the index of DIF for gender under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test?

5. What is the agreement between the index of DIF for gender within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test?

6. What is the agreement between the index of DIF for gender within the methods of IRT (IRT-3P and (Rasch model) for items in 2012 WASSCE mathematics multiple-choice test?

7. What is the agreement between the index of DIF for SES under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test?

8. What is the agreement between the index of DIF for SES within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test?

9. What is the agreement between the index of DIF for SES within the methods of IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test?

10. What is the agreement between the index of DIF for location under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test?

11. What is the agreement between the index of DIF for location within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test?

12. What is the agreement between the index of DIF for location within the methods of IRT (IRT-3P and Rasch model)?

Hypotheses

1. There is no significant agreement between the index of DIF for gender under the methods of Item Response Theory (Rasch model and IRT-3P) and Classical Test Theory (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test.
2. There is no significant agreement between the index of DIF for gender within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test.
3. There is no significant agreement between the index of DIF for gender within the methods of IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test.
4. There is no significant agreement between the index of DIF for SES under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test.
5. There is no significant agreement between the index of DIF for SES within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test.
6. There is no significant agreement between the index of DIF for SES within the methods of IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test.
7. There is no significant agreement between the index of DIF for location under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test.
8. There is no significant agreement between the index of DIF for location within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test.
9. There is no significant agreement between the index of DIF for location within the methods of IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test.

Purpose of the Study

The controversy of the supremacy of Item Response Theory (IRT) over Classical Test Theory (CTT) because of sample invariance of test parameter inherent in the former continues to reign in psychometry. This is in spite of the huge cost associated with the application of IRT because of requirement of large sample size and high cost of computer time. The general objective of this study is to compare the index of Differential Item Functioning (DIF) for a given sample under the methods of CTT and IRT for candidates with the same mathematics ability from different socio-economic status (SES), location, and gender

Specifically the objectives of the study are as follows:

1. To determine the index of DIF for gender under the methods of Item Response Theory (Rasch model and IRT-3P) and CTT (TID and M-H) for each item in 2012 WASSCE mathematics multiple-choice test.

2. To determine the index of DIF for SES under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for each item in 2012 WASSCE mathematics multiple-choice test.

3. To determine the index of DIF for location under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for each item in 2012 WASSCE mathematics multiple-choice test.

4. To find the agreement between the index of DIF for gender under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test.

5. To find the agreement between the index of DIF for gender within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test.

6. To find the agreement between the index of DIF for gender within the methods of IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test.

7. To find the agreement between the index of DIF for SES under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test.

8. To find the agreement between the index of DIF for SES within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test.

9. To find the agreement between the index of DIF for SES within the methods of IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test.

10. To find the agreement between the index of DIF for location under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test.

11. To determine the agreement between the index of DIF for location within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test.

12. To determine the agreement between the index of DIF for location within the methods of IRT (IRT-3P and Rasch model).

Significance of the Study

The study has much significance. The result of this study will be significant to measurement experts, practicing teachers, test developers, lecturers, public examination bodies, and government. Generally, the result of this study will help the measurement expert to know whether the same test item can be detected as DIF item by methods based on CTT and methods based on IRT.

The findings of this study will indicate whether the methods of detecting DIF under CTT will agree in the items flagged as DIF. It will also indicate whether the methods of detecting DIF under IRT will agree in the items flagged as DIF. This will guide practicing teachers and lecturers to take decisions in the appropriate method for detecting DIF. It will guide test developers to make sure that only non-DIF items are used for item banking.

Findings from this study will be useful to public examination bodies like WAEC, National Examination Council (NECO), Joint Admission and Matriculation Board (JAMB) in prescribing which DIF detection methods will be able to detect DIF in a test very well. The result of this study will encourage the government to hold more in service trainings, capacity buildings, and workshops for teachers, lecturers, and measurement and evaluation experts in particular on the various methods of detecting DIF.

Scope and Delimitation of the Study

This study focus on determining the agreement of the index of DIF under the CTT and IRT methods of detecting DIF. This is the dependant variable of the study. The following procedure of detecting DIF- TID and M-H which are based on CTT and Rasch model and IRT-3P which are based on IRT are the independent variables, will be used to find out how they can detect DIF in 2012 WASSCE mathematics object test. The secondary independent variables include gender (male and female), SES (high and low) and location (urban and rural). They were used as the focal and reference groups for each case (for instance gender: the focal group will be female and the reference group will be male). These variables were selected because they could show DIF between groups of the same ability. The various procedures for detecting DIF will be used to find out whether the 2012 WASSCE mathematics multiple-choice test show gender, socio-economic status and location related DIF as well as find out which procedure is able to detect DIF better with reference to the measurement theories (CTT and IRT).

The study was restricted to two states in South-South geopolitical zone, namely, Delta state and Edo state. Senior secondary three (SS3) students in public secondary schools were used. Such students were drawn from high and low socio-economic status, urban and rural area, male and female students.

Operational Definition of terms

Index of Differential Item Functioning- This is the value that shows whether DIF is present or not. The value 1 means there is no DIF while the value 2 means there is DIF.

Focal Group – This is the group of interest. It is the group you are studying to see whether it will differ from the reference group. They are female, low SES and rural.

Reference Group – This is the group that is to be used for the comparison. It is the group you want to compare the findings of the reference group with to see whether they are the same or different. They are male, high SES, and urban.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

The review of related literature is organized under the following subheadings:

1. Theoretical framework of the study
2. Concept of Differential Item Functioning
3. Method of detecting Differential Item Functioning
4. Gender issues
5. Socio-economic status
6. Location-urban/rural
7. Empirical studies
8. Appraisal of reviewed literature

Theoretical Framework of the Study

This work is based on the methods of detecting DIF that are based on two theories of measurement, namely; Classical Test Theory (CTT) and Item Response Theory (IRT). The classical test theory postulate that it is possible to construct two parallel test in which all the parameters are the same and in such test the observed score (X) is equal to a true component (T) and an error component E such that X=T+E. The assumptions in the classical test model are (a) the error and the true scores from the same test have a correlation of zero. (b) the average error score in the population of examinees is zero. This means that these random errors over many repeated measurements are expected to cancel out in the end leaving the expected mean of measurement errors to be equal to zero. Once the error is zero, the observed score is equal to the true score (X=T). (c) the error scores from parallel measurements are uncorrelated (Adegoke, 2013 & Ojerinde, 2013).

Classical test theory focused on test-level information. The two major item statistics that are used in item analysis and item selection in the development of achievement tests are item difficulty (P) and item discrimination (D). It is expected that the item difficulty and item discrimination indices are in the parallel tests. In spite of the popularity of CTT it has a lot of limitations. The person statistics (observe score) is (item) sample dependant and the item statistics (item difficulty and item discrimination) are (examinee) sample dependant. Another limitation of CTT has to do with the assumption that standard error of measurement is the same for all subjects and does not take into account variability in error at the different trait levels. Therefore, when individual differ in these extraneous variables, their performance will differ, not because of their ability but because of these extraneous variables. An awareness of the shortcoming of CTT and the benefits of IRT has led Joint Admissions and Matriculation Board to migrate from CTT to a full application of IRT in the analysis of its items (Ojerinde, 2014). Ugodulunwa (2014) emphasises that for there to be quality assurance in assessment in Nigeria there should be a shift of emphasis from CTT to IRT because of the promise it holds in solving most of the test design problems.

The primary interest of IRT is on item-level information as against that of CTT primary focus, which is on test-level information. IRT attempts to model the ability of an examinee and the probability of answering a test item correctly based on the pattern of responses to the items that constitute a test (Ojerinde, 2013). Nenty (2000) gave an alternative explanation. The score we observe (Xo) for an examinee can be resolved into that based on the ability which the test was designated to measure (Xint); and that based on other abilities (Xext), and of course, the ever present random error of measurement (Xe). When it is represented in an equation we have Xo=Xint+Xext+Xe. Nenty (2000) stated that the degree to which Xext factors influences the testing process and hence its results differ across examinees, schools, area (L.G.A). Hence, two examinees with the same ability on what is being measured may come out with significantly different scores in the same test depending on the extent to which these extrinsic factors (Xext) influence their respective performances. Such test is said to have items that shows differential item functioning (DIF). The recognition of systematic error (Xext) in IRT is a major point of deviation from CTT. There are three parameters associated with IRT; they are discrimination power (a), the difficulty parameter (b) and the guessing parameter (c). When a model contains these three parameters it is referred to as three parameter model (3P). When it contains the ‘a’ and ‘b’ parameters, it is referred to as the two parameter model (2P). However, when it contains only the ‘b’ parameter, it is referred to as one parameter model (1P). The a-parameter indicates the degree, to which examinees respond to an item, varies with, or relates to their trait level or ability. The b-parameter is the amount of trait inherent in an item. The c-parameter is the probability that a person completely lacking in the trait will overcome or answer the item correctly (Ojerinde, 2013 & Nenty, 2000).

There are four assumptions associated with IRT; (a) examinees performance on a test is a function of latent traits, or abilities. (b) the graphical relation between examinees latent traits and their probabilities of answering an item correctly is in the form of a monotonically increasing function called an item characteristics curve (ICC). (c) the probability of an examine getting an item correct is unaffected by the answer given to other items in the test (local independence). (d) the items measure one and only one area of knowledge or ability (unidimensionality). The use of IRT has helped to make sure that test measure only one unit trait, such that performance will now be dependent on examinees ability no matter where they are coming from.

Classical Test Theory

Classical test, commonly abbreviated as CTT, originates from the beginning of the 20th century. The final "classical model" was only published in the late 1960s (Lord & Novick, 1968). Classical test theory is regarded as the "true score theory". The theory started from the assumption that systematic effects between responses of examinees are due only to variation in ability of interest. All other potential sources of variation existing in the testing materials such as external conditions or internal conditions of examinees are assumed either to be constant through rigorous standardization or to have an effect that is nonsystematic or random by nature (Vander Linden & Hambleton, 1997).

The Wikipedia (2012) explained that Classical test theory assumes that each person has a true score T that would be obtained if there were no errors in measurement. A person's score is defined as the expected number-correct score over an infinite number of independent administrations of the test. Unfortunately, test users never observe a person's true score, only an observed score, X. Classical test theory assumes that each observed score (X) contains a True component (T) and an error component (E). Consequently, CTT is defined as follows

$X = T + E$, where

X = total score/observed score obtained

T = true score

E = error component.

Magno (2009) stated that CTT assumes that each individual has a true score, which would be obtained if there were no errors in measurement. However, because measuring instruments are imperfect, the score observed for each person may differ from an individual's true ability. The difference between the true score and the observed test score results from measurement error. Error is often assumed a random variable having a normal distribution.

Nenty (2005) discuss several assumptions of CTT, more than one version of the same test could be constructed. In other words, parallel tests could be constructed, such that scores from the two tests have the same standard deviation, the same correlation with the true scores and the variance on each test, which is not explainable by true score, is due purely to random error. The error scores on the two parallel tests do not relate and hence have a correlation of zero. The error scores on one of the parallel tests, and the true scores on the other test do not relate. Error and true scores from the same test have a correlation of zero and variance of the observed score is equal to the sum of the variances of the true and of the error scores.

De klerk (2008) emphasized that when measuring a psychological construct, unsystematic errors occur. These unsystematic errors could be anything, for instance distractions from outside the testing situation, physical well-being of the candidate or good/bad luck. Sometimes these influences have a positive effect on the test result; other times they have a negative influence. In other words, they cause a range of error around the true score. The true score can be seen as the systematic component of the raw score obtained. CTT assumes that the deviations occur equally to both sides of the true score. This means that the average error of measurement is zero, as the concurrent positive and negative deviations cancel each other out. He further emphasized that if we administered one test repeatedly to one candidate for an indefinite number of times, the range of measurement error is equal to the range of observed scores. In other words, the candidate will have an average score over all these repeated measures and the difference between his/her lowest score and this average score indicates the largest negative error. In this case, the error represents the situation where the candidate had the most negative external influences, which caused him/her to perform most poorly compared to the other administrations. This also work the other way around in that the largest positive error represents a situation where the candidate experienced the most positive influences that impacted positively on the performance on the test.

He continued by saying that unfortunately, in real life it is impossible in practice to have such a situation where repeated measures are possible. Furthermore, with most, if not all psychological constructs, learning, and memory processes are involved that will have a systematic, but undesirable influence on performance if a test is repeatedly administered. For instance, people could remember their previous test session and answer in a similar way, or they might figure out how to solve certain problems between test sessions and then perform better on the test the next time. (This is especially true for ability test). However, error of measurement will also average out over a large number of repeated measures, for a large group of people, where each individual has completed the test once.

Magno (2009) noted that the implication of the CTT for test takers is that tests are fallible imprecise tools. The score achieved by an individual is rarely the individual's true score. This means that the true score for an individual will not change with repeated application of the same test. The error influences the observed score to be higher or lower. Theoretically, the standard deviation of the distribution of random errors for each individual tells about the magnitude of measurement error. It is usually assumed that the distribution of random errors will be the same for all individuals. CTT uses the standard deviation of error as the basic measure of error, which is called the standard error of measurement. In practice, the standard deviation (S) of the observed score and the reliability of the test are used to estimate the standard error of measurement. The larger the standard error of measurement (Sm), the less certain is the accuracy with which an attribute is measured. Conversely, small standard error of measurement indicates that an individual score is probably close to the true score. The standard error of measurement is calculated with the formula: Sm = $s\sqrt{1-r}$. The reliability and standard error of measurement are central to CTT and with these two concepts, an estimate of the accuracy of a measurement can be obtained (De klert, 2008). The focus of CTT is on the total test score; frequency of correct responses (to indicate question difficulty); frequency of responses (to examine distractions); reliability of the test and item-total correction (to evaluate discrimination at the item level) (Impara & Plake, 1998).

At the item level, the CTT model is relatively simple. CTT does not invoke complex theoretical model to relate an examinee's ability to success on a particular item. Instead, CTT collectively considers a pool of examinees and empirically examines their success rate on an item (assuming it is dichotomously scored). This success rate of a particular pool of examinees on an item, well known as the P-value (the proportion of examinees responding in the keyed direction) of the item is used as the index for the item difficulty (actually, it is an inverse indicator of item difficulty, with higher value indicating an easier item) (Xitao, 1998; Wikipedia, 2012). The formula for doing this is:

Difficulty Index $=\frac{R}{T}$ x100 where

R =number of students who got the item right

T = total number of students tested

For CTT if the item difficulty index is the same for two populations of interest, then the item is said to be unbiased (Umobong, 2005). The ability of an item to discriminate between higher ability examinees and lower ability examinees is known as item discrimination, which is often expressed statistically as the Pearson product-moment correlation coefficient between the scores on the item (Xitao, 1998). However, the formula for finding the discrimination index for each of the items is

Discrimination index = $\frac{RU - RL}{\frac{1}{2}(T)}$ where

RU = number in the upper group who got the item right.

RL = number in the lower group who got the item right.

½(T) = half of the number of the test takers used in the item analysis.

In evaluating the quality of an assessment tool, a discussion of reliability and validity is essential. The reliability is the degree to which an instrument consistently measures the ability of an individual or group (Eluwa, Eluwa & Abang, 2011). De Klerk (2008) explained that reliability of test scores deals with the consistency of scores over replication. By this, we primarily mean the agreement between a candidate's scores when taking the test several times. The reliability or consistency of test scores can be estimated in two main ways:

1. Through repeated measures, that is, various administrations at different point in time.
    - Parallel form method (by comparing performance on two test that are parallel or equivalent/alternative form)
    - Test-retest method (by comparing several administrations of the same test)

2. Through a single measure, that is, one administration at a time.

    - Split half method (by comparing several administrations of the same text).
    - Internal consistency method (by comparing performance item by item with the same test, examples are Kuder-Richardson formula 20 and 21, Cronbach alpha, and so on).

A very high reliability figure indicates that the test has very similar items, that is it consists of quite a small number of items that all measure one very similar trait. A very low reliability figures indicates that the items are quite varied, that is the items are different from each other or might even be ambiguous (De klerk, 2008).

Validity is the degree to which an instrument measures what it is intended to measure (Eluwa et al, 2011). Validity of test scores can be described as the extent to which a test measures what it is supposed to measure (for instance, verbal reasoning ability). The most common forms of validity are face validity, construct validity, criterion validity (concurrent and predictive), content validity and so on. It is important to realize, however that a test with a high level of accuracy (reliability) does not necessarily imply that a test is measuring what it is supposed to measure. In a numerical test where very complex English was used, the test might actually measure a candidate's proficiency in English rather than his numerical reasoning ability. However, it might measure this English language ability quiet well and accurately. Therefore, the reliability of the test could be high, but the construct it is measuring is not the construct it was supposed to measure (De Klerk, 2008).

Classical test theory has the following merits:

1. Its relative weak theoretical assumptions, makes CTT easy to use and it adaptability in analysing practically all kinds of test renders it a popular choice.

2. The item statistics (item difficulty and item discrimination) has been an important tool for the measurement of psychological test.

3. Test equating can be accomplished empirically within the CTT framework (e.g. equipercentile equating) (Xitao, 1998).

4. The standard statistical packages like SPSS and SAS can be used to analyse it with ease.

However, CTT is faced with numerous limitations:

1. Item difficulty changes when a shift is made from a sample whose mean ability is high to one whose mean ability is low. Consequently, the same individual tested with two different groups of examinees may obtain two different errors of measurement and estimates of true score (Weiss & Davison, 1981).
2. Comparisons of examinees on ability measured by a set of test items comprising a test is limited to situation in which examinees are administered the same test items.
3. The classical test model does not provide basis for determining what a particular examinee might do when confronted with difficult test or speed test.
4. Izard and White (1980) stated that in classical test model two or more forms of achievement test cannot be made equivalent in level and range of difficulty.
5. It provides only one overall standard error of measurement for all the items composed in a test. Guiton and Ironson (1983) said that it is unreasonable to assume that scores throughout an entire test have the same degree of precision.
6. The reliability is defined in terms of parallel forms. The concept of parallel measures is difficult to achieve in practice.
7. It provides no basis to determine how an examinee might perform when confronted with a set of items.

(Hambleton & Swaminathan, 1985; Orluwene & Ukwuije, 2009).

Item Response Theory

The concept of Item Response Function was before 1950. Three pioneers were the Educational Testing Service psychometricians Frederic M. Lord, the Danish Mathematician Georg Rasch, and Austrian sociologist Paul Lazarsfeld, who pursued parallel research independently. Item response theory did not become widely used until the late 1970s and

1980s, when personal computers gave many researchers access to the computing power necessary for Item response theory (Wikipedia, 2012).

It is popularly abbreviated as IRT. The Item Response Theory (IRT) is also known as Item characteristic curve theory, Latent trait theory, Strong true score theory, Modern test theory and so on. Rudner (2001) stated that IRT is the study of test and item scores based on assumptions concerning the mathematical relationship between abilities (or other hypothesized traits) and item responses. Lord (1980) had earlier defined IRT as a mathematical function that relates the probability of correctly answering an item to an examinees position on the underlying ability continuum. Fulcher & Davidson (2007) emphasized that IRT rests on the premise that a test taker's performance on a given item is determined by two factors: one is the test taker's level of ability; the other is the characteristics of the item. It assumes that an observed score is indicative of a person's ability.

An individual possess a given amount of latent trait in a given subject for instance, there is latent trait for mathematics, English, Chemistry and so on. An individual has latent trait for all subject, but the latent trait is not equal for all the subjects thus the amount of latent trait an individual has in a subject depends on the level of knowledge he/she has on that subject. Unlike physical concepts that can be felt, seen, heard, tested, perceived through smelling, touching and, most human traits are latent. Being latent, such characteristics cannot be measured exactly, that is, they cannot be measured by bringing about some form of physical or direct conduct with the measurement device during the process of measurement (Nenty, 2005). In IRT latent trait is given the Greek letter θ. Its value ranges from -3 to +3. It has no zero position. The zero point is taken as the mean and standard deviation respectively. An examinee goes into the test room with his θ, the purpose of testing is to measure the amount of θ inherent in the test takers (Warm, 1978).Testing is the amount of latent trait in an individual. During testing, there is an encounter between the latent demanded by the test item and the amount of latent trait possess by the individual. If the amount of latent trait possess by the individual is lower than the amount of latent trait demanded by the test item, the individual will not be able to overcome the item i.e. he will fail the item. In testing an individual observe score (X_i):

$X_i = \theta_i + \lambda i + \epsilon_i$

θ_i = true ability component for the individual

λ_i = Is the systematic error component.

ϵi = Random Error

A random error is an error that is not operating in one way and can be eliminated by the process of sampling. While systematic error is the one that is operating in a given direction. The recognition of systematic error in IRT is a major point of deviation from classical test theory. Systematic errors are those aspects of error which when present in a test gives advantage to a group of test takers and disadvantage to another group of test takers. Some of the sources of systematic error (λi) could be the languages of the test item, culture, race, gender and so on.

Nenty (2000) gave an alternative explanation. The score we observe (X_0) for an examinee can be resolved into that based on the ability which the test was designated to measure (X_{int}) ;and that based on other abilities (X_{ext}), and of course, the ever present random error of measurement (X_e). When it is represented in an equation, we have

$X_0 = X_{int} + X_{ext} + X_e$

Sources of X_{ext} to include; the extent of familiarity with the languages and culture of the test, level of familiarity with the test stimuli, test sophistication motivation to perform well; value attached to rapid performance rapport with examinee (Anastesi, 1988) ; as well as the ability to guess. Others are: degree to which test instruction has been understood and followed, the extent to which the strategy of solving problem has been understood and followed (Van der flier, 1977), cheating or copying during testing, differences in content colleges and institutional emphasis, background experience factors, institutional history and poor testing scoring (Nenty 2000). Others are no equal scoring format, no equal access to relevant textbooks, instruments, laboratories, and workshop.

Nenty (2000) stated that the degree to each of these extrinsic factors influences the testing process and hence its results differs across examinees, schools, classrooms, sex of examinees, school location, local government areas (L.G.A) or districts. If a test through its designs, administration and scoring, is unable to control for these extrinsic factors, the differences in test scores observed for examinees might not reflect their abilities on which it is being measured, which is represented by X_{int}. Hence, two examinees with the same ability on what is being measured may come out with significantly different scores in the same test depending on the extent to which these extrinsic factors influence their respective performance. Such test is said to have items that shows DIF, scores from them are not valid, and decisions taken from them are questionable.

The primary interest of IRT is in whether an examine gets each individual item correct or not, rather in the total raw score. Each item of a set of items measures the underlying traits or traits (Verstralen, Bechger & Maris, 2001). It assumes that the probability of a given response is determined by both the person's ability and the item's difficulty (Orluwene & Ukwuije, 2009). It is worthy to note that as the difficulty value of the item rises, a test taker must be more able to ensure a higher probability of getting the item correct (Haiyang, 2010).

Morales in 2009, puts forward two primary postulates of IRT: first, a more able person should have greater possibilities of success on assessment item than a less able person. Secondly, any person should be more likely to do better on an easier item than on a more difficult one. Yen (1992) gave three major characteristics of IRT as it focuses on performance on individual items, rather than only on intact test, it describes item performance at each level of student's ability and it is model based.

The Assumptions of Item Response Theory (IRT)

1. Unidimensionality:

Only one ability or trait is able to elicit an examinees test performance (Hambleton & Swaminathan, 1985). In other words, each item should measure not more than one latent trait. From this, each item should elicit the same response from all examiners with the same ability on the trait under measurement, irrespective of other factors on which they may differ (Nenty 1979 as cited in Nenty 2005). If a test is not unidimensional, the scores that result from it are more or less meaningless. This is because it is not known to which of the many traits that sustained the responses to its items (Lord & Novick, 1968). The unidimentionality of a scale can be evaluated by performing an item-level factor analysis.

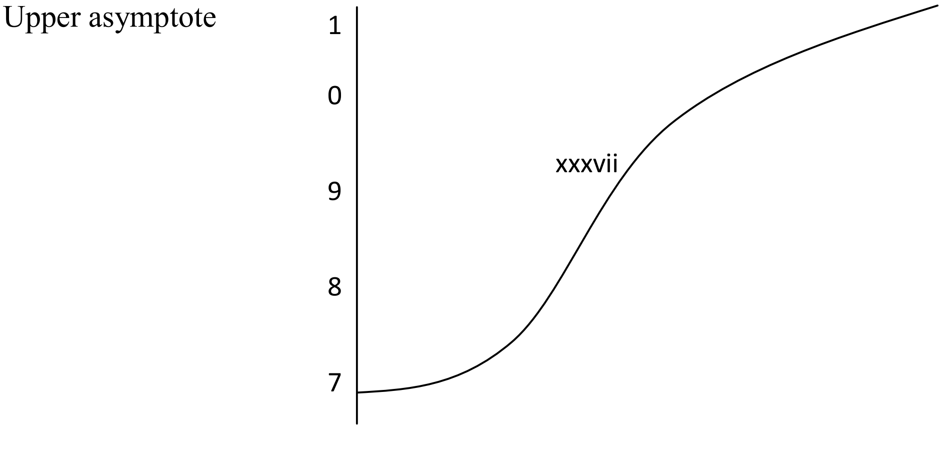
2. Local Independence

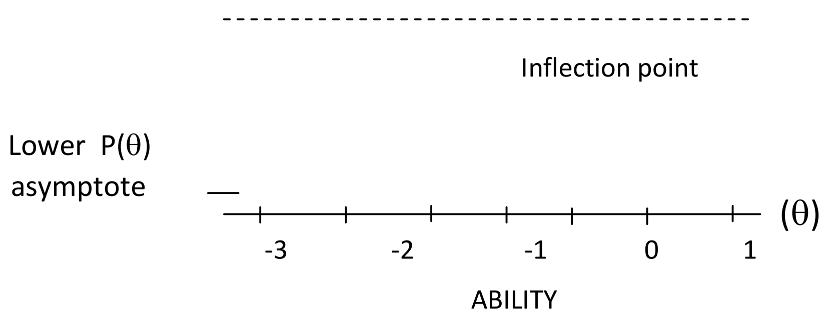
It means that the probability that a student will answer correctly any two items is the product of the probabilities that the students will answer correctly each separate item, and the psychometric contribution of an item to a test can be evaluated without knowledge of the other item in the test (Yen, 1992). It also means that the only relationship among the items is explained by the conditional relationship with the latent variable θ. That is, if the trait level is held constant, there should be no association among the item responses (Thissen & Steinberg, 1988). Simply put, the probability of a test taker getting an item correct in a test is unaffected by the answer given to other test items in the test. Hence, test item should be constructed in a way that no one item gives an insight to the answer to another item.

3. Normal ogive assumption

If one plots the probability of the examinees giving a correct answer to an item P(θ) as a function of ability, the result would be a smooth S-shaped curve, which is the shape of a normal ogive. In IRT it is referred to as item characteristic curve (ICC). Each item in a test will have its own ICC.

Figure 1: The item characteristic curve of an item





The probability of correct response is near zero at the lowest level of ability -3 (lower asymptote). It increases until at the highest level of ability +3, the probability of correct response approaches 1. The inflection point is the rising point of the ICC, which represents the difficulty index of the item.

Baker (2001) explained that the ICC is the basic building block of item response theory; all the other constructs of the theory depend upon this curve. Two technical properties of an ICC are used to describe it. The first is the difficulty of the item, which describes where the item functions along the ability scale. For example, an easy item functions among the low-ability examinees and a hard item functions among the high-ability examinees; thus, difficulty is a location index. The second technical property according to him is discrimination, which describes how well an item can differentiate between examinees having abilities below the item location and those having abilities above the item location. This property essentially reflects the steepness of the ICC on its middle section. The steeper the curve, the less the item is able to discriminate since the probability of correct response at low-ability levels is nearly the same as it is at high-ability level.

According to Baker (2001) when the item discrimination is less than moderate, the ICC is nearly linear and appears rather flat. When discrimination is greater then moderate, the ICC is S-shape and rather steep in its middle section. When the item difficulty is less than medium, most of the ICC has a probability of correct response that is greater than 0.5. When the item difficulty is greater than medium, most of the ICC has a probability of correct response less than 0.5. Regardless of the level of discrimination, item difficulty locates the item along the ability scale. Therefore, item difficulty and discrimination are independent of each other.

Item Response Theory Models

IRT models have been developed to deal with responses to either items that are scored in a dichotomous or a polytomous form. The models for dichotomous scoring pattern are the one parameter model (Rasch model), two parameter logistic model and three parameter logistic model. The models for polytomous scoring pattern are graded model, nominal model, partial credit model and rating scale model. There are three parameter needed to differentiate among the one parameter model (1P), two parameter model (2P) and three parameter model (3P). The΄B̀ parameter is known as the difficulty parameter. This value tells us how easy or how difficult an item is.΄A` parameter is called the discrimination parameter. This value tells us how effectively the item can discriminate between highly proficient students and less proficient students, the΄C` parameter is known as the΄G` parameter or the guessing parameter. This value tells us how likely the examinees are to obtain the correct answer by guessing (Chong, 2010). The one-parameter model uses only the΄B` parameter. The equation for the one-parameter logistic model or Rasch model is

$P(\theta) =1/1+e^{-1(\theta-b)}$

The two-parameter logistic model uses both the΄A` and΄B` parameters. The equation for the two-parameter logistic model is

$P(\theta) = 1/1+e^{-L} = 1/1+e^{-a(\theta-b)}$

Where: e is the constant 2.718

b is the difficulty parameter

a is the discrimination parameter

$L = a(e^{1}-b)$ is the logistic deviate (logit)

e is an ability level

The three-parameter logistic model uses΄A`,΄B` and΄C` parameters. The equation for the three-parameter logistic model is

$P(\theta) = c+(1-c)1/(1+e^{-a(\theta-b)})$

Where; c is the guessing parameter

A side effect of using the guessing parameter c is that the definition of the difficulty parameter is changed (Baker, 2001). The choice of IRT model is data dependent. Embretson and Reise (2000) suggest one should use the Rasch family model when each item carries equal weight in defining the underlying variable, and when strong measurement model properties are desired. If one desires fitting an IRT model to existing data or desires highly accurate parameter estimate, then a more complex model such as the two-parameter logistic model should be used (Reeve, 2012). The 2P is equivalent to the 3P model with c=0 and is appropriate for testing items where guessing the correct answer is highly unlikely, such as fill-in-the-blank item or where the concept of guessing does not apply, such as personality, attitude, or interest items (Wikipedia, 2012).

As with any use of mathematical model, it is important to assess the fit of the data to the model. Misfit is an observation that cannot fit into the overall structure of the examination. Misfit can be caused by many reasons, for example confusing distractors in a multiple-choice test, or if a test developer attempts to create an exam pertaining to Nigerian history, but accidentally an item about American history was included. In the context of CTT, this type of item is typically detected by either point-biserial correlation or factor analysis. In IRT it is identified by examining the misfit indices. If an item is identified as misfit, such item may be removed from the test form and rewritten or replaced in future test forms. If, a large number of misfitting items occur with no apparent reason for the misfit, the construct validity of the test will need to be reconsidered and the test specifications may need to be rewritten. Thus, misfit provides invaluable diagnostic tools for test developers, allowing the hypotheses upon which test specifications are based, to be empirically tested against data (Wikipedia, 2012 and Chong, 2010).

Item response theory has the following advantages:

1. It provides item level information at each level of student ability.

2. It estimates of examinees ability do not depend on the pool of items administered.

3. It estimates of item level information do not depend on the sample of examinees.
4. It focuses on performance on individual items, rather than only on intact tests.
5. It is model based and provides strong assumptions, which make IRT to be superior to other measurement theories.
6. In the place of reliability, IRT offers the test information function, which shows the degree of precision at different values of theta, θ which makes it clear that precision is not uniform across the entire range of test scores.
7. Item and ability parameters are invariant under a linear transformation (i.e. it is possible to change the means and variance estimates for different subgroups so that they lie on the same metric).
8. Its applications are numerous: item and scale analysis, DIF, instrument equating and computerized adaptive tests.

The notable limitations of IRT are as follows:

1. It is complex and highly mathematical; many psychometricians find it very difficult to understand the principles, postulates, and assumptions behind IRT.
2. The software for manipulating IRT is not readily available most especially in Nigeria.
3. In cases where this software is available, it is very difficult to operate.
4. It is difficult to use IRT models when items do not fit into it. In other words, it is very selective.
5. IRT needs large samples and relatively large number of items.

Comparison of Classical Test Theory and Item Response Theory

Classical test theory (CTT) and Item response theory (IRT) are largely concerned with the same problems but are different bodies of theory and therefore entail different methods. Although the two theories are generally consistent and complementary, there are a number of points of differences (Wikipedia, 2012). It is worth also mentioning some specific similarities between CTT and IRT, which will help to understand the corresponding differences between them. Lord (1980) showed that under the assumption that θ is normally distributed; discrimination in the 2P model is approximately a monotonic function of the point-biserial correlation. In particular, if the assumption holds, where there is a higher discrimination there will generally be a higher point-biserial correlation. Another similarity is that while IRT provides for a standard error of each estimate and an information function, it is also possible to obtain an index for a test, as a whole, which is directly analogous to cronbach,'s alpha, just as CTT does (Andrick, 1982).

However, there are numerous differences between CTT and IRT with IRT having a number of advantages over CTT. Wikipedia (2012) identify some specific differences between CTT and IRT: IRT makes stronger assumptions than CTT and in many cases provide correspondingly strong findings. These results only hold when the assumptions of the IRT models are actually met. Although CTT results have allowed important practical results, the model-based nature of IRT affords many advantages over analogous CTT finding. CTT test scoring procedures have the advantage of being simple to compute (and to explain) whereas IRT scoring generally requires relatively complex estimation procedure. IRT provides several improvements in scaling items and people. The specifics depend upon the IRT model, but most models scale the difficulty of items and the ability of people on the same metric. Thus, the difficulty of an item and the ability of a person can be meaningfully compared. Another improvement provided by IRT is that the parameters of IRT models are generally not sample-or test-dependent whereas true-score is defined in CTT in the context of a specific test. Thus, IRT provides significantly greater flexibility in situations where different samples or test forms are used. These IRT findings are foundational for computerized adaptive testing.

In addition, CTT yield only a single estimate of reliability and corresponding standard error of measurement, whereas IRT models measure scale precision across the underlying latent variable being measured by the instrument (Hays, Morales, & Reise, 2000). Another disadvantage of CTT method is that a participant's score is dependent on the set of questions used for analysis, whereas, an IRT-estimated person's trait level is independent of the questions being used (Reever, 2012). The IRT estimates score is sensitive to differences among individual response pattern and is a better estimate of the individual's level on the trait continuum than CTT's summed scale score (Santor & Ramsay, 1998).

Harvey and Hammer (1999) explained that one of the potentially most important differences between CTT and IRT concerns the issue of administrative efficiency (i.e., reducing testing time) and item-banking (i.e., developing calibrated item pools from which subtests of items can be selected for each individual tested). Whereas CTT-based indices of test functioning- and especially, scoring- are fundamentally based on the assumption that the entire item pool is going to be administer to each examinees, IRT-based methods can easily deal with the situation in which different examinees are presented with entirely different listings of items, or different numbers of items. This is because scoring methods used in IRT to estimate each examinee's θ score can produce estimates that lie on a common θ score metric. In contrast, the "number right" scoring methods typically used in CTT-based approaches are highly dependent on having the same list of items been presented to each examinee.

The CTT gives a very simple way of determining the validity and reliability of test. The classical item analysis provides us a way of doing this. By subjecting the whole test results to simple statistical tests, one can determine the validity and reliability of the test. On the other hand, IRT offers a more complex but more reliable way of determining validity and reliability of test. If the focus of CTT is on the test as a whole, IRT focuses on each item and each individual test takers (Morales, 2009). CTT requires minimum sample of 200 to 500 while IRT needs a minimum sample of 500 to 1000 (Hermandez, 2009). In addition, IRT requires a relatively large number of items.

Generally, measurement of precision is fixed for all scores in CTT but it veries across scores in IRT. There are graphical tools for item and scale analysis in IRT. In CTT mixed item formats lead to unbalanced impact on total scales but IRT easily handles mixed item formats. Longer scales increases reliability in CTT but this is not so in IRT, since both short and longer scales can be equally reliable. CTT summed scores are on ordinal scale while that of IRT are on interval scale.

Magno (2009) carried out a work to demonstrate the difference between CTT and IRT using derived test data. The sample was made up of 219 students. The instrument was made up of 70 items in chemistry. The result demonstrates certain limitations of the CTT and advantages of using the IRT. The IRT estimates of item difficulty do not change across samples as compared with CTT, difficulty indices were also more stable across forms of test than the CTT approach, IRT internal consistencies are very stable across sample while CTT internal consistencies failed to be stable across samples and IRT had significantly less measurement error than the CTT approach. Morales (2009) in his work on evaluation of mathematics achievement test: a comparison between CTT and IRT, he used 80 students with a mathematics achievement test consisting of 40 multiple-choice test items. The result shows that items, which were found to be “bad item” in CTT, came out not fitting also in Rasch model.

Eluwa, Eluwa and Abang (2011) carried out a study titled evaluation of mathematics achievement test: a comparison between CTT and IRT. They used a sample of 80 students. The mathematics achievement test was made up of 40 multiple-choice test items. The result showed that although CTT and IRT methods are different in so many ways outcome of data analysis using the two methods in this study did not say so. Items, which were found to be “bad item” in CTT, came out not fitting also in the Rasch model.

Progar and Socan (2008) carried out an empirical comparison of item response theory and classical test theory. A data set from the third international mathematics and science study was used. The findings indicated that the CTT and IRT item/person parameters are very comparable. The CTT and IRT item parameters show similar invariance property when estimated across different groups of participants that the IRT person parameters are more invariant across different item sets, and that the CTT item parameters are at least as much invariant in different item set as the IRT item parameters. The results furthermore, demonstrated that concerning the invariance property, IRT item/person parameters are in general empirically superior to CTT parameters but only if the appropriate IRT model is used for modeling the data.

A proper empirical comparison between CTT and IRT demands that the necessary conditions that will yield good results from these theories must be adhered to. The correct sample size and model for IRT must be applied. In addition, the data must fit into the IRT model to be used. A careful look at the empirical studies discuss above shows that in most cases the sample is usually below 500 and also most of the studies limited their used of IRT model to that of Rasch which is a one parameter model. This is in line with the comment of many researchers (Angoff, 1993; Camilli & Shepard, 1994) who believe that investigation of DIF in the framework of Rasch measurement is limited. Non-consideration of the possible differences in respect to discrimination power or differences in respect to pseudo-guessing will result in undetected DIF items and will lead to the ultimate removal of the most useful items in a measure (Angoff, 1993). Therefore, applications of the Rasch model limit researchers understanding of group differences in responding to item in a measure (Reever, 2012).

Item Response Theory, Classical Class Theory and Differential Item Functioning

Systematic errors are different from unsystematic errors. A systematic error refers to a characteristic of the test or the testing situation that will affect all measurements equally (De Klerk, 2008). For example, if there is a mistake in one or more of the test items that is presented to all the candidates completing the test, it will influence all the candidates in the same way. As psychological tests are mainly used to determine individual differences, the influence of systematic errors is unimportant and is not included in the CTT concepts (De Klerk, 2008). However, it is important to note that when the performance of candidates who experience some systematic error on a test is compared to the performance of candidates who completed a test free from such error, the comparison will be unfair. This is referred to as differential item functioning. A major advantage of IRT to CTT is that IRT is interested in systematic error, which can lead to DIF (Odili, 2003).

Harvey and Hammer (1999) stated that one of the important issues faced by counselling psychologist is that of responding to the diversity of clients. In particular, it is important that the tests used by counselling psychologists be free of systematic demographic subgroup bias. IRT techniques provide a powerful means of testing item for bias, using what is known as DIF, as well as assessing the cumulative effect of any item-level bias on the test's total score. In contrast, CTT based methods of assessing bias are fundamentally limited, especially approaches that base their assessment of bias on the presence of group mean differences in total tests scores across demographic groups or on differential item-passing/endorsement rates between subgroups (Drasgow, 1987). He further said that in essence, such methods cannot distinguish between the situation in which the subgroups have different means, and the test is biased, versus the means difference, but the test is not biased. Previously, Hunter (1975) and Lord (1977) have demonstrated that bias techniques based on CTT such as p-value differences or point-biserial correlation coefficients produces invalid indices of bias in the presence of group mean differences. Other variable beside item bias can contribute to mean differences.

According to Abedalaziz (2011) invariance means that item parameters (e.g. difficulty, discrimination and guessing) are not dependent on the ability distribution of any particular group of examinees and the examinee ability parameter (θ) is not dependent on a specific set of test items. This implies that for a correctly specified IRT model, the ICC for two subpopulation examinees will be the same regardless of the groups' ability distribribution (Humbleton, Swaminathan & Rogers, 1991). This property makes IRT an attractive framework for examining DIF since the occurrence of non-coinciding ICCs is an indicator of Differential item functioning between two groups (Abedalaziz, 2011)

Warm 1978 explained that studies of item bias using CTT often compare the P-values for one group with the p-values of another group. Item with significantly different p-values between the two groups are thought to show DIF. Such an approach is inappropriate because the method assumes that the two groups have the same average ability that is probably false if the groups are matched on moderator variables such as educational levels, since the quality of education varies considerably from school to school. He also said that the comparison of p-values across groups assumes that bivariate distribution of the p-values is linearly related but under CTT p-values are not linearly related. The same is true of other CTT item parameters, such as 'corrected' p-values, the inverse normal transformation of the p-values, and 'delta' (Lord & Novick, 1968). Therefore, the use of CTT to detect DIF may be in appropriate. A better and more reliable theory to use is Item Response Theory.

The concept of Differential Item Functioning

The extent to which a test measures what it purports to measure, or the extent to which specified inferences from the test's scores are justified is of paramount importance to psychometricians. In measurement, whether physical or mental, some errors are involved. Jencks et al (1979) as cited by Umoinyang (2011) classified these possible errors into three broad categories: conceptual errors, consistent errors and random errors. Conceptual errors are committed when a wrong measuring instrument is used to measure an attribute. For instance, when mental ability is measured using some test in vocabulary, it is obvious that an inference from the scores of that test will be bias. Consistent, systematic, or extraneous errors are those aspects of error which when present in a test gives advantage to a group of test takers and disadvantage to a group of test takers. A random or non-systematic error is an error that is not operating in one way. It could be because of temporary fluctuations in respondents, interview settings and so on. Umoinyang (2011) further observed that testing practices have tried to reduce random and conceptual errors but consistent errors in achievement test have not been addressed because construct validation is construed not to be a priority of achievement testing because it is conceived to be content based. Consistent error typifies itself in differential item functioning (DIF).

Odili (2010) also observed that recognition of extraneous error variable in test performance is one of the major shifts in Item Response Theory (IRT) of measurement from Classical Test Theory (CTT). Extraneous error variables are those errors that can bring about difference in performance of test takers in a test item other than their ability in the trait that is being measured. He explained that an item writing process that fails to check for influence of the sources of extraneous error variable will give rise to test items that will differentially function for different subgroups of test takers.

DIF occurs when testees from different groups who have been match on ability levels are performing differently in test items. Atar (2006) explained that it is critical that test items do not differentiate among examinees based on their gender, race, and ethnic background but rather differentiate between them based on their abilities. A fair test is one that is comparably valid for all groups and individuals and that affords all examinees an equal opportunity to demonstrate the skills and knowledge, which they have acquired and which are relevant to the test's purpose (Roever, 2005). The presence of large numbers of items with DIF is a serious threat to the validity of a test and any inference made from such test scores may not be valid.

Odili (2010) revealed that interest in analysis of differential item functioning in test derives from the consideration that nations all over the world perceive education as instrument for achieving egalitarianism among persons. Achieving this demands that test items should measure traits which, are taught in school subjects and not those traits that are alien to it. He further revealed that, the violation of this reasoning was responsible for criticism of use of tests in United States of America. The argument was that tests items discriminated unfairly against minority groups. The result of such criticism gave rise to legislations that sought to protect the minority groups in the use of the test results as well as taking steps to detect and reduce DIF.

The precise definition of DIF varies across methods, and according to whether binary or polytomous items are being examined. However, DIF can be defined broadly as conditional probabilities or conditional expected item scores that vary across groups (Teresi, 2004). According to Crane et al (2007), DIF could be defined as the different probability of giving the right answer to a test item by two individual with the same ability but from different groups. Augemberg and Morgan (2008) further put it that, DIF is observed when comparable (i.e., matched on ability) examinees from different groups have a different probability of answering a given item correctly. Thus, DIF implies that even after controlling for ability, an item appears to be more difficult for examinees from one group, as compared to examinees in other groups.

Fidalgo (2011) observed that suppose we have a test intended to measure spatial ability but having items that, because they are written in a complicated manner, also assess linguistic ability. Suppose this test is administered in schools having children who are native speakers and children who are immigrants and are still learning the language. It is easy to see that the former will do better than the later. Odili (2010) also stated that Differential item functioning (DIF) is the tendency of test takers of the same standing in the latent trait to perform differently in a test item.

There are two main types of DIF, namely uniform DIF and non-uniform DIF. Uniform DIF is said to occur when difference in correct response probability are found across all ability level for a particular item. In order words, it occurs when the item is more difficult at all ability levels for one group than the other. On the other hand, non-uniform DIF occurs when there is an interaction between the ability and group membership such that an item may seem difficult for those at the higher level in one group and after a particular point, it becomes more difficult for those at the lower level in the other group. In item response theory (IRT) uniform DIF occurs when two item characteristics curves (ICC's) differ but are more or less parallel to each other while non-uniform occurs when the ICC's for the two subgroups cross at some θ value. Before the cross over point, the item is favouring one subgroup and after the ICC's cross, the item starts to favour the other group. Uniform DIF is likely to occur when two ICC's have different b (difficulty} parameters and similar a (discrimination or slope) parameters. Non-uniform DIF is likely to occur when the two ICC's have similar 'b' parameters and different 'a' parameters (Swaminathan & Rogers, 1990).

In 2005, Zumbo & Gelin identifies some uses of DIF as:

1. Fairness and equity in testing
2. Dealing with a possible "threat to internal validity"
3. Trying to understand the (cognitive and /or psych-social) processes of item responding, test performance, and investigating whether these processes are the same for different groups of individuals.

Distinction between Terms: Impact, Item Bias and Differential Item Functioning

It is necessary to make a distinction among some related terms such as impact, item bias and DIF Item bias as expressed by Camilli & Shepard (1994) is an indication of serious errors or flaws in the measurement of ability for members of a specific group. Item bias can also take place when performance on a test requires other knowledge different from those the test is supposed to measured, thereby giving rise to test scores that are less valid for a particular group. Test item are biased when they contain sources of difficulty that are irrelevant or extraneous to the construct being measured, and these extraneous or irrelevant factors affect performance.

Zumbo (1999) explained that item impact is evident when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item because there are true differences between the groups in the under lying ability being measured by the item. Consequently, a difference in the performance of groups of examinees with different abilities on specific item is not indicative of test bias, but item impact (Schumacher, 2005).

DIF studies focus on the identification of item with differential performances. Upon identifying those items, the next step is to expose such item to further item bias analysis (e.g. by empirical evaluation or content analysis) so as to determine the potential causes of DIF as well as whether it is item impact or item bias. An item might show DIF, but not considered biased if the difference is because of the actual difference in the groups' ability to respond to the item (i.e., if one group of test takers is at a high level than the other group). It is only when differences in a group's ability to respond to a test item are caused by construct-irrelevant factors can DIF be considered as bias (Roever, 2005; Zieky, 2003; Zumbo, 1999)

An item is said to flag DIF:

1. If it contains language or content that is unequally difficult for different subgroups of test takers.
2. When the test item, item stem, test instruction or distracter is not good enough or/and can be understood in more than one way by the test takers.
3. If there are no equal learning opportunities, so much that one group is more exposed to the information being tested than the other group.

4. No equal access to relevant test books, instruments, equipment, laboratories and workshops.

5. When there is no equal scoring format for the test takers.

6. When a topic is of greater importance to one group than the other is.

However, an item can only be considered bias if it shows 1 and/or 2 above. Bias has to do with what is in the questions that tend to favour one group against the other. In other words, an item must show differential item functioning before it can be said to be bias. That is why test takers ability is matched in other to reduce 3 to 6 above, such that DIF could easily be used to pick out bias items. Nevertheless, an item that shows DIF does not necessarily mean it is bias. Yet, all items that show bias are DIF items. No statistic can determine whether or not a test item is biased, DIF helps us to sort out items that may be unfair which are then subjected to further analysis to find out whether they are bias or not. Roever (2005) wrote that, bias is usually a characteristic of a whole test, whereas DIF is a characteristic of an individual item.

Procedures for Detecting Differential Item Functioning

1. Locate examination items, where one group performs better than the other does.

2. Examinees are usually divided into two groups for comparison: focal group and reference group. The focal group can be defined as the group of interest and the reference group can be defined as the group that is to be used for the comparison. For example, females maybe the focal group males may be the reference group in a DIF study.

3. In order to distinguish DIF from item impact, a matching variable is required to match examinees on the underlying construct of interest (e.g. mathematics achievement, mental ability, spatial ability) so as to compare performance across groups. Atar (2006) identified two types of matching variable- Internal matching variable and external matching variable. When the performance on a test that the DIF is being study is used as the matching variable, it is referred to as internal matching variable. On the other hand, when the performance on another test that measures the same construct with the item of

interest is used as the matching variable, it is referred to as external matching variable. Note that examinees' ability levels (performance levels) are based upon their total scores on the examination. As such, the DIF analysis of one specific test item is as independent as possible from the DIF analysis of the other test item (Zumbo, 1999).

4. The type of scoring format used for a particular test usually determines the method of DIF to use. The two most commonly used scoring formats for tests are binary and ordinal. Binary scores are also referred to as dichotomous item responses and ordinal item responses are referred to as graded response, likert-type, or polytomous. The ordinal formats are commonly found in personality, social or attitudinal measures. It is important to note that it is not the question format that is important here but the scoring format. Items that are scored in a binary format are either:

   I. Items (e.g., multiple choice) that are scored correct/ incorrect in aptitude or achievement tests

   II. Items (e.g. true/ false, yes/no) that are dichotomously scored according to a scoring key in a personality scale (Zumbo. 1999).

Differential Item Functioning (DIF) Method

Method for detecting DIF has proliferated in recent years. Teresi in 2004 reported that differences among DIF methods could be characterized according to whether they:

1. Are parametric or non-parametric.

2. Are based on latent or observed variable.

3. Treat the disability dimension as continuous.

4. Can model multiple traits.

5. Can examine polytomous responses.

6. Can detect both uniform and non-uniform DIF.

7. Can include covariates in the model.

8. Must use a categorical studied (group variable).

Atar (2006) characterized the method in the following ways.

Non-Item response DIF procedures for dichotomously scored items

1. Mental-Haenszel procedure
2. Standardization procedure
3. Logistic Regression procedure
4. Simultaneous item bias test procedure (SIBTEST)

Non-Item response based DIF procedures for polytomously scored items

1. Mantel procedure
2. Generalized mantel-Heanszel procedure
3. Standardized mean difference procedure
4. Ordinal logistic Regression procedure
5. Poly-SIBTEST procedure

Item Response based DIF procedures

1. IRT likelihood-Ratio test procedure
2. Kamata's multilevel Rasch model
3. Multilevel logistic Regression model
4. GLLAMM procedure
5. Two-Parameter logistic IRT model
6. Graded Response model

Umoinyang (2011) characterized the method in the following ways.

1. Item-Parameter related methods
    - Transformed item difficulties-Major Axis(TID-MA)
    - Transformed item difficulties-45^0 line(TID-45^0)
2. The chi-square χ^2/Probability- Related methods
    - Cochran's chi-square(CTχ^2) approach
    - Mantel Haenszel (M-H) statistics
3. Analysis of variance, regression, and log-linear related methods.
    - This includes all the methods which involve point-biserial correlation, test re-test reliabilities, inter-correlation among test items as well as groups item interaction derived from analysis of variance, The logit model is also part of it.
4. Method based on IRT
    - 3 parameter item characteristics curve (ICC-3)
    - One-parameter item characteristics curve (ICC-1)

The American Board of Internal Medicine (2012) subdivided the method in the following ways.

1. Mantel-Heanszel: condition on raw score, statistical test of contingency tables
2. Logistic Regression: condition on raw score, model group-response relationship.
3. IRT methods: condition on ability (θ) compare item parameters on ICCs
    a. Compare item parameter estimates
        - Multivariate test (b, a, and c)
        - T-test on b-values

b. Area Method

- Total area
- Squared differences
- Weighted areas and differences

Several methods can be used to detect DIF in a test item. These methods could be characterized according to the two measurement theories namely, Item response theory (IRT) and Classical test theory (CTT). Some of the DIF procedures based on CTT are the transformation item difficulty (TID), point biserial correlation, P-value differences and so on. The DIF procedures based on IRT are the item characteristic curve (ICC), IRT-likelihood test, Rasch logistic model, 2- parameter logistic model, 3-parameter logistic model and so on.

If all the bias approaches were to identify the same items as biased, one could use the simplest and least expensive approach. However, if the approaches identify different items as biased, it becomes necessary to determine those methods which are most valid (Ironson & Subkoviak, 1979). A researcher carrying out DIF studies must put into consideration the types of data he is using, the scoring format, type of DIF needed (uniform or non-uniform), the variable (latent or observed), and the sample size before deciding on the method of DIF detection he wants to employ. It is usually better to use more than one method to give room for comparison of results. A table showing a summary of the different methods for detecting DIF and their characteristics is shown in appendix o.

Guidelines on Selection of DIF Methods to Use

1. Use parametric procedure if your data fits the model's assumptions if not, either another model is chosen or a non-parametric method is used instead.
2. The contingency table approaches have the advantages that they required small sample sizes when compared with IRT based model. It became problematic to use the IRT based model when the sample size of the focal group is small.

3. A researcher carrying out DIF studies must put into consideration the types of data he is using, the item scoring format, the type of DIF needed (uniform or non-uniform), the matching variable (latent or observed), the sample size and the availability of the software of the procedures before deciding on the procedure of DIF detection he wants to employ.

4. It is usually better to use more than one procedure in other to give room for comparison of results.

Mantel- Haenszel Method (M-H)

Mantel-Haenszel method (Mantel-Haenszel, 1959) is one of the most popular method used in detecting DIF, It is a non-parametric statistic using chi-square to test the null hypothesis of no relationship between the test performance on a given item and group membership. The null hypothesis can be expressed as H_o: $A_k/B_k = C_k/D_k$

The M-H statistical procedure consists of comparing the item performance of two groups (reference and focal) whose members were previously matched on the ability scale. The matching is done using the observed total test score as a criterion or matching variable (Holland & Thayer, 1988). The Mantel-Haenszel statistic is based on a contingency table analysis. For dichotomous items, K contingency table (2x2) is constructed for each item. The table of the data layout for the M-H method (see appendix o) shows the frequencies of correct response for reference and the focal groups.

For the k^{th} level of the matching variable, N_{1k} and N_{0k} are the number of examinees who answer the studied item correctly and incorrectly, respectively, N_{rk} and
N_{fk} is the number of examinees in the reference group and the focal group, respectively, and N_k is the total number of examinees.
A_k is the frequency of correct response in the reference group, B_k is the frequency of incorrect response in the reference group, C_k is the frequency of correct response in the k. focal group, and D_k is the frequency of incorrect response in the focal group for the k^{th} level of the matching variable.(Atar, 2006)
The M-H test statistic (from Atar, 2006) has this for

$$\text{MH-}X^2 = (|\sum_{k} A_k - \Sigma_k E(A_k)| - .5)^2/\Sigma_k \text{var}(A_k)$$

where

$$E(Ak) = \frac{N_{rk}N_{1k}}{Nk}$$

$$Var(Ak) = \frac{N_{rk}N_{fk}N_{1k}N_{0k}}{N^2{}_k(N_k\text{-}1)}$$

The common odds-ratio is calculated by

$$\alpha_{MH} = \frac{\Sigma_k A_k D_k/N_k}{\Sigma_k B_k C_k/N_k}$$

If this index is less than one, it indicates possible bias against the focal group. On the other hand, if the index is greater than 1, it indicates possible bias against the reference group (Atar, 2006).

Holland & Thayer (1988) proposed a logarithmic transformation of α for interpretive purpose, with the aim of obtaining a symmetrical scale in which a zero value indicates an absence of DIF, a negative value indicate that the item favour the reference group over the focal group, and a positive value indicates DIF in the opposite direction. This transformation is expressed as $\Delta\alpha_{MH}$= -2.35ln (α_{MH}).

To assess the degree of DIF present, the odds-ratio estimator can be transformed into the ETS (Educational Testing Service) "delta metric", which classify items as one of these three types.

A. Negligible DIF, where X^2 is non-significant or the absolute value of Δ is less than 1.0
B. Intermediate (moderate) DIF, when X^2 is significant and Δ ranges from 1.0 to 1.5 in absolute value
C. Large DIF, where X^2 is significant and the absolute value of Δ is more than 1.5

Roever (2005) advised that for test construction, A items are preferred, B items can be used where they are not enough A items and/or due to test specifications, but the use of C items requires an argued case.

The detection of DIF by the method of Mantel-Heanszel (M-H) is done by following these steps:

1. Divide examinees into two groups for comparison say focal group and reference.
2. Use the total scores of the test that the DIF is being studied as the matching variable to group the examinees into matching levels.
3. State the null hypothesis.
4. Determine A_k and C_k by counting the number of examinees that got the item correct in the reference and focal group respectively for each of the matching levels.
5. Determine B_k and D_k by counting the number of examinees that got the item wrong in the reference and focal group respectively for each of the matching levels.
6. Determine $E(A_k)$, $E(B_k)$, $E(C_k)$ and $E(D_k)$.
7. Calculate the M-H chi-square using the formula.
8. Find the table value of df = 1 at .05
9. If X^2-calculated is equal to or greater than X^2-table value the null hypothesis is rejected but if X^2-calculated is less than X^2-table value the null hypothesis is accepted.
10. The common odds-ratio is determined. When it is less than 1 it indicates a possible bias against the focal group but when it is more than 1 it indicates a possible bias against the reference group. While a value of 1.0 signifies no DIF.
11. The delta of the common odds-ratio is determined to indicate whether DIF is negligible, moderate, or large.

See appendix B for an illustration.

The WINSTEP statistical package can be used to carry out M-H analysis, see appendix B

Advantages of M-H Method

1. It is simple and easy to implement, it does not require highly specialized soft ware; it can be computed with SPSS through cross tabs with the grouping variable (gender, language) in rows, the item in columns and the matching variable (scores) as a layer.
2. It's availability of a hypothesis testing is a plus mark.
3. The M-H method is ideal because it does not rely solely on the X^2 statistic, which can be overly sensitive when large samples are used, which is customary in DIF analysis.
4. The Δ statistic not only complements the X^2 statistic, but also allows assessments of the degree of DIF to be made.
5. It requires few model assumptions

6. It performs favourably in simulations.
7. Umoinyang (2011) said it was recommended as the best optimal chi-square statistic method because it matches subjects most precisely, and provides a powerful test of significance.

Disadvantages of M-H Method

1. No covariates, other than the total score, which is the construct the item purports to measure.
2. Requires collapsing that is breaking it down into score groups.
3. More difficult to model multiple attributes.
4. It is less powerful in some studies than parametric method such as logistic regression (Rogers and Swaminathan, 1993).
5. It lacks power to detect non uniform DIF (Hambleton & Rogers, 1989; Swaminathan & Rogers, 1990; Uttaro & Millsap, 1994)
6. Mantel-Haenszel can be affected by item discrimination and it performs better with large group sizes (Roever, 2005). Camilli & Shepard (1994) claim that it needs similar number as IRT methods but Muniz et al (2001) shows that it functions well with 500 in the smaller group.
7. When test contain small number of items, observed scores may not represent true scores well and one of the assumptions made in the M-H method that observed scores are representatives of latent trait or true scores of examinees maybe violated, resulting in poor estimation of statistics (Pommerich, 1995 as cited by Teresi, 2004).

Mantel Method

The Mantel is a polytomous (ordered response) extension of the Mantel- Haenszel method (Mantel, 1963). Calculation is based on a comparison of the item means for matched groups. The null hypothesis is that at a fixed level of total score, there is no conditional association between the item score and group membership, When Mantel procedure is applied to the DIF context, it can be considered that there are in ordered levels for the item score variable: where m= 1, 2, …m. There are 2 levels for the group membership: the focal group and the reference group, There are K levels for the matching variable, where K= 1, 2… K. (Atar, 2006; and Krisjanssen et al, 2005).

It is more difficult to make selection for the matching variable than in the case of dichotomous test items. It can be done by using external measurement such as scores that are obtained based on the dichotomously scored item and polytomously scored item (Zwick et al, 1993).

Atar (2006), explained that for M response categories of the studied item and K levels of the matching variable, is three-dimensional 2xMxK table is constructed for each item. In the table of data layout for the mantel method (see appendix) Y_1, Y_2, Y_3, …, and Y_m indicates the score that is obtained for the first, the second. The third, …, and the M^{th} response category for the studied item. For the K^{th} level of the matching variable, N_{mk} is the total number of examinees who received an item score of Y_m, N_{rk} and N_{fk} is the total number of examinees in the reference group and the focal group respectively and N_k is the total number of examinees. N_{mrk} is the frequency of each item score of Y_m for the reference group, N_{mfk} is the frequency of each item score of Y_m for the focal group, at the K^{th} level of the matching variable.

He said that the null hypothesis to be tested is that the performance on the studied item of examinees in the reference group and the focal group is the same across all level of the matching variable. The Mantel chi-square test with one degree of freedom associated with the null hypothesis is

$$\text{Mantel-}X^2 = \frac{[|\Sigma_k\Sigma_m Y_m n_{mfk} - \Sigma_k E\ (\Sigma Y_m n_{mfk})|]}{\Sigma_k \text{var}\ (\Sigma_m Y_m n_{mfk})}$$

Where

$$E\ (\Sigma_m Y_m n_{mfk}) = \frac{N_{fk}\ \Sigma_m Y_m N_{mk}}{N_k}$$

$$\text{Var}\ (\Sigma_m Y_m n_{mfk}) = \frac{N_{rk} N_{fk}\ \ [N_k \Sigma_m Y_m 2_{Nmk} - (\Sigma_m Y_m N_{mk})^2]}{N_k^2\ (N_k-1)}$$

Zwick et al (1993) express the fact that when the items are scored as 0 or 1, the Mantel chi-square statistics is identical to the M-H chi-square statistics without the continuity correction.

The Mantel can detect uniform DIF well (power ranging from 0.5 to 1.00). Generally, its power for detecting uniform DIF is comparable to, or higher than, the power of most other techniques. However, due to Mantel test differences in mean item scores, it cannot detect non-uniform DIF (Krisjanssen et al, 2005).

Generalized Mantel-Haenszel Method (GMH)

The generalized Mantel-Haenszel (GMH) procedure is a generalized Mantel-Haenszel statistic for nominal response (polytomously scored items) data (Mantel-Haenszel, 1959). It is based on group differences in the entire response distribution.

Atar (2006) emphasized that the GMH statistic is viewed as the multivariate generalization of the M-H chi-square statistic. The GMH statistic with M-1 degree of freedom associated with the null hypothesis is defined using a matrix formulation,

$$\text{GMH-}X^2= [\Sigma F_k-\Sigma E(F_k)]^1\ [\Sigma V(F_k)]^{-1}\ [\Sigma F_k-\Sigma E(F_k)]$$

Where F_k is a 1x(M-1) vector of the frequencies of the reference group examinees for M-1 item score category at the Kth level of the matching variable. In this case, $F_k{}^1= [N_{1rk}, N_{2rk}, \ldots, N_{(m-1)rk]}$

$E(F_k{}^1)$ is also a 1x (M-1) vector that is formulated as

$$E(F_k{}^1) = \frac{Nrk}{Mk}\ T_k{}^1$$

Where T_{k1} is a 1x(M-1) vector of the frequencies of the examinees in both the reference group and the focal group for M-1 item score category at the Kth level of the matching variable,

$T_{k1} = [N_{1k}, N_{2k}, N_{(m-1)k}]$

Var (F_k) is a (M-1) x (M-1) covariance matrix

$$\text{Var } (F_k) = \frac{N_{rk}N_{fk}}{N_k{}^2(N_k-1)}\ [N_k \text{ diag } T_k-T_k\ T_k{}^1]$$

For the fact that GMH test differences across the entire response scale, it should be sensitive to both uniform and non-uniform DIF. However, it does not yield separate coefficients for uniform and non-uniform DIF (Krisjanssan et al, 2005).

Zwick et al (1993) reported that M-H method is a special case of the GMH procedure, when the item are scored as 0 or 1, the GMH chi-square statistic is identical to the M-H chi-square statistic without the continuity correction,

Standardization Method

Standardization method (Dorans & Kulick, 1986) is usually used to complement M-H method, while the M-H method helps to describe DIF; the standardization method is used to describe DIF. The standardization procedure has similar characteristics with the M-H method and both procedures provide similar results in a DIF (Atar, 2006). Observed total scores are used to match examinees in the reference group and in the focal group. When the examinees in the focal group and the reference group are matched with respect to their abilities that are intended to be measured by the test. The difference between the two groups for the same ability level is viewed as an "unexpected" DIF since it is expected that the two groups with the comparable ability level perform the same (Atar, 2006; Dorans & Kulick, 1986).

According to Atar (2006), the DIF measure in the standardization method is the observed proportion correct differences on an item between the focal group and the reference group at the K^{th} matching variable level, defined as

$D_k = P_{fk} – P_{rk}$

Where P_{fk} and P_{rk} are the proportion correct of the studied item for the focal group and reference group respectively at the K^{th} level of the matching variable. The standardized P-difference is defined as:

$SPD = \Sigma W_k(P_{fk}-P_{rk}) / \Sigma W_k$

Where W_k / ΣW_k is the weighting factor at the K^{th} level of the matching variable the performance differences between the focal group and the reference group (P_{fk}-P_{rk}). Wk is usually the number of examinees in the focal group at the K^{th} level of the matching variable.

When SPD DIF index is positive it shows that, the studied item favoured the focal group but when it is negative, it shows that the studied item favoured the reference group. The value of SPD DIF index is between -.05 to .05. The proportion correct difference between two levels is considered as negligible. When SPD DIF index is smaller than -.05 and greater than .05, such items are further subjected to examination.

Angeff (1995) posited that considering proportion correct differences between the focal group and the reference group for each level of the matching variable and weighing their differences are two properties of the standardization procedure that distinct the standardization procedure from the M-H procedure. In addition, null hypothesis is not available for the standardization method. Standardization method can be used to identify distractors that differentially attract examinees choice (Dorans & Holland, 2012).

Advantages of Standardization Method

1. Few model assumptions
2. Provides empirical item-scale regressions, so that non-uniform DIF is detected directly from these plots.
3. Comparing plots across score levels allows visual inspection of item by group by score level interactions.
4. It provides magnitude measures with guidelines.
5. Is not labour intensive or complex.(Teresi, 2004)

Disadvantages of Standardization Method

1. No covariates other than the total score are used.
2. Requires group variable

3. Formal tests of hypothesis of uniform DIF is not available
4. Inspection of plots is used.
5. More difficult to model multiple attributes.
6. Less effective with much skewed data.
7. Along with other observed score methods, it might not be optimal with less than 20 items.

   (Teresi, 2004)

8. It does not detect as DIF all items with larger aberrant group differences.
9. It considers as DIF items with average differences in passing the item equal to the average difference in ability between the two groups.
10. In normal circumstances, any item will be considered as DIF if the difference in passing the item at each ability level is 8% beyond the difference in ability level measured by the percentage answering all items correctly.

    (Gonzalez-Tamago, 1988).

Standardized Mean Difference Method (SMD)

The standardized Mean Difference method is a version of the standardization method, which is meant for polytomous scored items. (Dorans and Schmitt, 1993). It compares the mean item score between the reference group and the focal group, standardized as if the reference group distribution across strata were the same as the focal group distribution across the level of the matching variable (Zwick & Thayer, 1994). The DIF statistic is referred to as the standardized P-difference. The standardized mean difference is defined as

$SMD = \Sigma W_k(M_{fk} - M_{rk})/\Sigma W_k$

Where M_{fk} and M_{rk} are the mean item score at the K^{th} level of the matching variable of the focal group reference group respectively. $W_k/\Sigma W_k$ is the weighting factor at the K^{th}

level of the matching variable to weight the performance differences between the reference group and the focal group ($M_{fk} - M_{rk}$), and Wk is the number of examinees in the focal group at the K^{th} level of the matching variable (Atar, 2006).

When the mean item score of the focal group is smaller than that of the reference group, the SMD index is negative but it is positive when the focal group mean item score is greater than that of the reference group for the comparable ability levels.

Scheuneman Chi-Square Method

This is a modified chi-square method used to detect DIF (Scheunaman, 1979). According to Odili (2003), an item is defined as non-differential functioning if the probability of correct response is the same for all persons of a giving ability regardless of their group membership. Ability is measured by the total score in a homogenous test item that measures only mathematics ability. Abedalaziz (2010) explained that Scheunaman's version of the chi-square method is concerned not only with frequencies of persons in each category as the usual chi-square is but with the number of correct responses made by persons in each group (or sub population) of interest. This is evident in the degrees of freedom for this method, which is (k-1) (r-1) where k is the number of subpopulations and r is the number of score groups, or categories. Scheuneman's (1979) modified chi-square formula is

$$X^2 = \Sigma[(R_e - R_o)^2] / R_e + \Sigma[(F_e - F_o)^2]/F_e$$

Where R stands for reference group (R_e: expected frequencies, R_o: observed frequencies) and F stands for focal group (F_e: expected frequencies, F_o: observed frequencies). According to Scheuneman (1979) four or five score intervals can be created. The factors that determine number of interval are: difficulty of items, length of the test and size of the sample.

The detection of DIF by the method of Schrunaman chi-square method is done by following these steps

1. Divide examinees into two groups for comparison say focal group and reference group.
2. Use the total scores of the test that the DIF is being studied as the matching variable to group the examinees into matching levels.

3. State the null hypothesis

4. Determine the observed frequencies by counting the number of examinees that got the item correct in the reference group and the focal group respectively

5. Determine the expected frequencies by dividing the product of the total number of examinees who got the item correct and the total number of examinees in the group with the total number of examinees in both the reference and focal group.

6. The chi-square for each of the matching levels is calculated

7. The chi-square for all the matching levels are added up to get the final chi-square for the studied item

8. The degree of freedom is determine using (k-1) (r-1) where k is the number of groups and r is the number of matching levels.

9. A table value is got for the df value at .05 alpha level.

10. If X^2-calculated is equal to or greater than X^2-table value, the null hypothesis is rejected but if X^2-caculated is less than X^2-table value, the null hypothesis is accepted.

Shephard, Camilli & Averil (1981) as cited by Odili (2002) recommended that Scheuneman chi-square method is a suitable DIF detection method. This recommendation is based on its practical utility, and the fact that results obtained using the technique closely relates to those obtained with the ICC-3 parameter method. See appendix C for calculation.

Logistic Regression

Swaminathan and Rogers (1990) introduced a method of detecting DIF called Logistic regression. Zumbo(1999) explained that Logistic regression is based on the statistical modeling of the probability of responding correctly to an item by group membership (i.e. reference group and focal group) and a criterion or conditioning variable. This criterion or conditioning variable is usually the scale or subscale total score but sometimes a different measure of the same variable.

The Logistic regression method is applied on binary items but it can naturally be extended to ordinal items. In Logistic regression the following are put into consideration

1. The item response (0 or 1) which is the dependent variable.
2. The group variable (reference=1 and focal=2)
3. The total scale score for each subject tag variable TOT
4. A group by TOT interaction (i.e. the interaction of 2 and 3 above) is the independent variable

The Logistic regression equation is

$F_j = \ln [P_j /1\text{-}P_j] = \beta_o+\beta_1 X_j+\beta_2 G_j+\beta_3 (XG)_j$

Where P_j is the probability of getting item correctly for person J; $(XG)_j$ is the interaction term between the observed ability level (TOT) and the group membership (variable) for j. βo, β_1, β_2, and β_3 are the coefficients of the Logistic regression DIF model. β_o is the intercept of the model, β_1, β_2,and β_3 are the slopes of the model. The item reflects uniform DIF if β_2 is non-zero and β_3 is zero whereas the item reflects non-uniform DIF if β_3 is non-zero (Atar, 2006).

This method provides a test of hypothesis to test two hypotheses. The first hypothesis is the effect of group membership on the log odds of probability of correct response for the item is equal to zero, while the second is stated as there is no interaction between group membership and TOT. These hypotheses can be tested with likelihood-ratio test statistic that has chi-square distribution. The first hypothesis is usually used to detect non-uniform DIF. Zumbo and Thomas (1997) indicated that an examination of both the 2-df chi square test (of the likelihood ratio statistic) in Logistic regression and a measure of effect size is needed to identify DIF. For an item to be classified as displaying DIF, the two-degree-of-freedom chi-squared test in Logistic regression had to have a P-value less than or equal to 0.01 and the Zumbo-Thomas effect size measure of at least an R-squared of 0.130 (Zumbo, 1999).

The detection of DIF by the method of Logistic regression is done by following these steps:

1. First enter the conditioning variable (i.e., the total score).

2. The group variable is entered.
3. The interaction term is entered into the equation.
4. One obtains the chi-square value for step 3 and subtracts from it the chi-square value for step 1.
5. The resultant chi-square value from 4 can then be compared to 2 degrees of freedom chi-square test.
6. The 2-degree of freedom is got by finding the difference between the model chi-square statistic at step 3 (which is 3) and the model chi-square statistic at step 1 (which is 1), This two- degree of freedom is a simultaneous test of uniform and non-uniform DIF (Swaminathan and Rogers, 1990).
7. The cooperation of the R-squared values of step 2 and 1 will give you how much of the DIF is uniform while the cooperation of the R-squared values of step 3 and 2 will give you how much of the DIF is non-uniform.

   Zumbo (1999) gave an illustration of Logistic regression for binary item score test containing 20 item for 200 males and 200 females. The table on logistic regression for binary item (see appendix o) shows the result. Item 1 was without DIF whereas item 2 was with DIF. The difference in R-square from step 2 and step 3 for item 2 was quite small suggesting that DIF was predominantly uniform.

Advantages of Logistic Regression

1. There is no need to categorize a continuous criterion variable.
2. It can model uniform and/or non-uniform DIF.
3. It can generalize the binary Logistic regression model for use with ordinal item scores.
4. It is easily available on the SPSS platform.

5. Logistic regression procedure is as powerful as the M-H method for items that show uniform DIF (Swaminathan and Rogers, 1990).
6. Performs well in simulations, its detection rates is better than that of M-H and Rasch logit, in the presence of non-uniform DIF and when the reference and focal groups have unequal ability distributions.
7. Provide measurement of magnitude of DIF.
8. It is easy to perform, unless when IRT ability estimates are used (Teresi, 2004).

Disadvantages of Logistic Regression method

1. Requires more model assumptions
2. It is sensitive to misfit.
3. Item scoring may impact DIF detection.
4. Low item variability may result in false DIF detection.
5. Use of total score as conditioning variable is not optimal, but other estimate can be used (Camilli and Shepard, 1994; Crame and Colleagues, 2004 as cited by Teresi, 2004).

Ordinal Logistic Regression Procedure

This is an extension of Logistic regression procedure usually used to detect DIF in polytomous scored items (Zumbo, 1999). It follows the same process as Logistic regression method for binary-scored items.

Zumbo, 1999 gave an illustration using a 20 items likert-type; each item had a four point scale ranging from 0 to 3. There are 249 females (focal group) and 262 males (reference group)

The difference in chi-square values and degrees of freedom results in a 2-degree of freedom chi-square test

1.077 With 2 d.f; P= 0.299.

The P-value was obtained from a standard statistics textbook or a computer program like EXCEL, Minitab or statistica. Since 0.299>0.01 this item is not demonstrating DIF.

The effect size measures are as follows

For the model with only the conditioning variable (total score) at step 1

R-squared (%) is 40.82 or 0.4082

For the model with the conditioning and grouping variable at step 2

R-squared (%) is 40.82 or 0.4082

For the model with conditioning, grouping and interaction variables (the model with both uniform and non-uniform DIF) at step 3

R-squared (%) is 41.02 or 0.4102

The DIF effect size for both uniform and non-uniform DIF (step 1 Vs step 2) is R-square = 0.002 which is less than 0.130. Hence, it has a trivial effect size.

For item 2

Step 1: Model with total score X^2 (1) = 111.181, R-square d= 0.3135

Step 2: Uniform DIF X^2(2) = 159.121, R-squared = 0.5010

Step 3: Uniform and non-uniform DIF X^2 (3) = 161.194, R-squared = 0.5637

Examining the difference between steps 1 and 3 above we have

X^2 (2) = 50.013, P = 0.00001, R-squared = 0.2502 for the DIF test P =0.00001<0.01 and R-squared = 0.2502>0.130. Clearly, this item is statistically significant and shows a large DIF effect size. Moreover, comparing the R-squared values of step 2 and 3, the data suggest that item 2 shows predominantly uniform DIF.

Simultaneous Item Bias Tests Method (SIBTEST)

Simultaneous item bias test (SIBTEST) employs the non-parametric multidimensional DIF model of Shealy and Stout (1993). Which looks at the differences in probability of correct responses between focal and reference groups (beta index), after matching respondents on true ability score. Previous DIF detection procedures focus on each item separately but with SIBTEST method, multiple items can be tested to detect the amount of DIF in the entire subtest (Bolt, 2002).

To operate SIBTEST on standardized achievement test, the test items are divided into studied (suspect) subtest and the matching (valid) subtest. The studied subtest is comprised of the items believed to measure the primary and secondary dimensions based on the substantive analysis in the first stage (comprised of the items in the test that are suspected to exhibit DIF), whereas the matching subtest contains the items believed to measure only the primary dimension. The matching (valid) subtest is used as the internal matching criterion to control for the group differences in the "target ability" that is intended to be measured by the test in the detection of DIF or DTS. That is it is used to place individual in the focal and reference groups at each score level (Gierl et al, 2002; Atar, 2006). The estimate of unidirectional SIBTEST DIF index given by Atar (2006) is

$B_u = \Sigma P_{fk} (Y_{rk} – Y_{fk})$

Where

K = number of score levels on the valid subtest

B_u = maximum score level on the valid subtest

P_{fk} = proportion of the focal group examinees that obtain a valid subtest score of K

Y_{rk} = mean suspect subtest score for reference group

Y_{fk} = mean suspect score for focal group at the K^{th} valid subtest score level

The null hypothesis of no unidirectional DIF is

Ho: $\beta_u = 0$

The SIBTEST test statistic associated with the null hypothesis is

SIBTESTu = $\beta_u / \sigma\ (\beta_u)$

Where $\sigma\ (\beta_u)$ is the estimated error for unidirectional SIBTEST DIF index, β_u

Roussos and Stout (1996) classified the strength of DIF as

1) A- level DIF: the absolute value of beta index is less than 0.059
2) B- level DIF: the absolute value of beta index is between 0.059 and 0.088
3) C- Level DIF: the absolute value of beta index is equal or higher than 0.088.

The A, B and C level of DIF is also categorized as negligible, moderate, and large respectively. Sometimes group differences in the ability distribution might lead to biased estimate of the SIBTEST DIF index indicating the presence of DIF when there is no DIF. In this case, regression correction is used to compute an unbiased estimate of the SIBTEST DIF index (Atar, 2006)

Atar (2006) also reported that Narayanan and swaminathan (1996) compared the SIBTEST method with other two dichotomous DIF methods- the M-H method and the logistic regression method- to detect non-uniform DIF. They investigated the power of the three DIF methods in detecting non uniform DIF and the type 1 error rates. As a result, they found that the SIBTEST procedure was as powerful as the LR method in detecting non-uniform DIF. Type 1 error rates were within the expected normal level for the SIBTEST and the LR methods.

Advantages of SIBTEST method

1. It is non-parametric, so model fit is not an issue in DIF detection.
2. It allows modelling of multidimensional abilities.
3. Provides DIF significance test and magnitude estimates.

4. Can detect crossing DIF with crossing SIB.
5. Can measure impact by adjusting means
6. Simulations show superior performance of Poly-SIB (in comparison with IRTLR and DFIT under several IRT models) in terms of false positives when groups have different ability distributions and the correct model is not known.

   (Teresi, 2004; Bolt, 2002)

Disadvantages of SIBTEST method

1. There is no covariates
2. Usually requires a group or categorical variable
3. Used an observed “valid” score that may not be easy to construct
4. Poly SB can detect only uniform DIF
5. May not be powerful with smaller sample size

(Bolt, 2002; Teresi, 2004; Shealy and Stout, 1993)

Poly-Sibtest Method

Poly-SIBTEST method is an extension of SIBTEST to detect DIF for polytomously scored items (Chang, Mazzeo, & Roussos, 1996). It is defined as the expected group difference on the suspected item at each valid subtest score level. The Poly-SIBTEST DIF index is

$\beta = \Sigma P_k \, (Y_{rk} - Y_{fk})$

It can be interpreted in the same way as for the SIBTEST. The poly-SIBTEST test statistic associated with the null hypothesis is same as the one with SIBTEST procedure.

Poly-SIBTEST = $\beta / \sigma \;(\beta)$

Unlike the SMD method and the Mantel, the studied item is not included in the matching variable in the Poly-SIBTEST method (Chang et al, 1996). The Poly-SIBTEST procedure detects DIF for each item but does not detect DIF for each score category (Haender, 2001 as cited by Teresi, 2004)

Transformed Item Difficulty (TID) Method

Abedalaziz (2010) explained that an item is considered biased in this approach if, compared to other items on the test, it is relatively more difficult for one group than for another. The method involves computing the difficulty or P-value (proportion of subjects getting item right) for each item separately for each group. Using tables of the standardized normal distribution the normal deviate Z is obtained corresponding to the (1-P) th percentile of the distribution. A data value is calculated from the Z-value by the equation $A = 4Z - 13$.

He stated that a large delta value indicates a difficult item. For two groups, there will be a pair of delta values for each item. These pairs of delta values can then be potted on a graph, each, item represented by a point on the graph. A line can be fitted to the plot of points; and the deviation of a given point from the line is taken as measure of that item's bias. A large deviation indicates much bias (Subkoviak et al, 1987). This procedure has been used to study cultural differences in a wide variety of contexts (Angoff, 1975; Angoff & Ford, 1973; Angoff & Modu, 1993; Angoff & Sharon, 1972; Breland, Stocking, Pinchak & Abrams, 1974; Gultikson, 1964; Rudner, 1976, as cited by Abedalaziz, 2010)

The equation used for the major ellipse is $Y = AX + B$ (the best fitting line) in which : Y could for example represent male's delta values (Δ_m), X represents female's value (Δ_f), and

$\beta = \mu_x - A\mu_y$ where

A: Represents a line slope

B: The line sector of Y-axis

μ_y: The mean of delta value for female (Δ_f)

μ_y: The mean of delta values for males (Δ_m)

The perpendicular distance (D_i) that each point deviates from the major axis is calculated from this formula:

$$D_i = \frac{AX_i - Y_i + B}{A^2 + I}$$

Where X_i= Represents males delta value for item i

Yi= Represents female delta value for item i

Those items with (D_i) values in excess of ± one unit reveal DIF. The larger (D_i) is, the more biased the item. A signed transformed difficulty measure of DIF, which preserved both the direction and magnitude of DIF is obtain by attaching a positive sign to (Di) if the item reveals DIF in favour of females and a negative sign if the item reveals DIF in favour of males,

However, Roever (2005) revealed that this method is no longer in use as it confounds item difficulty and discrimination: more discriminating items look more difficult than they really are. In addition, the TID method does not match test takers by ability: only where test takers of the same ability level show different likelihood of getting the item right does the item truly function differentially. An improvement over the TID is the conditional P-value aka the Standardization method, which compares P-value for the reference and focal groups at each score level.

Point Biserial Method

Point biserial method of detecting DIF is also referred to as Item discrimination method. Traditionally, the correlation between item score and the total test score has been used in standard item analyses to identify discriminating item (Ironson and Subkoviak, 1979).

Green and Draper (1972) were one of the first people to extend it to detection of DIF in test items. Items are considered to show DIF if they are in the best discriminating half of the items for one group and in the worst half for another group.

Item Characteristic Curve Method or Area between Item Characteristic Curves

This is an item response theory (IRT) method that rest on the fact that the item characteristic curves (ICCs) of two subgroup are identical when an item does not show DIF and the area between the curve is zero i.e. the two curves are as close as possible. However, when an item shows DIF, the ICCs are not the same and the area between the curves is not zero. The principal conceptual unit of IRT is the ICC. An ICC is the function that relates the probability of a correct answer on an item to the "ability" measured by the test containing the item (Abedalaziz, 2010).

The b parameter is the item difficulty in IRT. It is determined by locating the point on the ICC that corresponds to a 50% chance of getting the item right, and the value of θ is then determine on the X-axis that corresponds to that point. This means that difficult items will have higher values of b and will be located at the right or higher end of the θ scale, while those items that are easy will have lower value of b and will be located at the left or lower end of the θ scale. The items with lower value of b require less ability to answer them correctly but those items with higher value of b require more ability to answer them correctly.

A positive value of the differences in b parameters for two groups indicates DIF favouring the reference group, while a negative value of the difference indicates DIF favouring the focal group. The simple differences in b parameters for the two groups convey the "size" rather than the statistical significance of the DIF (Camilli & Shepard, 1994).

Abedalaziz (2010) went further to explain that in a study involving identifying the difficulty parameter for males and females by BILOG-MG program. The difficulty difference was defined as follows:

$\Delta b = b_f - b_r$

Where

b_f: Estimated difficulty parameter for male (reference group)

b_r: Estimated difficulty parameter for female (focal group)

Δ_b: Estimated difficulty difference.

To test the significant of Δb, the statistic d was defined as follows:

$D = \Delta b / S_{\Delta b}$

Where

$S_{\Delta b}{}^2 = S_f{}^2 + Sr^2$

$S_{\Delta b}$: The standard error of b-difference

$S_f{}^2$: The variance of estimating b-parameter for females group.

$S_r{}^2$: The variance of estimating b-parameter for males group

With the aid of the normal probability, distribution tables the null hypothesis H_0: $\Delta_b = 0$ can be tested (Lord, 1980). DIF is said to favour the reference group if the value of difference is positive. If the value of difference is negative, then DIF favours the focal group.

Zumbo (1999) in his handbook gave three ways DIF can be assessed by comparing the ICCs of different groups on an item.

Figure 2: An example of an item that does not display DIF

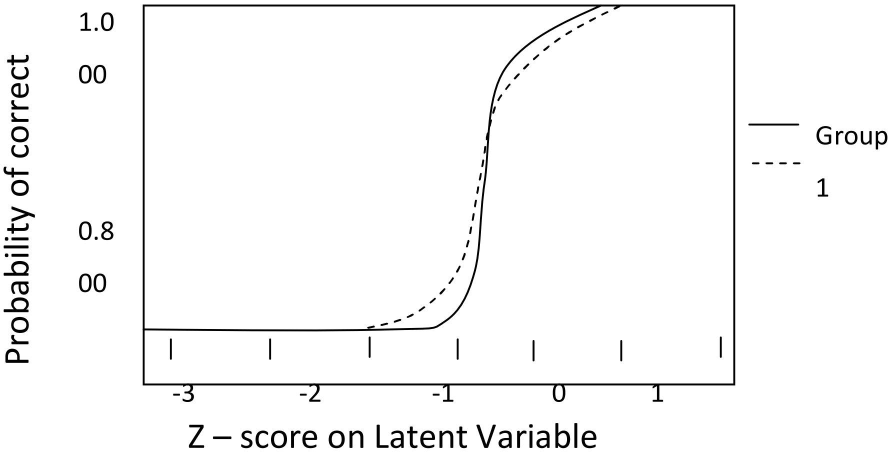


Figure 2 is an example of an item that does not display DIF because the area between the curves is very small and the parameters for each curve would be nearly equivalent.

Figure 3: An example of an item that displays substantial uniform DIF

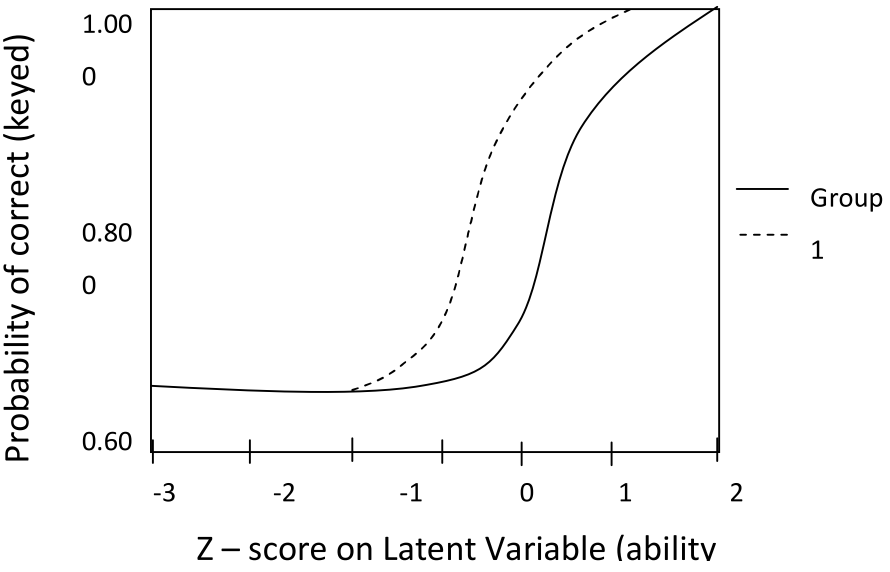


Figure 3, on the other hand, gives an example of item that displays substantial DIF with a very large area between the two ICCs. This type of DIF is known as uniform DIF because the ICCs do not cross.

Figure 4: An example of an item that displays substantial non uniform DIF

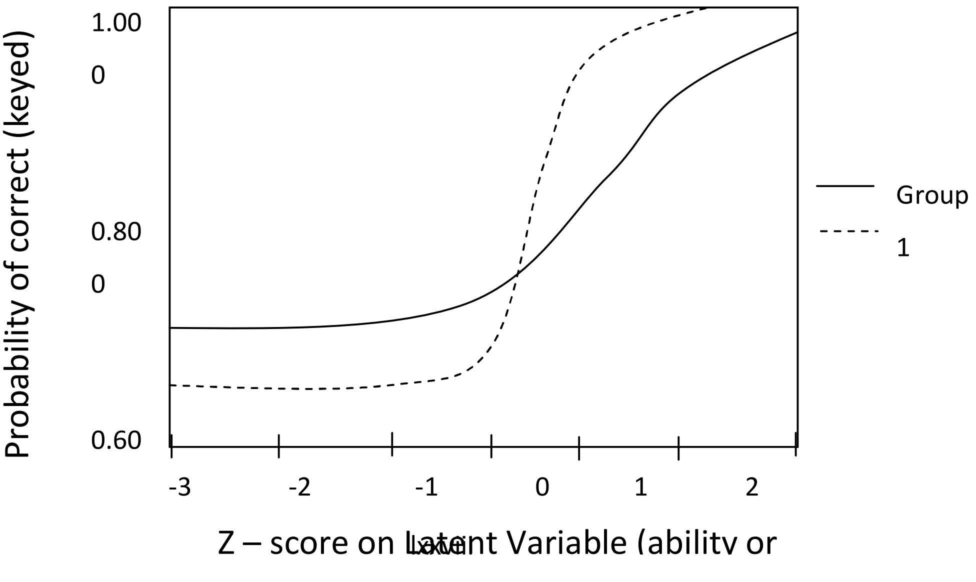


Figure 4 is an example of an item that displays substantial non uniform DIF because for those individual who score at or below the mean (i.e. $Z \leq 0$), group 1 is favoured whereas for those scoring above the mean (i.e. $Z > 0$), group 2 is favoured,

The America board of internal medicine (2012) defined the area between the ICCs as:

Area = $\Sigma\ \Delta\theta_k|P_{ref}(\theta) - P_{foc}(\theta)|$ Where $\Delta\theta_k$ is the width of a quadrature nodes

IRT-Likelihood Ratio Test Method (IRT-LR)

The IRT likelihood-ratio test procedure is one of the parametric and model-based methods used for detecting DIF (Thissen, 1991). Several models are available, like the logistic and graded response models. It can detect both uniform and non-uniform DIF, Differential item functioning is said to occur if item response functions differently between groups.

The null hypothesis to be tested is that the item parameters between the reference and the focal group do not differ. The difference in the item difficulty parameters between two groups is tested for the uniform DIF and the difference in the item discrimination parameter is tested for the non-uniform DIF. For the test of the null hypothesis of no DIF, two models are compared: a compact model and an augmented model. In the compact model, the item parameters for the common item or items across groups are constrained to be equal in the two groups. In the augmented model, the item parameters for the studied item are unconstrained and the remaining items are constrained to be equal in the two groups (Atar, 2006).

The likelihood ratio test statistic, G^2 according to Atar (2006) is computed by this equation:

$G^2 = -2LL_c - (-2LL_A)$

Where LL_c is the log likelihood for the compact model given the maximum likelihood estimates of the parameters of the compact model and LL_A is the log likelihood for the augmented model given the maximum likelihood estimates of the parameters of the augmented model. The value of G^2 is distributed as the chi-square with the degrees of freedom equal to the difference in the number of parameters in the two models. If the result of the test is found to be significant, then it is said that the study item exhibit DIF.

Advantages of IRT-likelihood Ratio Test Method

1. It has well developed theoretical models.
2. It can detect uniform and non-uniform DIF
3. No equating is required because of simultaneous estimation of group parameters.
4. It can model missing data
5. It can measure magnitude as differences in expected item scores.
6. It can measure impact of DIF on the total score using total (test) response function (TRF) which shows the relationship between expected scale scores and theta.
7. In simulation, it shows superior performance to non-parametric methods in terms of power, particularly with small samples e.g., 300 (Bolt, 2002). (Teresi, 2004)

Disadvantages of IRT-likelihood Ratio Test Method

1. Model must fit the data (Misfit can result in Type 1 error inflation false positive DIF detection)
2. Its assumptions must be met.
3. Categorical group variable as required
4. Its magnitude measures are not well integrated in DIF detection process.

Comparison Method for Item Parameter (1P, 2P, and 3P)

In this method both, the parameter difference statistics for item discrimination and difficulty are first calculated for each item from the reference and focal group. This method is based on IRT.

Item discrimination parameter difference = $a_i(R) - a_i(F)$

Item difficulty parameter difference = $b_i(R) - b_i(F)$

The value obtained from these parameters are then standardized. The square of the standardized difference value could be evaluated as X^2 statistics under 1 decree of freedom (Toit, 2003). If an item is significant at 0.01 or 0.05 alpha level of significance, the item is said to display DIF across group.

According to Thelk (2008), using the output generated by BILOG-MG, the appropriate values were input into the equation $(b_1-b_2)/\sigma bdiff$, where b_1 and b_2 are the difficulty values for groups and σbdiff is the standard error of the difference between the two b values in the numerator. The solution to this equation is distributed as a Z-score (M=0, SD=1). Based on the results of the equation above for each item, any item with an absolute value Z-score greater than 2.58 (corresponding to a two-tailed $p \leq 0.01$ or 1.96 (corresponding to a two-tailed $p \leq 0.05$), DIF exists.

The BILOG-MG statistical package can be used to detect DIF under the 1, 2, 3, parameter models. The difficulty index 'b' is use in the detection of uniform DIF while the discrimination index 'a' parameter is use in the detection of non-uniform DIF.

An example of BILOG-MG DIF analysis

*From the computer printout, pick the b-value of the focal and reference groups.

*Pick the b-difference and the standard error of the b-difference (SEb dif)

*Find the Z-score of each of the items.

Z-score =(b-dif)/(SEb-dif); from the table showing an example of BILOG-MG DIF analysis (see appendix o)

1. An item is said to reveal DIF if Z-score $\geq |1.96|$ at $p \leq 0.05$. Hence, the items 1 and 4 revealed DIF.

2. When Z-score is negative, it indicates DIF in favour of the focal group and when it is positive, it indicates DIF in favour of the reference group. This holds if in the analysis, the reference group comes first before the focal group as seen in the above example. The reverse is the case if the reference group comes first before the focal group. Hence, item 1 is in favour of the reference group while item 4 is in favour of the focal group.

3. This same procedure can be followed when non-uniform DIF is to be detected. In this case, make use of the discrimination index.

Rasch Model

The Rasch model states that for a dichotomously scored item j, with difficulty δj attempted by person I with ability βi, the probability of a correct response, $Yij = 1$, is model as

$P(Yij=1) = (\exp(\beta i - \delta j))/(1+\exp(\beta i - \delta j))$

As ability varies, the probability of a correct response to the item also varies. The probability that a person with low ability will respond correctly is correspondingly low. Symmetrically, the probability that a person with high ability will respond correctly is correspondingly high. Under the Rasch model, the discrimination parameter of the two-parameter logistic model is fixed at a value of a=1.0 for all items. There are two important parameters in the Rasch model, namely item difficulty and examinee ability. The item difficulty is calculated from the number of examinees who succeed in that item while the examinee ability is the estimate of the examinee underlying ability based on performance on a set of items. . An item is considered to flag DIF if the probability of the difference between the difficulty parameter is less than 0.05 level of significance.

The WINSTEPS statistical package can be used to carry out Rasch analysis for DIF. From the table showing an example of WINSTEPS DIF analysis (see appendix o)

Differential item functioning can be interpreted using the following methods:

1. Δ_b

2. Probability
3. 't' statistic

Using Δ_b

1. When $|\Delta_b|$ value is higher than 0.25 at α =.01 or 0.20 at α =.05, it indicates significant DIF. Hence items 1, 2 and 5 are DIF items at α =.05.
2. When Δ_b is negative, it indicates DIF in favour of the focal group and when it is positive, it indicates DIF in favour of the reference group. This holds if in the analysis, the reference group comes first before the focal group as seen in the example above. The reverse is the case if the reference group comes first before the focal group. Hence the DIF in item 1 and 5 are in favour of the reference group while the DIF in item 2 is in favour of the focal group.
3. DIF categorization in logit are as follows:

Large if $|\text{DIF}| \geq 0.65$ logits

Moderate if 0.42 logit $\leq |\text{DIF}| < 0.65$ logit

Negligible if $|\text{DIF}| < 0.42$ logit

Hence, the DIF in items 2 and 5 are large while the DIF in item 1 is negligible.

Using Probability

1. An item is said to revealed DIF if the probability of Δb is less than 0.05. Hence, item 1, 2 and 5 revealed DIF.
2. The Δ_b is used to identify the group the DIF favoured as well as the strength of the DIF.

Using't' statistic

1. An item is said to revealed DIF if the |t-value| is more than 2. Hence, item 1, 2 and 5 revealed DIF.

2. When t-value is negative, it indicates DIF in favour of the focal group and when it is positive, it indicates DIF in favour of the reference group. This holds if in the analysis, the reference group comes first before the focal group as seen in the example above. The reverse is the case if the reference group comes first before the focal group. Hence the DIF in items 1 and 5 are in favour of the reference group while the DIF in item 2 is in favour of the focal group.

3. Use the method of Δ_b to find the strength of the DIF.

Multilevel Logistic Regression Model

Swamson, Clauser, Case, Nungester and Featherman (2002) used two-level hierarchical logistic regression model in analysing differential item functioning (DIF) to explain possible causes of DIF in their studies. They used item characteristic variables at the second level of the hierarchical model to explain the effect of these variables on the differentiation of items between groups of interest and to statistically test the between-item variation in DIF index. Level-1 model can be though as the person-level

In their analysis, Swanson et al first conducted random coefficients model to estimate variances of intercept and slope coefficients, Then, they conducted intercepts and slopes-as-outcomes model to predict the variation of regression coefficients with item characteristic variables. They computed EB estimates of DIF coefficients for each item and compared EB estimates with M-H odds ratios (α indices) and log-odds ratios for each item. It was also concluded that the EB estimates of DIF coefficients were more accurate than the M-N estimates and standard logistic regression techniques.

Kamata's Multilevel Rasch Model

Atar (2006) explained that Kamata (2001) proposed an item analysis model using HGLM that is algebraically equivalent to the two-level Rasch model, which he later extended to two-level latent regression model in which he attempted to predict the variation of the intercept term with person characteristic variables. He also extended it to three-level Rasch model and latent regression model. Luppescu (2002) extended Kamata's two-level Rasch model to detect DIF. Williams (2003) extended it to polytomously scored item. She compared the performance of HGLM and GMH procedures and she found that both procedures were successful in detecting the items that exhibit DIF.

Parameter Comparisons Using T-Test on b-Values

After parameters have been equated, t-test of difference between b-parameters is employed. This is a very simple procedure which may be informative for identifying items which call for a closer look, but not too common to rely solely on this. It does not account for a- and c-parameters, which may very even for fixed b-value. It is useful in a Rasch situation, because this is the same as the multivariate test. It can be done automatically by BILOG-MG using a DIF command. DIF is only assessed in terms of item difficulty (b-parameter) i.e. uniform DIF. No consideration is given to non-uniform DIF possibilities.

(American Board of Internal medicine, 2012)

How Differential Item Functioning (DIF) Can Be Avoided

Roever (2005) reported that ETS has six guidelines:

1. Treat people with respect, avoid demeaning language, ethnocentrism, do not degrade or belittle a group.

2. Minimize the effect of construct-irrelevant knowledge or skills, be careful with charts/graphs don't use complex vocabulary where unnecessary, avoid elitism, specialized legal or business terms, regionalisms, specialized sports, tools, transportation terms. General terms are okay.

3. Avoid controversial, inflammatory, upsetting materials; entirely avoid abortion, genocide, torture, witchcraft. Use extreme caution with death, evolution, religion, violence, also be sensitive to cross-cultural issues.

4. Be careful with labels for people: Instead of 'the blind' use 'blind people', instead of 'mentally ill' use 'person with a psychological or emotional disability' instead of 'manmade' use 'synthetic' and instead of 'Black' use 'African'.

5. Avoid stereotype; do not stereotype groups of people with regard to their contribution to society, generosity, honesty, quality of cultures and so on. Mix depictions in tradition and non-tradition role should be avoided.

6. Represent diversity by showing various ethnic groups in items.

Roever (2005) further noted these facts:

1. Men tend to perform better on scientific/practical, sport-related, or military context, whereas women perform better on items dealing with human relationship/aesthetic.

2. Blacks and Latinos perform better than whites on reading passages that deal with minority concerns or contain reference to minorities.

3. Blacks perform worse than white on analogy items dealing with science but better on items dealing with human relationships, but this seems to be confounded with whether the item refers to concrete objects (easier for white) or abstract concepts (easier for black)

4. Blacks and Latinos perform worse than whites on analogy items that contain homographs.

Each of these methods of detecting DIF discussed above has its own advantages and disadvantages. However, the methods based on IRT seem to yield better results than those methods that are non-IRT based. Even with these wonderful advantages, they still have their limitations. Thus, it is important that the selection of the method to be used should reflect the unique conditions of the measurement instrument under study.

The Nature of Mathematics

According to Gittleman (1975), the word mathematics comes from the Greek word mathemata, meaning things learned or subject of instruction, around 390 BC. These subject learned were geometry, arithmetic, music and astronomy. Other practical skills were probably learned in a less formal manner. Gittleman (1975) further explained that mathematics was developed in response to need of early societies with growing numbers of people living, working and even fighting together came to the need to solve practical problems such as calculating quantity of material needed to build a store, house or the amount of food needed to provide for their army. It started from Babylonian but grew in ancient Greece. The ancient Greece mathematics was translated into Arabic and then Latin and later metamorphosed into the mathematics of Western Europe. Today it has become the mathematics of the world.

According to Hornby (2006), mathematics is science of numbers and shapes; it is the process of calculating using numbers. The branches of mathematics include arithmetic, algebra, geometry, and trigonometry. Traditionally the subject is divided into arithmetic, which studies numbers, geometry which studies structure, analysis which studies infinite processes(in particular calculus) and probability theory and statistics which study random process. Pilant (2008) defined mathematics as a way of describing relationship between numbers and other measurable quantities. Mathematics can express simple equations as well as interactions among them.

Viewed from the above perspective, mathematics is a subject that helps the individual to reason logically and sequentially when faced with everyday problems. Also the subject sharpens the intellectual ability of an individual. Just as English language forms the bedrock of liberal art, so also does the study and mastery of mathematics form the life of wire of technological development and advancement of a nation. With the coming of the computer age, mathematically literate workers are needed to handle the technological processes that go with it. Most students identify mathematics as their least favourite subject. It has become a barrier to some students' success in the school and in gaining admission into universities since it is compulsory for a student to have a credit in mathematics before he/she gain admission into any university in Nigeria. Oyedeji (1998) averred that as a school subject that is compulsory students see it as most difficult to learn. Mathematics the terror of school children and worry of teachers has shuttered the dream of many ambitious students.

One of the reasons why mathematics subject is difficult to learn is that the concepts in mathematics are abstract and difficult to understand. Jekami (1992) explained that mathematical concepts have unique characteristics of abstractness. For example, concepts of number, square or rectangle, and so on. are essentially intangible but portend significant meaning and implications when illustrated or attached to concrete objects or things. Mathematics relied greatly on deductive methods, axiomatic structure, hierarchical nature and a wide range of unfamiliar symbols. These characteristics call for scrutiny of a number of learners' issues. When mathematics knowledge does not relate directly to concrete or real objects and is filled with signs and symbols representing abstract relations, structures and patterns can lead to students giving different interpretations to concept. This can lead to differential item functioning among students.

Some problems facing the teaching and learning of mathematics as indicated by Odeyemi (1984) are as follows: students possess poor mathematics background caused from one level of education to another, students develop negative attitude towards mathematics. In addition, teachers are insufficient in quality and quantity to teach mathematics, teachers lack appropriate techniques to evaluate students' achievement and many teachers are not well motivated. Amoo (2007) also observed that many students fear the subject because some teachers handle it without considering individual difference. Furthermore Cummings et al (1993) have the view that another thing that comes up for mention is the problem of teaching aids.

Cummings et al (1993) indicated some ways we can improve the teaching and learning of mathematics: The teachers, parents, government, and the society should look for ways of sustaining the interest of students in mathematics. Students should be allowed to discover for themselves. Primary school teachers should try to understand the curriculum and help the students to understand its contents. There is the need to organize regular training course for teachers. The teachers should ask an open-ended question to engage students in thinking. Students could develop favourable attitude towards mathematics when parents, siblings, peers and other members of the community advise students on the importance of the subject rather than heart poisoning them on the issue. The students could also be given guidance on the different careers that demand the study of mathematics and the relationship between mathematics and other subjects (especially the sciences). Teachers should teach mathematics in a way to encourage the understanding of the required basic structure of mathematics. One way of improving the teaching and learning of mathematics is to make sure that in constructing mathematics item, care must be taking that the items do not show bias against a group of test takers.

Gender Issues

Gender does not mean sex (male and female) as men conceive. Rather, it is the psychological and socio-cultural dimensions of being male or female (Santrock, 2006). It is the term that is used to describe any individual due to the behaviour and character that is exhibited for the fact that the individual was born either male or female. In other words, gender is the socio-cultural interpretation of male and female based on their expected role, contributions and assigned duties (Ija, 2009). Simply put, gender refers to specific central patterns attributed to both males and females in terms of behaviour and mannerism (Okoro, 2011).

Udo, Uyoata, Inyon, & Ekanem (2011) explained that gender stereotyping is very much observed among Africans. From birth, the African child is restricted to the role expectations approved by the society. Because of this cultural practice, girls are discouraged from developing their individual potentials in various ways and disciplines. The girl-child faces a dilemma especially when she tries to venture into those areas culturally regarded as a "male reserved areas". Okoro (2011) identified some gender problems resulting from bias and prejudices on which males are favoured more than females in curriculum implementation, in particular and education in general. These problems include classroom practices, family practices, general stereotyping practices, cultural practices, textbooks illustrations and so on. He further explained that some textbooks illustrations portray boys as being critical thinkers, heroes, intelligent and the like while the girls are seen only as being soft or weak but excellent in keeping homes. These are utilized during instruction through illustrations, dramatizations, and role-plays in the class as well as during item generation. In support of this, Robert-Okah (2011) stated that gender stereotyping is equally noticed in academic writings where "man" is used to represent human beings; males are used as examples in textbooks more than females; stories of great men are told more frequently than those of greater women. In Nigerian cities, most pictures and statutes that adorn strategic places are that of men. Such men are celebrated for their valiant exploits whereas great women with equal or greater exploit are relegated to the background.

Accordingly, the works of (Dillon, 1962; Finn, 1980; Ansal, 1990 & Mboto. 2001 as cited by Mboto & Bassey, 2004) showed that males' performances are superior to their females' counterparts in the sciences, mathematics inclusive. Wozencraft (1963) found a superior school achievement in favour of girls, while Inomiesa (1989), Yilwa and Olarinoye (2004) and Alordiah (2010) found no significant difference in the performance of male and female subjects in science process skills, mathematics inclusive.

Gender differences achievement in mathematics has been found. These differences are likely to be both content and ability dependent. While males outperform females in scientific mathematical tasks, females outperform males in tasks involving verbal abilities (Abedalaziz, 2011). Men have a better spatial ability than women do which gives them advantage while solving certain kind of problems in geometry (Geary, 1996). Women score relatively higher on tests in mathematics that better match course work (Abedalaziz, 2010).

Gender related DIF is a regular issue with regard to achievement tests in mathematics because differences between females and males are often found (e.g. Bielinski & Davison, 2001; Boughlon, Gierl & Khalaq, 2000; Demars, 1998; Gamer & Engelhard,1999; Scheaneman & Grima, 1997; Willingham & Cole,1997; Abedalaziz, 2010 as citied by Abedalaziz, 2011). Uwadiae (2008) published that out of about 13.8% of the candidates who had credits and above in mathematics and English language plus three other subjects in senior secondary school examination, 7.32% were males while 6.43% were females. According to Ayodele (2011), this signifies that the males performed slightly better than the females. Viewed from the above perspectives, gender differences in mathematics is inconclusive and widely open to further investigation.

Socio-Economic Status

Socio-economic status is the way people are divided into groups in a society such that they have certain economic or/and social characteristics in common. In African countries and

Western World, socio-economic status (SES) of a family is usually linked with the family's income, parent's educational level, parent's occupation and social status (Okafor, 2007).

According to Evans (2004), lower income children have less stable families, greater exposure to environmental toxins and violence, and more limited extra-familial social support networks. There is no doubt that parents in such settings would report lower educational expectations, less monitoring of children's school work and less overall supervision of social activities compared to students from high socio-economic and intact families. Students who have a low SES earn lower test scores and are more likely to drop out of school (Eaman, 2005; Hocschild, 2003). It is believed that low SES negatively affects academic achievement because low SES prevents access to vital resources and creates additional stress at home (Eaman, 2005; Jegnes, 2002). Garson (2006) as cited by Blewins (2009) stated that socio-economic status is a determining factor on what strategies could be implemented in the curriculum to assist these particular students. It also could change the process on how these students are evaluated and assessed. The goal for all educators and in particular measurement and evaluation expert is to give each child equal opportunity to be successful in the educational process.

Studies have reportedly found that SES affects student's outcomes (Barry, 2005; Eaman, 2005; Jegnes, 2002; Hochschild. 2003; McNeal, 2001). It is generally well documented that higher family socio-economic status (SES) is related to higher educational expectations for their youths (Wentzel, 1998). Evans (2004) repeatedly discovered that low SES children are less cognitively stimulated than high SES children, because of reading less and experiencing less complex communications with parents involving more limited vocabulary. Okafor (2007) argued that while poverty and students' low SES background could be considered a concern regarding students' academic performance, they are not to be laboured because; the individual characteristics are variables that align to students' performance. There is no doubt that such conditions can impact students negatively, but the strongly determined and motivated students are likely to beat the odds of greater risk of academic failure and perform with distinction in school. Moreover, the argument should shift from closing the gap of social status of adults and focus on the integration of the SES classes into our teaching and learning process as well as putting it into consideration during evaluation.

Location (Urban/Rural)

Children attending rural schools face challenges of higher poverty than those attending urban schools. In Nigeria, the lingual Franca is English language, which in most cases is not widely spoken in rural schools. What obtains in most cases is the native language of that setting. This can greatly affect students' performance in mathematics since it is with English language mathematics is been taught and assessed in schools.

According to Odili (2003) because of an improved language, learning environment the students in urban area is likely to perform better than those in the rural area. In his study of location differential item functioning of WAEC/SSCE biology objective multiple choice questions in 1999, 2000 and 2001, it was revealed that the tests contain items with significant location DIF. In addition, Young (1998) found that "location of the school has a significant effect upon students' achievement, with students attending rural schools not performing as well as students from urban schools. Adeyemi (2011) stated that there is a significant difference between urban and rural achievement of students in public examinations.

Urban schools have main advantages namely availability of resources, library, opportunities, good environment, teachers and so on. However, one of the greatest advantages of rural schools is the tendency for smaller classes, which promise increased student-teacher interaction, allow for thorough and continuous student evaluation, and provide greater flexibility in teaching strategy.

Empirical Studies on Differential Item Functioning

Abedalaziz (2010) undertook a study on a gender-related DIF of mathematics test items. The instrument for data collection was a mathematics achievement test made up of 45 dichotomous scored items. The sample of the study is made up of 3390 students' comprising of 1600 males and 1790 females. The research made use of the Mental-Haenzel, Transformation item difficulty, and b-parameter difference procedures.

The study provides evidence that there are gender differences in performance on test items in mathematics that very according to content even when content is closely tied to curriculum. The Mental-Haenzel (MH) and b-parameter difference were agreeable in allocating nine item as revealing DIF and thirteen items as not revealing DIF. As such, the percentage of agreement between the two procedures is 55%. Transformation item difficulty (TID) and MH were agreeable in allocating seventeen items as revealing DIF, and nine items as not revealing DIF. As such, the percentage of agreement between the two procedures is 65%. TID and b-parameter difference were agreeable in allocating seven items as revealing DIF, and ten items as not revealing DIF. As such, the percentage of agreement between the two procedures is 43%. The highest agreement was between MH and TID while the lowest agreement was between TID and b-parameter differences.

This work was carried out outside Nigeria and the researcher did not make used of a nationwide instrument like WAEC. There is need to carry out a study like this to see whether the result will deviate from the one above using a nationwide examination instrument like WAEC/SSCE.

Abedalaziz (2012) carried out a study titled “Exploring DIF: comparison of CTT and IRT method. The instrument for data collection was a mathematics proficiency test contains 60 dichotomous scored items. The sample of the study was made up of 1280 students (656 males and 624 female). The DIF detection method used are Area index (IRT based), transformed item difficulty (CTT based), b-parameter difference (IRT based) and Scheuneman’s chi-square (CTT based).

TID shows that 35% of the items revealed DIF, b-parameter difference shows that 75% of the items revealed DIF, Area index shows that 77% of the items revealed DIF and Scheuneman’s chi-square shows that 50% of the items revealed DIF. The methods of area index and chi-square methods were agreeable in allocating 23 items as revealing DIF, and seven items as not revealing DIF. As such, the percentage of agreement between them is 56%. The b-difference and chi-square (Scheuneman) were agreeable in allocating 25 items as revealing DIF, and 21 items as not revealing DIF. As such, the percentage of agreement between them is 85%. The TID and chi-square were agreeable in allocating 16 items as revealing DIF, and 23 items as not revealing DIF. As such the percentage of agreement between them is 72%. The b-difference and Area index methods were agreeable in allocating 27 items as revealing DIF, and 5 items as not revealing DIF, As such, the percentage of agreement between them is 59%. Area index and TID methods were agreeable in allocating 16 items as revealing DIF, and 6 items as not revealing DIF. As such, the percentage agreement between them is 41%. TID and b-difference methods were agreeable in allocating 17 items as revealing DIF, and 20 items as not revealing DIF, As such; the percentage of agreement between them is 69%. The study pointed out that the highest agreement was between chi-square and b-parameter difference (85%) whereas the lowest agreement was between Area index and TID (41%). Females showed a statistically significant and consistent advantage over males on items involving relations and functions, polynomial, trigonometric functions, whereas men showed a less consistent advantage on items involving triangles, however, it was concluded that gender difference in mathematics may well be linked to content.

Again, this work was not carried out in Nigeria and the test used was not a nationwide test like WAEC/SSCE. Also, in this work socio-economic status and location was not considered. Therefore, there is need to carry out a test like this in Nigeria but using a nationwide test like WAEC/SSCE and putting in to consideration socio-economic status and location.

Ironson & Subkoviak (1979) carried out a study on a comparison of several methods of assessing item bias. The instrument of data collection is the National Longitudinal study (NLS) of 1972 in USA. It contains 150 dichotomous scored items. The sample of the study was made up of 3485 12th grade students (1691 blacks and 1794 whites). The DIF detection method used are TID, discrimination differences (Point biserial), chi-square and ICC.

The result of this study shows that for the 150 items analyzed, three of the methods (TID, chi-square, and ICC approaches) were moderately correlated. However, there was little agreement between the discrimination difference approaches (point-biserial) with others. The largest correlation was between chi-square and TID (0.370) and then between TID and ICC

(0.234). The discrimination differences approach did not correlate significantly with any other method.

In addition, this work was done outside Nigeria and it made used of black/white as its focal/reference group respectively, which is not applicable to Nigeria society. Hence there is need to carry out a similar study that would be of relevance to the Nigerian society.

Odili (2003) undertook a study on the effect of language manipulation on DIF of Biology multiple-choice test test. The instruments for data collection were four namely: WAEC/SSCE biology paper 2 1999, 2000, and 2001 made up of 60 items each. Differential functioning test items used in the original language (form A) made up of 30 items; Differential test items with simplified non-technical words used (form B) made up of 30 items; Questionnaire on student's background (SES) made up of 20 items. The sample of the study was made up of 3300 senior secondary three students (male 1762, female 1538; urban 1980, rural 1320; high SES 638, low SES 2662; experimental group 512, control group 513). The DIF detection method used was the scheuneman's modified chi-square. However, he used the dependent t-test and chi-square to test the significant difference existing between the two groups in the experimental study.

The result revealed that WAEC/SSCE biology paper 2 for 1990, 2000, and 2001 contains item with significant location, gender, and socio-economic status DIF, with location having more DIF items. In addition, the manipulation of differential functioning test questions did significantly reduce DIF for the test takers.

The researcher used scheuneman's modified chi-square, he did not make use of the IRT based DIF method or the purely CTT based DIF method. This could be so because it was not the focus of his study. He however, used gender, SES, and location in his study. Therefore, there is need to carry out a similar study but using several DIF detection methods and WAEC/SSCE mathematics objective questions.

Appraisal of Literature

The literature described DIF as a significant difference between the p-value of two groups matched for ability under classical test theory. DIF is described as the tendency of test takers of the same standing in the latent trait to have different probability of getting an item right under the Item response theory.

Methods of detecting DIF under CTT and their characteristics were discussed. Also discussed were the methods of detecting DIF under IRT. In addition, the problems inherent in the methods based on IRT and those based on CTT were also looked into. Other aspects covered by the literature review are the nature of mathematics, gender issues, socio-economic status, and location (rural/urban).

Finally, studies that compared methods were also looked at. What the researchers did and what they failed to do were also carefully examined. These are what formed the basis of the present study.

There are several methods of DIF detection. Some of these methods are based on IRT while others are based on CTT. Due to the inherent characteristics, advantages and limitation of these IRT and CTT based DIF detection methods, the disparity that exists between CTT and IRT as well as whether the methods under CTT and IRT will detect the same items as DIF items; the need to compare these methods becomes relevant. There seems to be a dire shortage of information on this area in the country but there are some works done outside the country in these works, their focus was mainly on race and gender, race is not relevant to our country, a knowledge gap that this study made effort to fill. Because of this scarcity of information in the course of gathering material for this work, it becomes imperative to carry out a comparison study of the index of DIF under the detection methods of CTT and IRT in Nigeria, and to extend it not only to gender but also to location and socio-economic status.

CHAPTER THREE

RESEARCH METHOD AND PROCEDURE

This chapter describes in detail the procedures that were used in the conduct of this study. It was organized into the following sections.

1. Design of the study
2. Population for the study
3. Sample and sampling techniques
4. Research Instrument
5. Validation of the instrument
6. Reliability of the instrument
7. Method of data collection
8. Method of data analysis

Design of the Study

The research design for this study is the Ex-post-facto design. It was used to collect data that will enable the researcher to determine the index of DIF in West African Senior School Certificate Examination (WASSCE) mathematics multiple-choice test as well as compare the index of DIF of those methods based on CTT with those methods based on IRT.

The primary variable of interest in this study is the comparison of index of DIF under the methods of IRT and CTT. The index of DIF is the dependent variable. The methods of IRT (Rasch and IRT-3P) and that of CTT (TID and M-H) are the independent variables. The secondary independent variables include gender (male/female), location (rural/urban) and socio-economic status (low SES/high SES). These variables were chosen because they are capable of becoming sources of systematic error. The DIF index of the four methods of detecting DIF for gender, SES and location was compared.

Population for the Study

The population of this study consists of all senior secondary class III students in public schools in Delta and Edo states. The students in SS3 classes were used because they run the same academic calendar. They were expected to be at the same level in coverage of WAEC senior secondary school class 3 mathematics syllabus upon which WASSCE is based. According to the statistics from Delta and Edo states ministries of education there are 723 secondary schools in the two states. Delta has 449 secondary schools while Edo has 274 secondary schools. The population of this study is made up of 65,961 senior secondary III students in both Edo and Delta states, of this number, 39,958 or 60.58% are in schools located in Delta state, while 26,003 or 39.42% are in schools located in Edo state.

The distribution of public secondary schools and population of SS3 students in Urban and Rural areas in Delta state and Edo states in 2012/2013 session is shown in appendix E. Also, from appendix E, 41509 or 62.93% of the SS3 students are in schools located in the urban areas, while 24,452 or 37.07% are in schools located in the rural areas. The distribution of the population of SS3 students according to gender is shown in appendix F. In addition, from appendix F 34,188 or 51.84% of the students in SS3 are male while 31,773 or 48.16% are female.

Sample and Sampling Techniques

The sample for the survey study was 1900 students or 2.88% of the population. This sample size is three times greater than the minimum sample size requirement based on Taro Yemen's formula (Ukwuije, 2003). Please see appendix M. To ensure adequate representation of the individual in the variable under investigation, the proportional stratified random sampling approach was used. The schools were first stratified according to the two states. A proportion of the sample size was taken from the states to reflect the number of students in the states. This is shown in appendix G. From Delta state 1152 students were sampled while 748 of the students sampled were located in Edo State.

At the second level, the schools in each state were stratified into urban and rural. A rural school as defined and as used in this study is one located in a community where the mother tongue (native language) is the medium of communication, and there is no government office (except primary and secondary schools) or private company that employs educated elite. A proportion of the sample size was taken from each state to reflect the number of students in the rural/urban locations. This is shown in appendix H.

In Delta State, 693 of the sampled students came from urban location while 459 came from rural location. In Edo state, 451 of the sampled students came from urban location while 297 came from rural location. Schools were randomly sampled. Every student in SS3 in the sampled school was used. This method was adopted to avoid the problem of keeping some students out of the test room. The list of the schools used in the study as well as the demographic characteristics of the respondents in this study are shown in appendix L

Research Instrument

Two instruments were used for data collection. They are the WASSCE mathematics mulpiple-choice test for 2012 examination (Appendix I), used by permission (Appendix E). It contains 50 multiple choice type questions, which cover the WASSCE mathematics syllabus of senior secondary schools in Nigeria. The second instrument is a socio-economic status questionnaire constructed by Adelusi (1982) and used in her investigation of factor of achievement in English language. The instrument has 20 items on a 3-point scale, most favourable (3), favourable (2) and least favourable (1). One of the items in the socio-economic (SES) questionnaire is item no. 6

At home, you speak English

(3) all the time (2) sometime (1)rarely/never

Odili (2003) reversed some of the items to check response set like item no. 3 and modified some to suit his study, e.g. item no. 13. At home you speak Yoruba or any other

Nigerian language was modified as: At home you speak your native language. This modification was done because the population in Odili (2003) study was not predominantly Yorubas. He also modifies item no. 20 on income of father or guardian to reflect the present minimum wage earning at the time of his study.

This researcher went further to modify items no. 20 originally in Odili version. It was modified because most students may not know the monthly wages of their parents.

Your father/guardian's pay your school fees and buy your school books promptly.

□ Every time □ Sometime □ Rarely/never

See appendix A for the questionnaire on SES.

Validity of Instrument

In order to ensure the validity of the 2012 WASSCE mathematics multiple-choice, the test items were examined to see whether they cover the mathematics syllabus for SS3 and it was found to be so.

Content validity for the socio-economic status questionnaire was established by making sure that the instrument contains items that measured the yardstick for classification of individuals into high and low socio-economic status. According to Okafor (2007), in African countries and western world, socio-economic status of a family is usually linked with the family's income, parent's educational level, parent's occupation, and social status. According to Odili (2003) other yardsticks relevant to SES are availability of electronic facilities at home, and ability of parents to guide the educational development of their children. However to ensure further content validity and face validity the SES questionnaire was given to my supervisors, measurement experts and experienced educationists who are both academically and professionally qualified. From their comments and recommendations, some of the items were modified and re-worded.

Reliability of the instrument

In order to ensure reliability of the 2012 WASSCE multiple-choice test questions, the test-retest method of establishing reliability was used. The 2012 WASSCE mathematics multiple-choice test was administered to the same group of person at the interval of two weeks. The responses of the students on the two occasions were correlated using Pearson product moment correlation. It yielded a value of .89 as a measure of stability of 2012 WASSCE mathematics multiple-choice test question as shown in appendix G. This will increase dependability of findings in the present study.

The reliability for the SES scale was also established using test-retest. The measure of stability was .70 as shown in appendix G. The two instruments were administered the same day to the same sample of students.

Method of Data collection

The socio-economic status questionnaire was used to collect the students' biographical data on sex, state, and location. The SES questionnaire has a 3 point scale-three indicates high socio economic status, two indicates middle socio economic status and one indicates low socio economic status. The maximum score of the SES questionnaire is 60 while the minimum score is 20 if all items are responded to. For the purpose of this study the students with 40 and below will be grouped as low SES while those with 41 and above will be grouped as high SES since the mid-point based on 3 point scale for 20 items is 40. The mathematics teachers in the schools were used to assist in the administration of the socio-economic status questionnaire (SESQ) and the 2012 WASSCE mathematics test was administered together. To ensure that a good testing environment was realized the mathematics teachers informed the students that the exercise would be part of their continuous assessment. The test was administered within the time limit specified by WAEC.

Method of Data Analysis

. The data were analysed using the BILOG-MG, WINSTEPS 3.75, SPSS 17, and Microsoft excel statistical packages. First, a preliminary observation was done to verify the two major assumptions that must be verified, they are the unidimensionality and model fit. The WINSTEPS statistical packaged was used to establish the model fit for Rasch model while the BILOG-MG was used to establish the model fit for IRT-3P model. The confirmatory factor analysis using the SPSS package was done to confirm the undimentionality of the 2012 WASSCE mathematics multiple-choice test, see appendix N.

The BILOG-MG, Microsoft excel and WINSTEPS were used to answer research questions 1-3 while SPSS 17 was used to answer research questions 4-12 and to test the nine hypotheses. Specifically, the BILOG-MG was used to detect DIF items based on the method of IRT-3P model. The WINSTEPS 3.75 was used to detect DIF items based on the methods of Rasch model and Mantel-Haenszel. The Microsoft excel was used to detect DIF items based on the method of transformed item difficulty. The criterion for decision rule under the Rasch model was that an item is said to reveal DIF if the probability is less than 0.05. The criterion for decision rule under the IRT-3P model was $|1.96|$ at $p \leq .05$, which implied that the Z-score of test items with $|1.96|$ and above are considered as differentially functional. The criterion for decision rule under the Mantel-Haenszel method was that an item is said to reveal DIF if the probability of the M-H chi-square is less than 0.05. The criterion for decision rule under the TID method was that an item is said to reveal DIF if $|\text{Di}|$ is more than one unit. Where Di is the perpendicular distance that each points deviates from the major axis.

Descriptive statistics of frequency count and percentage were used to answer research questions four to twelve. The chi-square test of independence was used to test the nine hypotheses at $\alpha = .05$. In order to determine the degree of agreement between the DIF detection methods, the contingency coefficient was used for the nine hypotheses. The maximum contingency coefficient value for 2x2 table is 0.707. Contingency coefficient value between 0.60-0.707 was considered to be high, 0.30-0.59 was moderate and 0.00-0.29 was low.

CHAPTER FOUR

PRESENTATION OF RESULT AND DISCUSSION

This chapter presents the summary of the analysis of data in this study. The results are presented according to the research questions and hypotheses. The Rasch model, Item Response Theory-3Parameter model, Mantal-Haenszel and transformed item difficulty were used to identify items that flagged differential item functioning (DIF). Percentage was used to identify the level of agreement that existed between DIF methods. The independent chi-square test was used to test the hypotheses at α=0.05 to further analysis if the agreements between DIF methods were significant. The Pearson's coefficient contingency was used to determine the degree of agreement between the DIF detection methods. The outputs of data analysis are presented in tables. The results and interpretations are presented under each table.

Index of Differential Item Functioning under the methods of CTT and IRT

To determine the index of DIF for gender, SES and location under the methods of CTT and IRT for items in 2012 WASSCE mathematics multiple-choice test; the data collected from 1900 students were analysed using BILOG-MG, WINSTEP 3.75 and SPSS 17 packages as explained in chapter three. Consequently, the following research questions were answered.

Research Question 1: What is the index of DIF for gender under the methods of Item Response Theory (Rasch model and Item Response Theory-3 Parameter model (IRT-3P)) and Classical Test Theory(Transformed item difficulty(TID) and Mantel-Haenszel (M-H)) for each item in 2012 WASSCE mathematics multiple-choice test?

Table 1: Rasch Model and IRT-3P Model Statistics for Gender

	IRT-3P MODEL (GENDER)					RASCH MODEL (GENDER)
ITEMS	b-	b-	SEb	Z-score	DIF	

	value		dif	dif		INDEX					
	M	F					bm	bf	Δb	Prob	DIF INDEX
1	-0.47	-0.73	0.26	0.13	2.00*	2	-1.35	-1.57	0.22*	0.03	2
2	-0.39	-0.31	0.18	0.13	1.38	1	-1.29	-1.25	-0.04	0.69	1
3	0.78	1.05	-0.27	0.14	-1.93	1	-0.38	-0.19	-0.19	0.06	1
4	0.28	0.61	-0.33	0.14	-2.36*	2	0.77	-0.53	-.24*	0.02	2
5	0.43	0.46	-0.03	0.13	-0.23	1	-0.65	-0.65	0	1	1
6	0.71	0.99	-0.28	0.14	-2.00*	2	-0.44	-0.23	-0.21*	0.04	2
7	1.17	0.86	0.31	0.14	2.21*	2	-0.08	-0.34	.26*	0.01	2
8	1.1	1.27	-0.17	0.15	-1.13	1	-0.13	-0.02	-0.11	0.29	1
9	1.02	1.3	-0.28	0.15	-1.87	1	-0.2	0	-0.2	0.07	1
10	0.04	0.46	-0.42	0.14	-3.00*	2	-0.95	-0.65	-31*	0	2
11	1.16	1.35	-0.19	0.15	-1.27	1	-0.09	0.05	-0.13	0.22	1
12	0.49	0.78	-0.29	0.13	-2.23*	2	-0.6	-0.4	-.20*	0.04	2
13	2.07	2.34	-0.27	0.17	-1.59	1	0.61	0.81	-0.2	0.12	1
14	0.93	1.25	-0.32	0.15	-2.13*	2	-0.26	-0.04	-0.22*	0.03	2
15	0.28	0.25	0.03	0.14	0.21	1	-0.77	-0.81	0.04	0.67	1
16	0.94	1	-0.79	0.14	-5.64*	2	-0.24	-0.24	0	1	1
17	1.84	1.73	0.11	0.14	0.79	1	0.44	0.34	0.1	0.39	1
18	1.3	1.51	-0.21	0.15	-1.4	1	0.02	0.17	-0.15	0.18	1
19	1.38	1.45	-0.07	0.15	-0.47	1	0.1	0.12	-0.02	0.85	1
20	0.94	0.87	-0.07	0.14	-0.5	1	-0.26	-0.33	0.07	0.51	1
21	1.22	1.24	-0.02	0.15	-0.13	1	-0.04	-0.04	0	1	1
22	0.89	0.69	0.2	0.13	1.54	1	-0.3	-0.47	0.17	0.11	1
23	1.33	1.47	-0.14	0.15	-0.93	1	0.04	0.14	-0.1	0.4	1
24	1.63	1.49	0.14	0.14	1	1	0.28	0.15	0.13	0.26	1
25	0.47	0.6	-0.13	0.14	-0.93	1	-0.62	-0.54	-0.08	0.43	1
26	1.15	1.37	-0.22	0.14	-1.57	1	-0.09	0.06	-0.15	0.16	1
27	0.63	0.83	-0.2	0.14	-1.43	1	-0.5	-0.36	-0.14	0.18	1
28	0.86	0.97	-0.11	0.14	-0.79	1	-0.32	-0.25	-0.07	0.51	1
29	1.65	2.02	-0.37	0.16	-2.64*	2	0.29	0.56	-.27*	0.02	2
30	1.48	1.25	0.23	0.14	1.64	1	0.16	-0.04	0.2	0.07	1
31	1.61	1.23	0.38	0.14	2.71*	2	0.26	-0.05	.31*	0	2
32	1.61	1.76	-0.15	0.16	-0.94	1	0.26	0.36	-0.1	0.38	1
33	1.46	1.49	-0.03	0.15	-0.2	1	0.15	0.15	0	1	1
34	2.48	2.47	0.01	0.19	0.05	1	0.92	0.92	0	1	1
35	1.83	1.3	0.53	0.15	3.53*	2	0.43	0	.43*	0	2
36	1.33	1.18	0.15	0.14	1.07	1	0.04	-0.09	0.13	0.23	1
37	1.35	1.35	0	0.15	0	1	0.05	0.05	0	1	1
38	1.89	2.03	-0.14	0.17	-0.82	1	0.47	0.57	-0.1	0.43	1
39	0.99	1.38	-0.38	0.15	-2.60*	2	-0.22	0.07	-.29*	0.01	2
40	1.01	1.06	-0.05	0.14	-0.36	1	-0.19	-0.19	0	1	1
41	2.31	1.86	0.45	0.14	3.21*	2	0.8	0.44	.36*	0.01	2
42	1.53	1.23	-0.3	0.15	-2.00*	2	0.2	-0.05	.25*	0.02	2
43	0.6	0.58	0.02	0.14	0.01	1	-0.53	-0.53	0	1	1
44	2.31	2.08	0.23	0.17	1.35	1	0.8	0.61	0.19	0.13	1
45	2.01	1.81	0.2	0.16	1.25	1	0.56	0.4	0.16	0.16	1
46	3.52	2.81	0.71	0.17	4.18*	2	1.74	1.19	.55*	0	2
47	3.01	2.74	0.27	0.18	1.5	1	1.34	1.12	0.22	0.13	1
48	1.54	1.34	0.2	0.14	1.43	1	0.21	0.04	0.17	0.13	1
49	3.38	2.98	0.4	0.19	2.11*	2	1.63	1.31	.32*	0.04	2
50	1.14	1.11	0.03	0.14	0.21	1	-0.12	-0.12	0	1	1

b-value=difficulty value, SEb dif=standard error of b-difference, bm=measure for male, bf=measure for female, Δb=DIF contrast.

Table 1 shows the DIF statistics of the Rasch model method for each of the 50 items for gender. An item is said to revealed DIF if the probability is less than 0.05. The Rasch model method flagged 15 items at the 0.05 level of significance. That is 30% of the

2012.WASSCE mathematics multiple-choice test items functioned differentially for male and female examinees. The DIF items are 1, 4, 6, 7, 10, 12, 14, 29, 31, 35, 39, 41, 42, 46 and 49

Table 1 shows the DIF statistics of the IRT-3P method for each of the 50 items. An item is said to reveal DIF if Z-score≥│1.96│ at p≤0.05. The IRT-3P method flagged 16 items. That is 32% of the 2012 WASSCE mathematics multiple-choice test functioned differentially for male and female examinees. The DIF items are 1, 4, 6, 7, 10, 12, 14, 16, 29, 31, 35, 39, 41, 42, 46, and 49

Table 2: Transformed Item Difficulty and Mantel-Haenszel Statistics for Gender

TRANSFORMED ITEM DIFFICULTY (GENDER)				MANTEL-HAENSZEL (GENDER)
P- Value	Z- Value	Delta		
			CV	

ITEMS	M	F	M	F	M	F	Di	DIF INDEX	χ^2	PROB	ODDS RATIO (IN LOGIT)	DIF INDEX
1	0.61	0.62	0.28	0.31	14.12	14.24	0.09	1	3.39	0.07	-0.19	1
2	0.6	0.55	0.26	0.13	14.04	13.52	-1.26*	2	0	0.97	-0.01	1
3	0.41	0.32	-0.22	-0.47	12.12	11.12	-1.09*	2	1.85	0.17	0.15	1
4	0.49	0.39	-0.02	-0.28	12.92	11.88	-1.08*	2	3.06	0.08	0.19	1
5	0.47	0.42	-0.07	-0.2	12.72	12.2	-1.01*	2	0.09	0.76	-0.04	1
6	0.43	0.33	-0.17	-0.44	12.32	11.24	-1.07*	2	3.15	0.08	0.2	1
7	0.36	0.35	-0.35	-0.36	11.6	11.56	-0.61	1	7.36*	0.01	-0.3	2
8	0.37	0.29	-0.33	-0.55	11.68	10.8	-1.12*	2	0.36	0.55	0.08	1
9	0.38	0.29	-0.3	-0.55	11.8	10.8	-1.13*	2	0.98	0.32	0.12	1
10	0.53	0.42	-0.08	-0.2	12.68	12.2	-1.09*	2	7.59*	0.01	0.29	2
11	0.36	0.28	-0.35	-0.58	11.6	10.68	-1.05*	2	0.01	0.91	0.02	1
12	0.46	0.36	-0.1	-0.35	12.6	11.6	-1.13*	2	4.96*	0.03	0.23	2
13	0.24	0.17	-0.7	-0.95	10.2	9.2	-1.14*	2	1.5	0.22	0.17	1
14	0.39	0.29	-0.28	-0.56	11.88	10.76	-1.20*	2	1.48	0.22	0.14	1
15	0.49	0.45	-0.02	-0.12	12.92	12.52	-0.76	1	0.45	0.5	-0.08	1
16	0.39	0.33	-0.28	-0.44	11.88	11.24	-1.22*	2	0.73	0.39	0.1	1
17	0.27	0.23	-0.61	-0.73	10.56	10.08	-0.7	1	0.84	0.36	0.11	1
18	0.34	0.26	-0.41	-0.64	11.36	10.44	-1.02*	2	1.26	0.26	0.13	1
19	0.33	0.27	-0.44	-0.61	11.24	10.56	-1.02*	2	0.06	0.77	-0.04	1
20	0.39	0.35	-0.28	-0.36	11.88	11.56	-0.44	1	0.09	0.77	-0.04	1
21	0.35	0.3	-0.36	-0.52	11.56	10.92	-1.03*	2	1.34	0.25	-0.15	1
22	0.4	0.38	-0.25	-0.3	12	11.8	-0.43	1	0.65	0.42	-0.09	1
23	0.34	0.27	-0.41	-0.61	11.36	10.56	-1.02*	2	0.28	0.59	0.07	1
24	0.3	0.26	-0.52	-0.64	10.92	10.44	-0.58	1	0.16	0.69	-0.05	1
25	0.46	0.39	-0.58	-0.28	10.68	11.88	-1.10*	2	0.65	0.42	0.09	1
26	0.36	0.28	-0.35	-0.58	11.6	10.68	-1.13*	2	2.14	0.14	0.17	1
27	0.44	0.36	-0.15	-0.35	12.4	11.6	-1.15*	2	1.94	0.16	0.15	1
28	0.4	0.34	-0.25	-0.41	12	11.36	-1.01*	2	0.11	0.73	0.04	1
29	0.29	0.2	-0.55	-0.84	10.8	9.64	-1.21*	2	3.42	0.06	0.24	1
30	0.32	0.3	-0.47	-0.52	11.12	10.92	-0.32	1	2.21	0.14	-0.17	1
31	0.3	0.3	-0.52	-0.52	10.92	10.92	0.45	1	5.66*	0.02	-0.27	2
32	0.3	0.23	-0.52	-0.73	10.92	10.08	-1.01*	2	0.01	0.96	0	1
33	0.32	0.26	-0.47	-0.64	11.12	10.44	-1.07*	2	0	0.97	-0.01	1
34	0.2	0.16	-0.84	-0.99	9.64	9.04	-0.63	1	2.91	0.09	-0.28	1
35	0.27	0.29	-0.61	-0.55	10.56	10.8	0.12	1	12.71*	0	-0.42	2
36	0.34	0.3	-0.41	-0.52	11.36	10.92	-0.71	1	0.38	0.54	-0.07	1
37	0.33	0.28	-0.44	-0.58	11.24	10.68	-1.02*	2	0.37	0.54	-0.08	1
38	0.26	0.2	-0.64	-0.84	10.44	9.64	-1.10*	2	0.19	0.66	0.07	1
39	0.38	0.28	-0.3	-0.58	11.8	10.68	-0.32	1	4.26*	0.04	0.24	2
40	0.38	0.32	-0.3	-0.47	11.8	11.12	1.04*	2	0.12	0.73	0.04	1
41	0.22	0.22	-0.77	-0.77	9.92	9.92	0.45	1	0.01	0.98	0.01	1
42	0.31	0.3	-0.49	-0.52	11.04	10.92	0.32	1	5.96*	0.01	-0.29	2
43	0.44	0.4	-0.15	-0.25	12.4	12	0.65	1	0.56	0.46	-0.09	1
44	0.25	0.19	-0.61	-0.87	10.56	9.52	0.75	1	3.91*	0.04	-0.28	2
45	0.25	0.22	-0.61	-0.77	10.56	9.92	0.86	1	1.36	0.24	-0.15	1
46	0.12	0.13	-1.17	-1.12	8.32	8.52	0.51	1	0.18	0.67	-0.07	1
47	0.15	0.13	-1.03	-1.12	8.88	8.52	0.69	1	0.01	0.93	0.02	1
48	0.31	0.28	-0.49	-0.58	11.04	10.68	0.72	1	0.95	0.5	-0.08	1
49	0.13	0.12	-1.12	-1.17	8.52	8.32	0.68	1	0.36	0.55	-0.1	1
50	0.36	0.32	-0.35	-0.47	11.6	11.12	0.67	1	0.01	0.99	-0.01	1

Di=perpendicular distance that each points deviates from the major axis.

Table 2 shows the DIF statistics of the TID method for each of the 50 items. An item is said to flag DIF if │Di│ values is in excess of one standard deviation. The TID method flagged 27 items at the 0.05 level of significance. That is 50% of the 2012 mathematics multiple-choice test functioned differentially for male and female examinees The DIF items are 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 16, 18, 19, 21, 23, 25, 26, 27, 28, 29, 32, 33, 37, 38, and 40.

Table 2 shows the DIF statistics of the Mantel-Haenszel method for each of the 50 items. An item is said to flag DIF if the probability is less than 0.05. The M-H method flagged 8 items at the 0.05 level of significance. That is 16% of the 2012 WASSCE mathematics multiple-choice test items functioned differentially for male and female examinees The DIF items are 7, 10, 12, 31, 35, 38, 42 and 44.

Research Question 2: What is the index of DIF for socio-economic status (SES) under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for each item in 2012 WASSCE mathematics multiple-choice test?

Table 3: IRT-3P and Rasch Model Statistics for Socio-Economic Status

IRT-3P				DIF INDEX	RASCH MODEL	DIF INDEX
ITEMS	b-	b	Z-			

	value		dif	SEbd	score						
	H	L					bh	bl	Δb	Prob	
1	-0.62	-0.49	-0.13	0.13	1	1	-1.51	-1.42	-0.09	0.41	1
2	-0.28	-0.32	0.04	0.13	0.31	1	-1.24	-1.29	0.05	0.61	1
3	0.73	1.12	-0.39	0.14	-2.79*	2	-0.44	-0.16	-.28*	0.01	2
4	0.45	0.48	-0.03	0.14	-0.21	1	-0.66	-0.66	0	1	1
5	0.28	0.63	-0.35	0.13	-2.69*	2	-0.79	-0.54	-.25*	0.01	2
6	0.93	0.82	0.11	0.14	0.79	1	-0.29	-0.38	0.09	0.36	1
7	1.08	1.01	0.07	0.14	-0.5	1	-0.17	-0.23	0.06	0.52	1
8	1.14	1.26	-0.12	0.14	-0.86	1	-0.12	-0.04	-0.08	0.45	1
9	0.98	1.34	-0.36	0.15	-2.40*	2	-0.25	0.02	-.27*	0.01	2
10	0.23	0.32	-0.09	0.14	-0.64	1	-0.83	-0.78	-0.05	0.63	1
11	1.03	1.48	-0.45	0.15	-3.00*	2	-0.2	0.13	-.33*	0	2
12	0.7	0.63	0.07	0.13	0.54	1	-0.47	-0.54	0.07	0.49	1
13	2.19	2.2	-0.01	0.17	-0.06	1	0.7	0.7	0	1	1
14	1.06	1.34	-0.28	0.14	-2.00*	2	-0.18	-0.14	-0.04	0.65	1
15	0.35	0.26	0.09	0.14	0.64	1	-0.74	-0.82	0.08	0.41	1
16	0.99	1	-0.01	0.14	-0.07	1	-0.24	-0.24	0	1	1
17	2.3	1.42	0.88	0.14	6.29*	2	0.79	0.08	.71*	0	2
18	1.31	1.52	-0.21	0.15	-1.4	1	0.01	0.16	-0.15	0.17	1
19	1.41	1.45	-0.04	0.15	-0.27	1	0.1	0.1	0	1	1
20	0.8	1.05	-0.25	0.14	-1.79	1	-0.39	-0.21	-0.18	0.09	1
21	1.09	1.39	-0.3	0.15	-2.00*	2	-0.16	0.06	-.22*	0.04	2
22	0.93	0.73	0.2	0.13	1.54	1	-0.29	-0.46	0.17	0.11	1
23	1.38	1.45	-0.07	0.15	-0.47	1	0.09	0.09	0	1	1
24	1.7	1.48	0.22	0.14	1.57	1	0.32	0.13	0.19	0.1	1
25	0.5	0.61	-0.11	0.13	-0.85	1	-0.62	-0.55	-0.07	0.5	1
26	1.26	1.29	-0.03	0.14	-0.21	1	-0.02	-0.02	0	1	1
27	0.75	0.75	0	0.14	0	1	-0.43	-0.43	0	1	1
28	1.19	0.74	0.45	0.14	3.21*	2	-0.08	-0.45	.37*	0	2
29	1.71	1.94	-0.23	0.16	-1.44	1	0.32	0.49	-0.17	0.16	1
30	1.24	1.52	-0.28	0.14	-2.00*	2	-0.04	0.16	-0.2	0.07	1
31	1.52	1.38	0.14	0.14	1	1	0.18	0.05	0.13	0.25	1
32	1.51	1.87	-0.36	0.16	-2.25*	2	0.17	0.44	-0.27	0.02	2
33	1.41	1.56	-0.15	0.15	-1	1	0.09	0.19	-0.1	0.38	1
34	2.23	2.76	-0.53	0.19	-2.79*	2	0.74	1.13	-0.39*	0.01	2
35	1.63	1.55	0.08	0.15	0.53	1	0.26	0.19	0.07	0.5	1
36	1.46	1.14	0.32	0.14	2.29*	2	0.13	-0.14	0.27*	0.02	2
37	1.25	1.47	-0.22	0.15	-1.47	1	-0.03	0.13	-0.16	0.14	1
38	1.95	1.97	-0.02	0.16	-0.13	1	0.52	0.52	0	1	1
39	1.15	1.23	-0.08	0.15	-0.53	1	-0.12	-0.06	-0.06	0.62	1
40	1.05	1.07	-0.02	0.14	-0.14	1	-0.19	-0.19	0	1	1
41	2.34	1.9	0.44	0.14	3.14*	2	0.82	0.46	.36*	0	2
42	1.27	1.53	-0.26	0.15	-1.73	1	-0.02	0.17	-0.19	0.08	1
43	0.56	0.67	-0.11	0.14	-0.79	1	-0.57	-0.5	-0.07	0.5	1
44	2.1	2.32	-0.22	0.17	-1.29	1	0.63	0.79	-0.16	0.22	1
45	1.89	1.96	-0.07	0.16	-0.44	1	0.49	0.49	0.00	1.00	1
46	3.84	2.66	1.18	0.18	6.56*	2	2.0	1.06	.94*	0.00	2
47	3.08	2.69	0.39	0.17	2.29*	2	1.4	1.09	.31*	0.03	2
48	1.57	1.38	0.19	0.14	1.36	1	0.21	0.05	0.16	0.15	1
49	3.41	2.98	0.43	0.18	2.39*	2	1.66	1.31	.35*	0.02	2
50	1.3	1.03	0.27	0.14	1.93	1	0.00	-0.22	.22*	0.03	2

Table 3 shows the DIF statistics of the Rasch model method for each of the 50 items for SES. An item is said to revealed DIF if the probability is less than 0.05. The Rasch model method flagged 15 items at the 0.05 level of significance. That is 30% of the 2012.WASSCE mathematics multiple-choice test items functioned differentially for examinees from high and low SES. The DIF items are 3, 5, 9, 11, 17, 21, 28, 32, 34, 36, 41, 46, 47, 49 and 50.

Table 3 shows the DIF statistics of the IRT-3P method for each of the 50 items. An item is said to reveal DIF if Z-score≥│1.96│ at p≤0.05. The IRT-3P method flagged 16 items. That is 32% of the 2012 WASSCE mathematics multiple-choice test functioned differentially for examinees from high and low SES. The DIF items are 3, 5, 9, 11, 14, 17, 21, 28, 30, 32, 34, 36, 41, 46, 47, and 49.

Table 4: Transformed Item Difficulty and Mantel-Haenszel Statistics for Socio-Economic Status

TRANSFORMED ITEM DIFFICULTY (TID) - (SES)								DIF INDEX	MANTEL-HAENSZEL			DIF INDEX
	P-Value		Z-Value		Delta							
ITEMS	H	L	H	L	H	L	Di		χ^2	PROB	ODDS RATIO	

											(IN LOGIT)	
1	0.65	0.58	0.39	0.21	14.6	13.84	-1.12*	2	1.59	0.21	0.13	1
2	0.6	0.55	0.26	0.13	14	13.52	-0.87	1	0.25	0.62	-0.06	1
3	0.44	0.31	-0.2	-0.49	12.4	11.04	-1.09*	2	5.19*	0.02	0.26	2
4	0.49	0.41	0	-0.22	13	12.12	-1.45*	2	0.02	0.88	-0.02	1
5	0.51	0.39	0.03	-0.28	13.1	11.88	-1.96*	2	7.97*	0.01	0.31	2
6	0.41	0.36	-0.2	-0.35	12.1	11.6	-0.95	1	2.08	0.15	-0.17	1
7	0.39	0.33	-0.3	-0.44	11.9	11.24	-0.67	1	0.63	0.43	-0.09	1
8	0.38	0.29	-0.3	-0.55	11.8	10.8	-1.10*	2	0.01	0.96	0.01	1
9	0.4	0.28	-0.3	-0.58	12	10.68	-1.24*	2	1.87	0.17	0.17	1
10	0.52	0.44	0.06	-0.15	13.2	12.4	-1.01*	2	0.01	0.93	0.01	1
11	0.4	0.26	-0.3	-0.64	12	10.44	-1.31*	2	3.42	0.06	0.23	1
12	0.45	0.39	-0.1	-0.28	12.5	11.88	-0.95	1	0.03	0.86	-0.02	1
13	0.24	0.18	-0.7	-0.91	10.2	9.36	-1.01*	2	0.54	0.46	-0.11	1
14	0.39	0.31	-0.3	-0.49	11.9	11.04	-0.97	1	0.15	0.69	-0.05	1
15	0.5	0.45	0	-0.12	13	12.52	-0.79	1	1.98	0.16	-0.16	1
16	0.4	0.33	-0.3	-0.44	12	11.24	-0.96	1	0.05	0.83	0.03	1
17	0.23	0.27	-0.7	-0.61	10.1	10.56	0.72	1	12.15*	0	-0.41	2
18	0.36	0.26	-0.4	-0.64	11.6	10.44	-1.11*	2	1.11	0.29	0.13	1
19	0.34	0.26	-0.4	-0.64	11.4	10.44	-1.08*	2	0.38	0.54	-0.08	1
20	0.43	0.32	-0.2	-0.47	12.3	11.12	-1.29*	2	4.23*	0.04	0.22	2
21	0.39	0.27	-0.3	-0.61	11.9	10.56	-1.24*	2	0.57	0.45	0.1	1
22	0.41	0.37	-0.2	-0.33	12.1	11.68	0.95	1	0.42	0.51	-0.07	1
23	0.35	0.27	-0.4	-0.61	11.5	10.56	-1.02*	2	0.04	0.84	-0.03	1
24	0.3	0.26	-0.5	-0.64	10.9	10.44	-0.17	1	0.09	0.76	-0.04	1
25	0.48	0.39	-0.1	-0.27	12.8	11.92	-1.01*	2	0.17	0.68	0.05	1
26	0.36	0.29	-0.4	-0.55	11.6	10.8	-0.94	1	0.04	0.84	-0.03	1
27	0.44	0.37	-0.2	-0.33	12.4	11.68	-0.91	1	0.03	0.87	-0.02	1
28	0.37	0.37	-0.3	-0.33	11.7	11.68	0.04	1	12.99*	0	-0.4	2
29	0.3	0.22	-0.5	-0.77	10.9	9.92	-1.03*	2	0.19	0.66	0.06	1
30	0.37	0.26	-0.3	-0.64	11.7	10.44	-1.03*	2	5.37*	0.02	0.26	2
31	0.33	0.27	-0.4	-0.61	11.2	10.56	-1.01*	2	0.41	0.52	-0.08	1
32	0.33	0.21	-0.4	-0.8	11.2	9.8	-1.65*	2	1.19	0.27	0.14	1
33	0.34	0.25	-0.4	-0.67	11.4	10.32	-1.29*	2	0.19	0.66	0.06	1
34	0.24	0.13	-0.7	-1.12	10.2	8.52	-1.86*	2	0.89	0.35	0.17	1
35	0.31	0.25	-0.5	-0.67	11	10.32	-1.08*	2	0.45	0.5	-0.09	1
36	0.34	0.31	-0.4	-0.67	11.4	10.32	-1.17*	2	2.7	0.1	-0.19	1
37	0.37	0.26	-0.3	-0.64	11.7	10.44	-1.17*	2	0.64	0.42	0.1	1
38	0.27	0.2	-0.6	-0.84	10.6	9.64	-1.10*	2	0.45	0.5	-0.1	1
39	0.38	0.29	-0.3	-0.55	11.8	10.8	-1.13*	2	0.22	0.64	-0.06	1
40	0.39	0.32	-0.3	-0.47	11.9	11.12	-0.69	1	0.27	0.6	0.06	1
41	0.23	0.21	-0.7	-0.8	10.1	9.8	-0.19	1	1.44	0.23	0.14	1
42	0.36	0.25	-0.4	-0.67	11.6	10.32	-1.31*	2	1.35	0.25	0.14	1
43	0.47	0.38	-0.1	-0.3	12.7	11.8	-1.14*	2	0.11	0.74	0.04	1
44	0.25	0.17	-0.6	-0.95	10.6	9.2	-1.23*	2	0.01	0.92	0.02	1
45	0.28	0.2	-0.6	-0.84	10.7	9.64	-1.07*	2	0.09	0.76	0.04	1
46	0.1	0.14	-1.3	-1.08	7.88	8.68	1.20*	2	4.09*	0.04	-0.31	2
47	0.16	0.13	-1	-1.12	9.04	8.52	-0.31	1	0.31	0.58	-0.09	1
48	0.32	0.27	-0.4	-0.61	11.6	10.56	-1.05*	2	0.32	0.57	-0.07	1
49	0.13	0.11	-1.1	-1.22	8.52	8.12	-0.47	1	0.19	0.67	-0.07	1
50	0.36	0.32	-0.4	-0.47	11.6	11.12	-0.22	1	1.85	0.17	-0.16	1

Table 4 shows the DIF statistics of the TID method for each of the 50 items. An item is said to flag DIF if │Di│ values is in excess of one standard deviation. The TID method flagged 32 items at the 0.05 level of significance. That is 64% of the 2012 mathematics multiple-choice test functioned differentially for examinees from high and low SES. The DIF items are 1, 3, 4, 5, 8, 9, 10, 11, 13, 18, 19, 20, 21, 23, 25, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 42, 43, 44, 45, 46, and 48.

Table 4 shows the DIF statistics of the Mantel-Haenszel method for each of the 50 items. An item is said to flag DIF if the probability is less than 0.05. The M-H method flagged 7 items at the 0.05 level of significance. That is 14% of the 2012 WASSCE mathematics multiple-choice test items functioned differentially for examinees from high and low SES. The DIF items are 3, 5, 17, 20, 28, 30 and 46

Research Question 3: What is the index of DIF for location under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for each item in 2012 WASSCE mathematics multiple-choice test?

Table 5: IRT-3P and Rasch Models Statistics for Location

IRT-3P							RASCH MODEL (LOCATION)				
ITEMS	b-value U	R	b dif	SEbd	Z-score	DIF INDEX	bu	br	Δb	Prob	DIF INDEX

1	-0.69	0.32	-1.01	0.2	-5.05*	2	-1.68	-1.2	-.48*	0	2
2	-0.4	0.81	-1.21	0.2	-6.05*	2	-1.53	-0.94	-.59*	0	2
3	1.79	2.35	-0.56	0.22	-2.55*	2	-0.39	-0.13	-.26*	0.02	2
4	0.51	2.56	-2.05	0.22	-9.32*	2	-1.05	-0.03	-1.02*	0	2
5	0.72	2.25	-1.53	0.21	-7.29*	2	-0.95	-0.19	-.76*	0	2
6	1.81	2.06	-0.26	0.21	-1.24	1	-0.38	-0.28	-0.1	0.39	1
7	1.2	2.54	-1.34	0.22	-6.09*	2	-0.3	-0.03	-.27*	0.02	2
8	2.29	2.62	-0.33	0.22	-1.5	1	-0.13	0.01	-0.14	0.23	1
9	1.2	3.12	-1.92	0.23	-8.35*	2	-0.3	0.27	-.57*	0	2
10	0.81	1.31	-0.5	0.2	-2.50*	2	-0.9	-0.67	-.23*	,03	2
11	2.09	3.33	-1.24	0.24	-5.17*	2	-0.23	0.38	-.61*	0	2
12	1.49	1.75	-0.26	0.2	-1.3	1	-0.54	-0.45	-0.09	0.36	1
13	3.7	4.4	-0.7	0.28	-2.50*	2	0.6	0.94	-.34*	0.01	2
14	2.21	2.33	-0.12	0.22	-0.55	1	-0.16	-0.16	0	1	1
15	0.76	1.47	-0.71	0.21	-3.38*	2	-0.92	-0.59	-.33*	0	2
16	2.12	2.08	0.04	0.21	0.19	1	-0.22	-0.27	0.05	0.63	1
17	4.28	1.93	2.35	0.22	10.68*	2	0.9	-0.36	1.26*	0	2
18	2.81	2.62	0.19	0.23	0.83	1	0.14	0.01	0.13	0.25	1
19	2.55	3.16	-0.61	0.24	-13.17*	2	0.01	0.29	-.28*	0.01	2
20	2.09	1.86	0.23	0.23	8.09*	2	-0.23	-0.39	0.16	0.14	1
21	2.29	2.81	-0.52	0.52	-1	1	-0.13	0.11	-.24*	0.04	2
22	2.34	1.1	1.24	0.2	6.20*	2	-0.1	-0.79	.69*	0	2
23	2.61	2.95	-0.34	0.33	-1.03	1	0.04	0.18	-0.14	0.23	1
24	2.87	3.19	-0.32	0.23	-1.59	1	0.17	0.31	-0.14	0.26	1
25	1.41	1.5	-0.09	0.2	-0.45	1	-0.58	-0.58	0	1	1
26	2.62	2.35	0.27	0.22	1.23	1	0.04	-0.13	0.17	0.11	1
27	1.64	1.86	-0.22	0.21	-1.05	1	-0.47	-0.39	-0.08	0.46	1
28	2.23	1.7	0.53	0.21	2.52*	2	-0.16	-0.47	.31*	0	2
29	3.24	3.59	-0.35	0.25	-1.4	1	0.36	0.51	-0.15	0.23	1
30	2.92	2.32	0.6	0.22	-2.73*	2	0.2	-0.15	.35*	0	2
31	3.04	2.35	0.69	0.22	3.14*	2	0.26	-0.13	.39*	0.01	2
32	3.1	3.25	-0.15	0.25	-0.6	1	0.31	0.34	-0.03	0.79	1
33	2.98	2.62	0.36	0.23	1.57	1	0.23	0.01	0.22	0.06	1
34	4.11	4.96	-0.85	0.32	-2.66*	2	0.81	1.24	-.43*	0.01	2
35	3.42	2.3	1.12	0.22	5.09*	2	0.45	-0.16	.61*	0	2
36	2.87	2	0.87	0.21	4.13*	2	0.17	-0.32	.49*	0	2
37	2.35	3.25	-0.9	0.24	-3.75*	2	-0.1	0.34	-.44*	0	2
38	3.68	3.33	0.35	0.25	1.4	1	0.59	0.38	0.21	0.09	1
39	2.11	2.91	-0.81	0.23	-3.52*	2	-0.22	0.16	-.38*	0.01	2
40	2.4	1.88	0.52	0.21	2.48*	2	-0.07	-0.38	.31*	0.01	2
41	4.47	2.63	1.84	0.22	8.76*	2	1	0.01	.99*	0	2
42	2.82	2.56	0.26	0.23	1.13	1	0.14	-0.02	0.16	0.14	1
43	1.26	1.96	-0.7	0.21	-2.5*	2	-0.66	-0.34	-.32*	0.01	2
44	3.78	4.27	-0.49	0.28	-1.75	1	0.54	0.87	-0.23	0.09	1
45	3.63	3.25	0.38	0.24	1.58	1	0.57	0.34	0.23	0.07	1
46	6.52	3.61	2.91	0.27	10.78*	2	2.07	0.52	1.55*	0	2
47	5.15	4.49	0.66	0.29	2.28*	2	1.35	0.99	.36*	0.02	2
48	3.18	2.22	0.96	0.22	0.04	1	0.33	-0.2	.53*	0	2
49	5.87	4.43	1.44	0.29	4.97*	2	1.73	0.95	.78*	0	2
50	2.6	1.9	0.7	0.21	3.33*	2	0.03	-0.37	.40*	0	2

Table 5 shows the DIF statistics of the Rasch model method for each of the 50 items for location. An item is said to revealed DIF if the probability is less than 0.05. The Rasch model method flagged 31 items at the 0.05 level of significance. That is 62% of the

2012.WASSCE mathematics multiple-choice test items functioned differentially for examinees from urban and rural environment. The DIF items are 1, 2, 3, 4, 5, 7, 9, 10, 11, 13, 15, 17, 19, 21, 22, 22, 28, 30, 31, 34, 35, 36, 37, 38, 40, 41, 43, 46, 47, 48 and 50

Table 5 shows the DIF statistics of the IRT-3P method for each of the 50 items. An item is said to reveal DIF if Z-score≥│1.96│ at p≤0.05. The IRT-3P method flagged 30 items. That is 60% of the 2012 WASSCE mathematics multiple-choice test functioned differentially for examinees from rural and urban location. The DIF items are 1, 2, 3, 4, 5, 7, 9, 10, 11, 13, 15, 17, 19, 20, 22, 28, 30, 31, 34, 35, 36, 37, 39, 40, 41, 43, 46, 47, 49, and 50.

Table 6: Transformed Item Difficulty and M-H Statistics for Location

TID												
									MANTEI-HAENSZEL			
	P-Value		Z-Value		Delta				LOCATION			
								DIF INDEX			ODDS RATIO (IN	DIF INDEX
ITEMS	U	R	U	R	U	R	Di		χ^2	PROB		

											LOGIT)	
1	0.71	0.46	0.56	-0.1	15.24	12.6	-1.92*	2	40.47*	0	0.69	2
2	0.68	0.41	0.46	-0.22	14.84	12.12	-1.86*	2	30.16*	0	0.61	2
3	0.46	0.24	-0.1	-0.7	12.6	10.2	-1.73*	2	0.98	0.32	0.13	1
4	0.6	0.22	0.26	-0.77	14.04	9.92	-1.98*	2	65.27*	0	0.95	2
5	0.57	0.25	0.18	-0.67	13.72	10.32	-1.86*	2	43.87*	0	0.78	2
6	0.46	0.27	-0.1	-0.61	12.6	10.56	-1.89*	2	0.13	0.72	-0.05	1
7	0.44	0.23	-0.15	-0.73	12.4	10.08	-1.91*	2	4.00*	0.04	0.25	2
8	0.41	0.22	-0.22	-0.77	12.12	9.92	-1.98*	2	0.25	0.62	-0.07	1
9	0.44	0.18	-0.15	-0.91	12.4	9.36	-1.99*	2	5.21*	0.02	0.3	2
10	0.56	0.35	0.16	-0.36	13.64	11.56	-1.94*	2	1.47	0.22	0.14	1
11	0.43	0.17	-0.17	-0.95	12.32	9.2	-1.91*	2	3.04	0.08	0.25	1
12	0.49	0.3	-0.01	-0.52	12.96	10.92	-1.72*	2	4.37*	0.04	0.24	2
13	0.28	0.1	-0.58	-1.28	10.68	7.88	-1.93*	2	2.41	0.12	0.25	1
14	0.41	0.24	-0.22	-0.7	12.12	10.2	-1.74*	2	3.29	0.07	-0.24	1
15	0.57	0.33	0.18	-0.44	13.72	11.24	-1.82*	2	4.39*	0.04	0.24	2
16	0.42	0.27	-0.2	-0.61	12.2	10.56	-1.87*	2	0.62	0.43	0.1	1
17	0.23	0.28	-0.73	-0.58	10.08	10.68	0.67	1	43.67*	0	-0.84	2
18	0.36	0.22	-0.35	-0.77	11.6	9.92	-1.89*	2	4.13*	0.04	-0.27	2
19	0.38	0.18	-0.3	-0.91	11.8	9.36	-1.80*	2	0.12	0.73	0.06	1
20	0.43	0.29	-0.17	-0.55	12.32	10.8	-1.74*	2	0.02	0.89	-0.02	1
21	0.41	0.2	-0.22	-0.84	12.12	9.64	-1.93*	2	0.88	0.35	-0.13	1
22	0.4	0.37	-0.25	-0.33	12	11.68	-1.70*	2	15.44*	0	-0.45	2
23	0.38	0.19	-0.3	-0.87	11.8	9.52	-1.96*	2	0.09	0.75	0.05	1
24	0.35	0.17	-0.36	-0.95	11.56	9.2	-1.86*	2	8.95*	0	0.39	2
25	0.5	0.33	0	-0.44	13	11.24	-1.65*	2	0.01	0.98	.0.1	1
26	0.37	0.26	-0.33	-0.7	11.68	10.2	-1.35*	2	1.03	0.31	-0.13	1
27	0.47	0.28	-0.07	-0.58	12.72	10.68	-1.79*	2	1	0.32	0.12	1
28	0.41	0.31	-0.22	-0.49	12.12	11.04	-1.42*	2	13.69*	0	-0.45	2
29	0.32	0.15	-0.28	-1.03	11.88	8.88	-1.67*	2	0.05	0.82	-0.04	1
30	0.35	0.25	-0.36	-0.67	11.56	10.32	-1.21*	2	2.71	0.09	-0.21	1
31	0.33	0.24	-0.44	-0.7	11.24	10.2	-1.19*	2	6.46*	0.01	-0.33	2
32	0.33	0.17	-0.44	-0.95	11.24	9.2	-1.58*	2	4.74*	0.03	-32	2
33	0.34	0.22	-0.41	-0.77	11.36	9.92	-1.73*	2	4.64*	0.03	-0.29	2
34	0.25	0.08	-0.67	-1.4	10.32	7.4	-1.86*	2	1.21	0.27	-0.25	1
35	0.3	0.25	-0.52	-0.67	10.92	10.32	-0.46	1	17.72*	0	-0.5	2
36	0.35	0.28	-0.36	-0.58	11.56	10.68	-1.16*	2	9.58*	0	-0.38	2
37	0.61	0.17	-0.28	-0.95	11.8	9.2	-1.63*	2	3.81*	0.05	0.27	2

					8							
38	0.28	0.17	-0.58	-0.95	10.68	9.2	-1.35*	2	7.71*	0.01	-0.42	2
39	0.42	0.2	-0.2	-0.84	12.2	9.64	-1.94*	2	1.35	0.24	0.16	1
40	0.4	0.29	-0.25	-0.55	12	10.8	-1.49*	2	4.32*	0.04	-0.25	2
41	0.22	0.22	-0.77	-0.77	9.92	9.92	0.45	1	0.24	0.63	-0.07	1
42	0.36	0.22	-0.35	-0.77	11.6	9.92	-1.77*	2	2.17	0.14	-0.2	1
43	0.51	0.28	0.03	-0.58	13.12	10.68	-1.67*	2	3.80*	0.05	0.23	2
44	0.27	0.11	-0.61	-1.22	10.56	8.12	-1.59*	2	0.45	0.5	0.12	1
45	0.28	0.17	-0.58	-0.95	10.68	9.2	-1.23*	2	2.94	0.09	-0.25	1
46	0.1	0.45	-1.28	-0.12	7.88	12.52	1.97*	2	11.54*	0	-0.51	2
47	0.17	0.1	-0.95	-1.28	9.2	7.88	-1.45*	2	0.52	0.47	0.13	1
48	0.32	0.26	-0.28	-0.64	11.88	10.44	-1.48*	2	3.82*	0.05	-0.24	2
49	0.13	0.1	-1.12	-1.28	8.52	7.88	0.98	1	2.88	0.09	0.3	1
50	0.38	0.29	-0.3	-0.55	11.8	10.8	-1.03*	2	9.77*	0	0.38	2

Table 6 shows the DIF statistics of the TID method for each of the 50 items. An item is said to flag DIF if ⎜Di⎜ values is in excess of one standard deviation. The TID method flagged 46 items at the 0.05 level of significance. That is 92% of the 2012 mathematics multiple-choice test functioned differentially for examinees from rural and urban location. The DIF items are 1,2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48 and 50.

Table 6 shows the DIF statistics of the Mantel-Haenszel method for each of the 50 items. An item is said to flag DIF if the probability is less than 0.05. The M-H method flagged 25 items at the 0.05 level of significance. That is 50% of the 2012 WASSCE mathematics multiple-choice test items functioned differentially for examinees from rural and urban location. The DIF items are 1, 2, 4, 5, 7, 9, 12, 15, 17, 18, 22, 24, 28, 31, 32, 33, 35, 36, 37, 38, 40, 43, 46, 48, and 50.

The Agreement between the index of DIF for Gender, SES and Location under the methods of IRT and CTT mathematics multiple-choice test

Research Question 4: What is the agreement between the index of DIF for gender under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test?

Hypothesis 1: There is no significant agreement between the index of DIF for gender under the methods of Item Response Theory (Rasch model and IRT-3P) and Classical Test Theory (Transformed Item Difficulty and M-H) for items in 2012 WASSCE mathematics multiple-choice test.

In order to answer research question 4 and test its corresponding null hypothesis 1, frequency count, percentage, chi-square independent test and contingency coefficient were employed. The results obtained from the analysis are presented in table 7.

Table 7: Level of Agreement of methods of detecting DIF between CTT and IRT for Gender

Methods	ADIF	ANDIF	%	χ^2	df	Sig.	C
Rasch Vs TID	6	15	42	0.97	1	0.324	0.14
Rasch Vs M-H	6	34	80	10.44	1	0.001*	0.42
IRT-3P Vs TID	7	14	42	0.99	1	0.318	0.14
IRT-3P Vs M-H	6	32	76	8.09	1	0.004*	0.37

*Significant at α≤.05, ADIF=No of agreed DIF items, ANDIF=No of agreed non DIF items, %=Percentage of agreement, C=Contingency coefficient.

Table 7 shows that the Rasch and TID methods were agreeable in allocating 6 items as revealing DIF, and 15 items as not revealing DIF. As such, the percentage of agreement between Rasch and TID is 42% [i.e. ((15+6)/50)x 100 = 42%]. This shows that there is a low agreement between the index of DIF for gender under the Rasch and TID methods of detecting DIF.

Table 7 shows that the Rasch and M-H methods were agreeable in allocating 6 items as revealing DIF, and 34 items as not revealing DIF. As such, the percentage of agreement between Rasch and M-H is 80%. This shows that there is a high agreement between the index of DIF for gender under the Rasch and M-H methods of detecting DIF.

Table 7 shows that the IRT-3P and TID methods were agreeable in allocating 7 items as revealing DIF, and 14 items as not revealing DIF. As such, the percentage of agreement between IRT-3P and TID is 42%**.** This shows that there is a low agreement between the index of DIF for gender under the IRT-3P and TID methods of detecting DIF.

Table 7 shows that the IRT-3P and M-H methods were agreeable in allocating 8 items as revealing DIF, and 32 items as not revealing DIF. As such, the percentage of agreement between IRT-3P and TID is 76%. This shows that there is a high agreement between the index of DIF for gender under the IRT-3P and M-H methods of detecting DIF.

Table 7: shows the chi-square statistics analysis between Rasch model and TID for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 0.97 was found not significant at df =1, p>.05. The null hypothesis which states that there is no significant agreement between the index of DIF for gender under the methods of IRT (Rasch model) and CTT (TID) for items in 2012 WASSCE mathematics multiple-choice test was accepted.

Table 7: shows the chi-square statistics analysis between Rasch model and M-H for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 10.44 was found significant at df =1, p<.05. The null hypothesis is rejected. There is a significant agreement between the index of DIF for gender under the methods of IRT (Rasch model) and CTT (M-H) for items in 2012 WASSCE mathematics multiple-choice test. The contingency coefficient value is 0.42, which is a moderate agreement between the two methods since for a 2x2 table; the maximum value of C is 0.707.

Table 7: shows the chi-square statistics analysis between IRT-3P and TID for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 0.99 was found not significant at df =1, p>.05. The null hypothesis which states that there is no significant agreement between the index of DIF for gender under the methods of IRT (IRT-3P) and CTT (TID) for items in 2012 WASSCE mathematics multiple-choice test was accepted.

Table 7: shows the chi-square statistics analysis between IRT-3P and M-H for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square value of 8.09 was found significant at df =1, p<.05. The null hypothesis is rejected. There is a significant agreement between the index of DIF for gender under the methods of IRT (IRT-3P) and CTT (M-H) for items in 2012 WASSCE mathematics multiple-choice test. The contingency

coefficient value is 0.37 which indicates a moderate agreement between the two methods in DIF detection since for a 2x2 table the maximum value of C is 0.707.

Research Question 5: What is the agreement between the index of DIF for gender within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test?

Hypothesis2: There is no significant agreement between the index of DIF for gender within the methods CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test.

In order to answer research question 5 and test its corresponding null hypothesis 2, frequency count, percentage, chi-square independent test and contingency coefficient were employed. The results obtained from the analysis are presented in table 8.

Table 8: Level of Agreement of methods of detecting DIF within CTT for Gender

Methods	ADIF	ANDIF	%	χ^2	df	Sig.	C
M-H Vs TID	3	18	42	1.04	1	0.307	0.14

ADIF=No of agreed DIF items, ANDIF=No of agreed non DIF items, %=Percentage of agreement, C=Contingency

Table 8 shows that the M-H and TID methods were agreeable in allocating 3 items as revealing DIF, and 18 items as not revealing DIF. As such, the percentage of agreement between M-H and TID is 42%. This shows that there is a low agreement between the index of DIF for gender under the M-H and TID methods of detecting DIF.

Table 8 shows the chi-square statistics analysis between M-H and TID for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 1.04 was found not significant at df =1, p>.05. The null hypothesis which states that there is no significant agreement between the index of DIF for gender under the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test was accepted.

Research Question 6: What is the agreement between the index of DIF for gender within the methods of IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test?

Hypothesis 3: There is no significant agreement between the index of DIF for gender within the methods of IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test.

In order to answer research question 6 and test its corresponding null hypothesis 3, frequency count, percentage, chi-square independent test and contingency coefficient were employed. The results obtained from the analysis are presented in table 9.

Table 9: Level of Agreement of methods of detecting DIF within IRT for Gender

Methods	ADIF	ANDIF	%	χ^2	df	Sig.	C
IRT-3P Vs Rasch	15	34	98	45.54	1	0.000*	0.69

*Significant at α≤.05, ADIF=No of agreed DIF items, ANDIF=No of agreed non DIF items, %=Percentage of agreement, C=Contingency

Table 9 shows that the IRT-3P and Rasch methods were agreeable in allocating 15 items as revealing DIF, and 34 items as not revealing DIF. As such, the percentage of agreement between IRT-3P and Rasch is 98%. This shows that there is a high agreement between the index of DIF for gender under the IRT-3P and Rasch methods of detecting DIF.

Table 9 shows the chi-square statistics analysis between IRT-3P and Rasch model for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 45.54 was found significant at df =1, p=.05. The null hypothesis is rejected. There is a significant agreement between the index of DIF for gender under the methods of IRT-3P and Rasch model for items in 2012 WASSCE mathematics multiple-choice test. The contingency coefficient value is 0.69 which indicates a high agreement between the two methods since for a 2x2 table the maximum value of C is 0.707

Research Question 7: What is the agreement between the index of DIF for SES under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test?

Hypothesis 4: There is no significant agreement between the index of DIF for SES under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics test.

In order to answer research question 7 and test its corresponding null hypothesis 4, frequency count, percentage, chi-square independent test and contingency coefficient were employed. The results obtained from the analysis are presented in table 10

Table 10: Level of Agreement of methods of detecting DIF between CTT and IRT for SES

Methods	ADIF	ANDIF	%	χ^2	df	Sig.	C
Rasch Vs TID	9	12	42	0.15	1	0.700	0.06
Rasch Vs M-H	5	33	76	6.65	1	0.010*	0.34
IRT-3P Vs TID	10	12	44	0.02	1	0.880	0.02
IRT-3P Vs M-H	6	33	78	10.79	1	0.001*	0.42

*Significant at α≤.05, ADIF=No of agreed DIF items, ANDIF=No of agreed non DIF items, %=Percentage of agreement, C=Contingency coefficient.

Table 10 shows that the Rasch and TID methods were agreeable in allocating 9 items as revealing DIF, and 12 items as not revealing DIF. As such, the percentage of agreement between Rasch and TID is 42%. This shows that there is a low agreement between the index of DIF for SES under the Rasch and TID methods of detecting DIF.

Table 10 shows that the Rasch and M-H methods were agreeable in allocating 5 items as revealing DIF, and 33 items as not revealing DIF. As such, the percentage of agreement between Rasch and M-H is 76%. This shows that there is a high agreement between the index of DIF for SES under Rasch and M-H methods of detecting DIF.

Table 10 shows that the IRT-3P and TID methods were agreeable in allocating 10 items as revealing DIF, and 12 items as not revealing DIF. As such, the percentage of agreement between IRT-3P and TID is 44%. This shows that there is a low agreement between the index of DIF for SES under the IRT-3P and TID methods of detecting DIF.

Table 10 shows that the IRT-3P and M-H methods were agreeable in allocating 6 items as revealing DIF, and 33 items as not revealing DIF. As such, the percentage of agreement between IRT-3P and M-H is 78%. This shows that there is a high agreement between the index of DIF for SES under the IRT-3P and M-H methods of detecting DIF.

Table 10 shows the chi-square statistics analysis between IRT (Rasch model) and CTT (TID) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square is 0.15 was found not significant at df =1, p>.05. The null hypothesis which states that there is no significant agreement between the index of DIF for SES under the methods of Rasch model and TID for items in 2012 WASSCE mathematics multiple-choice test was accepted.

Table 10 shows the chi-square statistics analysis between IRT (Rasch model) and CTT (M-H) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square is 6.65 was found significant at df =1, p<.05. The null hypothesis is rejected. There is a significant agreement between the index of DIF for SES under the methods of Rasch and M-h for items in 2012 WASSCE mathematics multiple-choice test. The contingency coefficient value is 0.34 which indicate a small agreement between the two methods since for a 2x2 table the maximum value of C is 0.707.

Table 10 shows the chi-square statistics analysis between IRT (IRT-3P) and CTT (TID) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 0.02 was found not significant at df =1, p>.05. The null hypothesis which states that there is no significant agreement between the index of DIF for SES under the methods of IRT-3P and TID for items in 2012 WASSCE mathematics multiple-choice test was accepted.

Table 10 shows the chi-square statistics analysis between IRT (IRT-3P) and CTT (M-H) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 10.79 was found is significant at df =1, p<.05. The null hypothesis is rejected. There is a significant agreement between the index of DIF for SES under the methods of IRT-3P and M-H for items in 2012 WASSCE mathematics multiple-choice test. The contingency coefficient is 0.42, which is a moderate agreement between the two methods since for 2x2 table the maximum value of C is 0.707.

Research Question 8: What is the agreement between the index of DIF for SES within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test?

Hypothesis 5: There is no significant agreement between the index of DIF for SES within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics test.

In order to answer research question 8 and test its corresponding null hypothesis 5, frequency count, percentage, chi-square independent test and contingency coefficient were employed. The results obtained from the analysis are presented in table 11.

Table 11: Level of Agreement of methods of detecting DIF within CTT for SES

Methods	ADIF	ANDIF	%	χ^2	df	Sig.	C
M-H Vs TID	5	16	42	0.20	1	0.659	0.06

ADIF=No of agreed DIF items, ANDIF=No of agreed non DIF items, %=Percentage of agreement, C=Contingency coefficient.

Table 11 shows that the M-H and TID methods were agreeable in allocating 5 items as revealing DIF, and 16 items as not revealing DIF. As such, the percentage of agreement between M-H and TID is 42%. This shows that there is a low agreement between the index of DIF for SES under the M-H and TID methods of detecting DIF.

Table 11 shows the chi-square statistics analysis between CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 0.20 was found not significant at df =1, p>.05. The null hypothesis which states that there is no significant agreement between the index of DIF for SES under the methods of M-H and TID for items in 2012 WASSCE mathematics multiple-choice test was accepted.

Research Question 9: What is the agreement between the index of DIF for SES within the methods of IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test?

Hypothesis 6: There is no significant agreement between the index of DIF for SES within the methods of IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics test.

In order to answer research question 9 and test its corresponding null hypothesis 6, frequency count, percentage, chi-square independent test and contingency coefficient were employed. The results obtained from the analysis are presented in table 12.

Table 12: Level of Agreement of methods of detecting DIF within IRT for SES

Methods	ADIF	ANDIF	%	χ^2	df	Sig.	C
IRT-3P Vs Rasch	14	33	94	37.05	1	0.000*	0.65

*Significant at α≤.05, ADIF=No of agreed DIF items, ANDIF=No of agreed non DIF items, %=Percentage of agreement, C=Contingency coefficient.

Table 12 shows that the IRT-3P and Rasch methods were agreeable in allocating 14 items as revealing DIF, and 33 items as not revealing DIF. As such, the percentage of agreement between IRT-3P and Rasch is 94%. This shows that there is a high agreement between the index of DIF for SES under the IRT-3P and Rasch methods of detecting DIF.

Table 12 shows the chi-square statistics analysis between IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 37.05 was found significant at df =1, p<.05. The null hypothesis is rejected. There is a significant agreement between the index of DIF for SES under the methods of IRT-3P and Rasch model for items in 2012 WASSCE mathematics multiple-choice test. The contingency coefficient value is 0.65 which indicates a high agreement between the two methods since for a 2x2 table the maximum value of C is 0.707.

Research Question 10: What is the agreement between the index of DIF for location under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test?

Hypothesis 7: There is no significant agreement between the index of DIF for location under the methods of IRT (Rasch model and IRT-3P) and CTT (TID and M-H) for items in 2012 WASSCE mathematics multiple-choice test.

In order to answer research question 10 and test its corresponding null hypothesis 7, frequency count, percentage, chi-square independent test and contingency coefficient were employed. The results obtained from the analysis are presented in table 13.

Table 13: Level of Agreement of methods of detecting DIF between CTT and IRT for Location

Methods	ADIF	ANDIF	%	χ^2	df	Sig.	C
Rasch Vs TID	27	0	54	2.67	1	0.103	0.23
Rasch Vs M-H	19	13	64	4.16	1	0.041*	0.28
IRT-3P Vs TID	26	0	52	2.90	1	0.089	0.23
IRT-3P Vs M-H	18	13	62	3.00	1	0.083	0.24

*Significant at α≤.05, ADIF=No of agreed DIF items, ANDIF=No of agreed non DIF items, %=Percentage of agreement, C=Contingency coefficient.

Table 13 shows that the Rasch and TID methods were agreeable in allocating 27 items as revealing DIF, and 0 items as not revealing DIF. As such, the percentage of agreement between Rasch and TID is 54%. This shows that there is a moderate agreement between the index of DIF for location under the Rasch and TID methods of detecting DIF.

Table 13 shows that the Rasch and M-H methods were agreeable in allocating 19 items as revealing DIF, and 13 items as not revealing DIF. As such, the percentage of agreement between Rasch and M-H is 64%. This shows that there is a moderate agreement between the index of DIF for location under the Rasch and M-H methods of detecting DIF.

Table 13 shows that the IRT-3P and TID methods were agreeable in allocating 26 items as revealing DIF, and 0 item as not revealing DIF. As such, the percentage of agreement between IRT-3P and TID is 52%. This shows that there is a moderate agreement between the index of DIF for location under the IRT-3P and TID methods of detecting DIF.

Table 13 shows that the IRT-3P and M-H methods were agreeable in allocating 18 items as revealing DIF, and 13 items as not revealing DIF. As such, the percentage of agreement between IRT-3P and M-H is 62%. This shows that there is a moderate agreement between the index of DIF for location under the IRT-3P and M-H methods of detecting DIF.

Table 13 shows the chi-square statistics analysis between IRT (Rasch model) and CTT (TID) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 2.67 was found not significant at df =1, p>.05. The null hypothesis which states that there is no significant agreement between the index of DIF for location under the methods of Rasch model and TID for items in 2012 WASSCE mathematics multiple-choice test was accepted.

Table 13 shows the chi-square statistics analysis between IRT (Rasch model) and CTT (M-H) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 4.16 was found is significant at df =1, p<.05. The null hypothesis is rejected. There is a significant agreement between the index of DIF for location under the methods of Rasch model and M-H for items in 2012 WASSCE mathematics multiple-choice test. The contingency coefficient value is 0.28 which indicate a low agreement between the two methods since for a 2x2 table the maximum value of C is 0.707.

Table 13 shows the chi-square statistics analysis between IRT (IRT-3P) and CTT (TID) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 2.90 was found not significant at df =1, p>.05. The null hypothesis is which states that there is no significant agreement between the index of DIF for location under the methods of IRT-3P and TID for items in 2012 WASSCE mathematics multiple-choice test was accepted.

Table 13 shows the chi-square statistics analysis between IRT (IRT-3P) and CTT (M-H) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 3.00 was found not significant at df =1, p>.05. The null hypothesis which states that There is no significant agreement between the index of DIF for location under the methods of IRT-3P and M-H for items in 2012 WASSCE mathematics multiple-choice test was accepted.

Research Question 11: What is the agreement between the index of DIF for location within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test?

Hypothesis 8: There is no significant agreement between the index of DIF for location within the methods of CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test.

In order to answer research question 11 and test its corresponding null hypothesis 8, frequency count, percentage, chi-square independent test and contingency coefficient were employed. The results obtained from the analysis are presented in table 14.

Table 14: Level of Agreement of methods of detecting DIF within CTT for Location

Methods	ADIF	ANDIF	%	χ^2	df	Sig.	C
M-H Vs TID	23	2	50	0.01	1	0.999	0.01

ADIF=No of agreed DIF items, ANDIF=No of agreed non DIF items, %=Percentage of agreement, C=Contingency coefficient.

Table 14 shows that the M-H and TID methods were agreeable in allocating 23 items as revealing DIF, and 2 items as not revealing DIF. As such, the percentage of agreement between M-H and TID is 50%. This shows that there is a moderate agreement between the index of DIF for location under the M-H and TID methods of detecting DIF.

Table 14 shows the chi-square statistics analysis between CTT (M-H and TID) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of 0.00 was found not significant at df =1, p>.05. The null hypothesis which states that there is no significant agreement between the index of DIF for location under the methods of M-H and TID for items in 2012 WASSCE mathematics multiple-choice test was accepted.

Research Question 12: What is the agreement between the index of DIF for location within the methods of IRT (Rasch model and IRT-3P) for items in 2012 WASSCE mathematics multiple-choice test?

Hypothesis 9: There is no significant agreement between the index of DIF for location within the methods of IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test.

In order to answer research question 12 and test its corresponding null hypothesis 8, frequency count, percentage, chi-square independent test and contingency coefficient were employed. The results obtained from the analysis are presented in table 15.

Table 15: Level of Agreement of methods of detecting DIF within IRT for Location

Methods	ADIF	ANDIF	%	χ^2	df	Sig.	C
IRT-3P Vs Rasch	29	18	94	38.26	1	0.000*	0.66

*Significant at α≤.05, ADIF=No of agreed DIF items, ANDIF=No of agreed non DIF items, %=Percentage of agreement, C=Contingency coefficient.

Table 15 shows that the Rasch and IRT-3P methods were agreeable in allocating 29 items as revealing DIF, and 18 items as not revealing DIF. As such, the percentage of agreement between Rasch and IRT-3P is 94%. This shows that there is a high agreement between the index of DIF for location under the Rasch and IRT-3P methods of detecting DIF.

Table 15 shows the chi-square statistics analysis between IRT (IRT-3P and Rasch model) for items in 2012 WASSCE mathematics multiple-choice test. The calculated chi-square of38.26 was found significant at df =1, p<.05. The null hypothesis is rejected. There is a significant agreement between the index of DIF for location under the methods of IRT-3P and Rasch model for items in 2012 WASSCE mathematics multiple-choice test. The contingency coefficient value is 0.66, which indicates a high agreement between the two methods since for a 2x2 table the maximum value of C is 0.707.

Discussion of Results

Index of Differential Item Functioning (DIF) under the methods of CTT and IRT

The four methods used in this study namely Rasch model, IRT-3P, M-H, and TID; all revealed the presence of gender DIF items in 2012 WASSCE mathematics multiple-choice test. In other words mathematics multiple-choice test used by WASSCE contained items that function differently for students with the same mathematics ability for different sex (male/female). That is, there are items in 2012 mathematics multiple-choice test that measured different things for boys and girls with the same mathematics ability. This result agrees with similar research result reported by Odili (2003). His result showed that there was evidence of gender DIF in WAES/SSCE biology paper 2 for 1999, 2000 and 2001. According to Umoinyang (1991) the mathematics multiple choice test used by WAEC in the 1990 General certificate Examination also contains test items with significant gender DIF. Incidence of gender DIF was also reported by Abedalaziz (2010) in mathematics.

All the four methods of detecting DIF used in this study revealed SES differential item functioning. In order words 2012 WASSCE mathematics multiple-choice test contained items that function differentially among examinees from low and high SES. There are items in 2012 WASSCE mathematics that measure differentially among examinees from low and high SES. This result is in agreement with similar result reported by Odili (2003). His study showed evidence of the presence of SES differential item functioning items in WAEC/SSCE biology paper 2 for 1999 and 2001. According to Green (1980) as cited by Odili (2003), if test requires that students have knowledge and skills not taught in school, difference in performance in the test will no longer be based on the achievement of the common knowledge and skills taught in school. Literature on socio-economic DIF status is scanty; this was also observed by Odili (2003). However, studies have reported that SES affects students' academic achievement (Barry, 2005; Eamon, 2005; and Hachschid, 2003). Consequently, argument should not be limited to just closing the gap between the social status of the populace but should also extend to making allowance for differences in SES in our teaching and learning processes. The SES factor should also be put into consideration during evaluation.

Analysis of students' responses to 2012 WASSCE mathematics multiple-choice test revealed that the test contains items with significant location DIF. The results showed that the 2012 WASSCE mathematics multiple-choice test measured different things for students from urban and rural schools. This result is in agreement with the result of the study carried out by Odili (2003). Also, Umoinyang (1991) analysed mathematics multiple choice test used by West African Examination Council (WAEC) in the 1990 General certificate of Examination. His study revealed 29 test items that differentially functioned in favour of candidates from educationally developed states. This finding also agrees with Schmitt (1983), who reported that mathematical subtest of SAT showed evidence of location DIF among white population of examinees.

One striking results from this study; is that location had the highest number of DIF items when compared with SES and gender. This also agreed with Odili (2003) findings, where location had more DIF items. There is need for test writers to adopt test-writing procedures to address the incidence of DIF in mathematics test this is so because of the glaring evidence of gender, location, and socio-economic status DIF in test used by public examination bodies.

The Agreement between the Index of DIF for Gender, SES and Location under the methods of IRT and CTT

The result of the study generally showed that the agreement between the index of DIF for gender under the methods of IRT and CTT ranges from 42% to 98%. The CTT (M-H) had 76% to 80% agreement with the IRT methods, as against the CTT (TID) method that had just 42% agreement with the IRT methods. Similarly for SES, the result of the study showed that the agreement between the four methods ranges between 42% and 94%. The CTT (M-H) had 76% to 78% agreement with the IRT methods as against the CTT (TID) method which had 42% to 44% agreement with the IRT methods. Likewise for location the agreement between the four methods ranges from 54% to 94%. The CTT (M-H) had 62% to 64% agreement with the IRT methods as against the CTT (TID) method which had 52% to 54% agreement with the IRT methods. This result agrees with Abedalaziz (2012) study which showed that the strongest agreement among the various methods he used was between the CTT (chi-square) and IRT (b-parameter), while the lowest agreement was between IRT (Area index) and CTT (TID). In addition, Abedalaziz (2010) showed that the lowest agreement was between CTT (TID) and IRT (b-parameter difference). According to Baghi and Ferrara (1989) the chi-square techniques are considered approximate to Item Response Theory techniques. This could be why Mantel-Haenszel, which is one of the chi-square techniques, had a high agreement with the IRT methods. Another reason could be because the M-H method match ability as it is done in IRT methods of detecting DIF. Also, Roever (2005) explained that the TID method does not match test takers by ability and it tends to interpret difficulty and discrimination as DIF. This could be why the TID method flagged almost all the items as DIF items.

Surprisingly, the two IRT methods (IRT-3P and Rasch model) had the highest agreement (98%). This is unlike the two CTT methods (M-H and TID), they had the least agreement (42%). These show that there is a stronger agreement between the IRT methods than the CTT methods. The IRT methods tend to be more reliable than the CTT methods.

The result shows that the agreement between the CTT (M-H) and the IRT methods were found to be significant for gender, SES and location while the agreement between the CTT (TID) and the IRT methods were found to be not significant for gender, SES and location. In addition, the agreement between the two IRT methods was found significant while the agreement between the two CTT methods was found not significant for gender, SES and location. The implication of this result is that there is significant agreement between M-H and the IRT methods but there is no significant agreement between TID and The IRT methods.

In addition, the contingency coefficient yielded the highest agreement value between the two IRT methods, next was that between M-H and the other IRT methods (Rasch and IRT-3P) which yielded moderate and small value respectively. The least contingency coefficient value was between the two CTT methods (TID and M-H). This implies that the IRT methods of detecting DIF are more reliable than the CTT methods of detecting DIF. However, the Mantel-Haenszel method is more reliable than the Transformed item difficulty method.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

Based on the findings of this study, which resulted from the analysed data, the consequent interpretations and discussion on the previous chapters, the following summary, conclusion, and recommendation are hereby presented.

Summary

Differential item functioning test items have been an issue in testing. It can occur in national examinations conducted in a heterogeneous country like Nigeria. This has generated the proliferation of several methods that can be used to detect DIF items in a test. Whether these DIF methods can detect the same test items as DIF item is of much concerned to measurement and evaluation experts. More so that some of these methods of detecting DIF are based on Classical test theory while others are based on Item response theory. Literature has revealed that one of these theory seem to be more advantageous than the other. The CTT is sample and test dependent; this has been a major limitation to CTT.

Differential item functioning (DIF) implies that even after controlling for ability, an item appears to be more difficult for examinees from one group, as compared to examines in the other group. There are several methods of detecting DIF under the CTT and IRT. Whether these different methods from these two theories will be able to detect the same items as DIF items is the crux of this study. The researcher decided to compare the index of DIF for a given sample under the methods of CTT and IRT for candidates with the same mathematics ability from different socio-economic status (SES), location, and gender. Four DIF detection methods were used in this study. Two of these methods are based on CTT (Transformed item difficulty and Mantel-Haenszel) while the remaining two are based on the IRT (Item response theory three parameter and Rasch model).

The four DIF detection methods were used to analyse the responses of 1900 students for gender, SES and location in WASSCE mathematics multiple-choice test. The statistical packages used for the analysis were BILOG-MG, WINSTEPS, SPSS and Microsoft excel. Descriptive statistics were used to answer the research questions. The chi-square test of independence was used to test the hypotheses at $\alpha=0.05$. In other to determine the degree of agreement between the DIF detection methods, the contingency coefficient was used.

Summary of Findings:

The study found that:

1. Gender differential item functioning is present in 2012 WASSCE mathematics multiple-choice test.
2. Socio-economic status differential item functioning is present in 2012 WASSCE mathematics multiple-choice test.
3. Location differential item functioning is present in 2012 WASSCE mathematics multiple-choice test.
4. Location had the highest number of DIF items
5. There is no significant agreement between the Rasch method, 3-parameter method, which are IRT based and the TID method, which is CTT, based for gender, SES, and location.
6. There is a moderate significant agreement between Rasch method and 3-parameter method, which are IRT based and M-H method, which is CTT based for gender and SES.
7. There is a low significant agreement between Rasch method and3-parameter method, which are IRT based and M-H method, which is CTT based for location.
8. The two IRT methods (IRT-3P and Rasch model) had the highest agreement level for gender, SES and location.
9. The two CTT methods (M-H and TID) had the lowest agreement level for gender, location and SES.
10. Gender has the highest agreement level between the CTT and IRT methods of detecting DIF.

Conclusion

Based on the findings of this study, the method used in detecting DIF in a test is very important. An appropriate DIF detection method boosts the processes of selecting items that are actually flagging DIF. The DIF detection methods based on the Item response theory are the best methods that can be used to detect DIF items in a test because it does not depend on p-value but on individual student latent ability. The study indicates that Mantel-Haenszel and transformed item difficulty DIF detection methods are not as effective as the IRT methods in detecting DIF items in a test.

Recommendations

Based on the findings of this study, it is hereby recommended that:

1. Test writers are advised to use the IRT methods of detecting DIF. With a little effort the principles, postulate and assumptions behind IRT models can be easily understood.

2. Effort should be made by all concerned bodies to make IRT differential item functioning detection methods software available.

3. Seminars and workshops should be carried out to aid the proper understanding of IRT differential item functioning detection methods as well as how the IRT based software can be used for data analysis.

4. Examination bodies and even the classroom teachers can use the M-H method if the IRT based methods are not readily available.

5. Test writers should be sensitive to the disparities that exist between rural and urban examinees, low SES and High SES, and male and female.

6. WAEC and other public examination bodies should analyse students' responses to test items for differential functioning before releasing the test for public use.

7. Item writers should be trained on how to identify DIF items and on how to write DIF-free items.

8. A review panel should be set up for national/ state wide examinations to review DIF items

9. Test writers should construct items that are free from writing errors such as offensive, controversial and demeaning terms.

10. Test writers should be sensitive to the heterogeneous nature of Nigeria. They should write mathematics test items that would not unduly favour one group against the other.

Contributions to Knowledge

The study has the following as its contributions to knowledge:

i. Item response theory based methods are more sensitive in detecting differential item functioning than the classical test theory based methods.

ii. The Mantel-Haenszel method of detecting Differential Item Functioning under Classical Test Theory had highest agreement with the Item Response Theory based methods.

iii. The Item Response Theory based methods of detecting Differential Item Functioning yielded comparable results for students in respective of their socio-economic status, gender, and location.

Suggestion for Further Research

The researcher hereby gives suggestions for further study in the following areas:

1. This study used only two states (Delta and Edo). This study should be replicated in other parts of the country, using these DIF detection methods to find out whether it will yield similar result.

2. The efficacy of Item Response Theory DIF methods should be tested with other public examination bodies using their test items.

3. This study can be carried out using test items in tertiary institutions among students of various levels and groups.

4. The study can be carried out using other group of examinees like school type, disabled/abled, geographical region (like northern and southern Nigeria), educationally disadvantaged/ advantaged, different ethnic groups and religious groups in Nigeria.

5. Similarly, other detection methods not used in this study should be used to find out the level of agreement between the methods of detecting DIF based on CTT and IRT

REFERENCES

Abedalaziz, N. (2012). *Exploring DIF: comparison of CTT and IRT methods.* www.ontariointernational.org/icsD20...

Abedalaziz, N. (2011). Detecting DIF using item characteristic curve approaches. *The International Journal of Educational and Psychological Assessment, 8(2), 1-15.*

Abedalaziz, N. (2010). A gender-related differential item functioning of mathematics test items. *The International Journal of Educational and Psychological Assessment, 5.*

Adedoyin, O.O., & Adedoyin, J. A. (2013). Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters. *Herald Journal of Education and General Studies*, 2(3), 107-114.

Adegoke, B. A. (2013). Comparison of item statistics of physics Achievement Test using Classical Test Theory and Item Response Theory Frameworks. *Journal of Education and Practice,* 4(22), 87-96.

Adeyemi, B. E. (2011). A comparative study of students' academic performance in public examinations in secondary schools in Ondo and Ekiti states, *Nigeria. Current Research Journal of economic theory, 3(2), 36-42.*

Ahmad, Z. K., & Nordin, R. (2012). Advance in Educational Measurement: A Rasch model analysis of Mathematics Proficiency Test. International Journal of Social science and Humanity, 2(3), 248-251.

Alordiah, C. O. (2010). *Relative effectiveness of two assessment procedures on Junior secondary school students' achievement in mathematics in Agbor, Delta state.* An unpublished MEd dissertation of Delta state university, Abraka, Delta state, Nigeria.

American Board of Internal Medicine (2012*). Introduction to differential item functioning-item response theory course.* Jtemplin.coe.uga.edu/files/irt/irt07 abim_lecture10.pdf.retriveFeb2012.

Amoo(2007). *Improved skills in the conduct of continuous assessment in schools***.** A seminar paper presented at the Vanguard group of schools on Thursday 8th Agust 2007. Isiwo, Ogun state.

Anastasi, A. (1988*). Psychological testing.* (4th ed.). New York: Macmillan publishing co., Inc.

Andrick, D. (1982). An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttmann scale response pattern. *Education Research and perspectives, 9. 95-104.*

Angoff, W. H. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds). *Differential item functioning* (3-23. Hillsdale, N. J: Erlbaum.

Atar, B. (2006). *Differential Item Functioning analysis for mixed response data using IRT likelihood-ratio test, Logistic regression, and GLLAMN procedures*. Electronic these, Treatises and Dissertations (ETDs). Paper 248 http ://diginole. lib.fsu.edu/etd/248.

Augembery, K. E.,& Morgan, D. L. (2008*). Differential performance of test items by geographical regions.* Paper presented at the annual meeting of the National council on Measurement in Education in New York, NY.

Ayodele, O. J. (2011). Gender difference and performance of secondary school students in mathematics. European Journal of Educational Studies. 3(1), 173-179.

Baghi, H. (1988). *The stability of Item parameter estimates and item characteristic curves across time using different item response models and different estimation procedure*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Baghi, H., & Ferrara, S. (1989*). A comparison of IRT, Delta-plot, and Mantel-Haenszel techniques for detecting Differential item functioning across subpopulations in the Maryland test of citizenship skills.* A paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Baker, F. B (2001). *The Basics of Item Response Theory.*USA: ERIC clearinghouse on assessment and evaluation.

Barry, J. (2005). *The effect of Socio-economic status on academic achievement.* An unpublished MA thesis. Wichita state university. USA.

Blewins, B. E. (2009). *Effects of socioeconomic status on academic performance in Missouri public schools*. An unpublished PhD dissertation. Lindenwood university. USA.

Bolt, D. M. (2002) A monte carlo comparison of parametric and non-parametric polytomous DIF detection methods. *Applied Psychological measurement, 15, 113-141.*

Camilli, G., & Shepard, D. (1994). *Methods for identifying biased test items.* London: Sage publications LTD.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaption of the SIBTEST procedure, Journal of Educational Measurement, 33(3), 333-353.

Chong, H. Y. (2010). *A simple guide to the item response theory (IRT) and Rasch modeling.* http ://www .creative-wisdom .com.

Crane, P. K., Gibbons, L. E., Narasimhalu, K., Lai, J. S.,& Cella, D. (2007). Rapid detection of differential item functioning in assessments of health-related quality of life: The functional assessment of cancer therapy. *Quality of life Research, 16, 101-114.*

Cummings, C. O., Afam, L. E., Eboh, A. B., & Ekaonyewehe, F. (1993). *A model for teaching geometric progression in secondary schools in Ika south local government area of Delta state*. An unpublished B.Sc (Ed) project. University of Nigeria, Nsukka.

De klerk, G. (2008). *Classical Test Theory (CTT).* In M. Born, C. D. Foxcraft, & R. Butter (Eds), online Readings in Testing and Assessment, International Test commission, http ://www.intestcom.org /publications/ORTA.php.

Dorans, N. J., & Holland, P. W. (2012). *Differential Item Functioning*. Hillsdale, N. J : Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. Journal of Educational Measurement, 23(4), 355-368.

Dorans, N. J., & Schmitt, A. P. (1993). *Construction response and differential item functioning: A pragmatic approach* (ETS-RR-91-47) Princeton, NJ. Educational Testing service.

Drasgow, F. (1987). A study of the measurement bias of two standardized psychological tests. *Journal of Applied Psych*ology, 72, 19-29.

Eamon, M. K. (2005). Social-demographic, school, neighborhood, and parenting influences on academic achievement of Latino young adolescents. *Journal of Youth and Adolescence, 34(2), 163-175.*

Eluwa, O. I., Eluwa, A. N., & Abang, B. K. (2011). Evaluation of mathematics achievement test: A comparison between Classical Test Theory (CTT) and Item Response Theory (IRT). *Journal of Educational and Social Research, 1(4), 99-106.*

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologist*. Mahweh, New Jersey: Lawrence Erlbaum Associates.

Evans, G. N. (2004). The environment of childhood poverty. *American Psychologist, 59, 77-92.*

Federal Republic of Nigeria (2004). *National Policy on Education.* Lagos: NERDC press.

Fidalgo, A. M. (2011). A new approach for differential item functioning detection using mantel- Haenszel methods. The GMHDIF program. *The Spanish Journal of Psycholog*y. http:// readperiodicals. com/201107/2551774001.htm.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An advance book*. London and New York: Routeledge.

Geary, D. C. (1996). Sexual selection and sex differences in mathematics abilities. *Behavioural and Brain Science, 19. 229-281.*

Gier, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2002). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. A paper presented at the annual meeting of the National Council on Measurement in Education (NCME) at the symposium entitled, “New Approaches for identifying and interpreting differential bundle functioning” New Orleans, Louisiana. http://www.education. ualberta.ca/educ/psych/crame/

Gittleman, A. (1975). *History of Mathematics*. Columbus: Charles, G. M publishing co.

Gonzalez-Tamayo, E. (1988). Differential Item Functioning, item discrepancy and bias. Available at www.eric.ed.gov/./read.Detail.

Green, D. R., & Draper, J. F. (1972), *Exploratory studies of bias in achievement test.* Paper presented at the Annual Meeting of the American Psychological Association, Horolulu.

Guitton, R. M., & Ironson, G. H. (1983). Latent trait theory for organizational research. *Organizational Behaviour and Human performance, 31, 54-87.*

Guler, N., Uyanik, G. K., & Teker, G. T. (2013). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education*, 2(1), 1-6.

Haiyang, S. (2010). An application of classical test theory and many-facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics, 33(2), 87-102*.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test item: Comparison of IRT area and Mantel-Haenszel methods*. Applied Measurement in Education, 2, 313-334.*

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: principles and applications.* Boston, MA: Kluwer Nijhaff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundermentals of item response theory.* London: SAGE publications Inc.

Harvey, R. J., & Hammer, A. L. (1999). *Item Response Theory*. Virgina Polytecnic Institute & State University Virgina: consulting psychologists press, Inc.

Hays, R. D., Morales, L. S., & Reise, S. P (2000). Item Response Theory and health outcomes measurement in the 21st century. *Medical Care, 38(9 supplements), 28-42.*

Hernandez, R. (2009). Comparison of item discrimination and item difficulty of the Quick-mental Aptitude test using CTT and IRT methods. *The international Journal of Social issues 59(4), 821-840.*

Hochschild, J. L. (2003). Social Class in Public Schools. Journal of Social Issues 59(4), 821-840.

Holland, P. W., & Thayer, D. T. (1988). *An alternative definition of the ETS delta scale of item difficulty*. Educational Testing service, Technical report (85-64)/ research report 25-43.

Hornby, A. S. (2006). Oxford Advance Learner's Dictionary, 7th Edition. London: Oxford University Press.

Hunter, J. E. (1975). *A critical analysis of the use of item means and item test correlations to determine the presence or absence of content bias in achievement test items.* Paper presented at the National Institute of Education conference on Test bias, Annapolis.

Igbokwe, D. I (2003). An assessment of the foundation for a sustainable scientific and technological development in Nigeria. *Journal of Issues on Mathematics, 6(1), 18-30.*

Ija, C. N. (2009). Towards promoting gender equality and women empowerment for sustainable development in River state. *Journal of International Gender studies, 4,225-235.*

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35, 69-81.*

Inomiesa, E. A. (1989). Sex and school location as factor in science and student's achievement. *Journal of Science Teachers' Association of Nigeria, 21(2), 117-125.*

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16(4), 209-223.*

Izard, J. F., & White, J. D. (1980). The use of latent trait models in the development and analysis of classroom test. In Spearitt, D. (Ed). *The improvement of measurement in education and psychology: contribution of latent trait theories. Australian*: Australian council for Educational Research.

Jekami, E. O. (1992). *Issues, problems and prospects of mathematics curriculum development in Nigeria: An NERDC overview.* In National school curriculum review conference proceedings. Lagos: Macmillan.

Jegnes, W. H. (2002). Examining the effects of parental absence on academic achievement of adolescents: the challenge of controlling for family income. Journal of Family and Economic Issues. 23(2).

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. Journal of Educational Measurement, 38(1), 79-93.

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

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14, 117*-138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of Mental test scores*. Reading: Addison-Wesley.

Luppescu, S. (2002). *DIF detection in HLM*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The international Journal of Educational and Psychological Assessment, 1(1), 1*-11.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58(303), 690-700.*

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

Mboto, F. A. & Bassey, S. W. (2004). Attitude and gender in science, technology and mathematics (STM) student's performance. *International Journal of Research in Education 1(1&2), 34-37.*

McNeal, R. B. (2001). Differential effects of parental involvement on cognitive and behavioural outcomes by socioeconomic status. *Journal of Socio-economics, 30(2),171.*

Morales, R. A. (2009). Evaluation of mathematics achievement test: A comparison between CTT and IRT. *The International Journal of Educational and Psychological Assessment, 1(1), 19-26.*

Muniz, J., Hambliton, R. K., & Xmg, D, (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*&2, 115-135.

Nenty, H. J. (2005). From Classical Test Theory (CTT) to Item Response Theory (IRT): An introduction to a desirable transition. Afemikhe & G. T. Adewel (Eds), Issues in Educational measurement and Evaluation in Nigeria in honour of Wole Falayajo (371-384). Ibadan, Nigeria.

Nenty, H. J. (2000). Some factors that influence students' pattern of responses to mathematics examination items. Boleswa Educational Research Journal. 17, 47-58.

Nwankwo, O. C. (2011). A practical Guide to Research writing. Port Harcourt: Pam Unique publishers.

Odeyemi, J. O. (1984). An Evaluation of fielded aspects of elementary school teacher preparation for mathematics teaching. *Unpublished PhD thesis,* University of Ibadan.

Odili, J. N. (2003). Effect of language manipulation on differential item functioning of Biology multiple choice test. *An unpublished PhD thesis.* University of Nigeria Nsukka.

Odili, J. N. (2010). Effect of language manipulation on differential item functioning of test items in Biology in a multicultural setting. *Journal of Educational Assessment in Africa, 4, 268-286.*

Ogretmen, T. (2009). A comparison of the parametric methods based on the item response theory in determining differential item and test functioning. *Education and Science, 34(152), 113-125.*

Ojerinde, D. (2013). Classical Test Theory (CTT) Vs Item Response Theory (IRT): An evaluation of the comparability of item analysis results. A guest lecture presented at the Institute of Education, University of Ibadan on 23rd May.

Ojerinde, D. (2014). Innovations in Assessment: JAMB experience. A keynote address presented at the 16th Annual National conference of the Association of Educational Researchers and Evaluation of Nigeria (ASSEREN) at Calaber on 15th July.

Ojerinde, D., & Ifewulu, B. C. (2012). Item Unidimensionality using 2010 Unified Tertiary Matriculation Examination Mathematics Pre-test. A paper presented at the 2012 international conference of IAEA. Kazastan.

Okafor, P. C. (2007). A case study: Factors contributing to the academic achievement of low-socioeconomic status students in Anambra South country, Anambra state, Nigeria*. An unpublished PhD dissertation*. St John's university, Jamaican, New York.

Okoro, I. F. (2011). Curriculum Implementation: A driving force for achieving gender equity in education. *Journal of International Gender studies (JIGS), 6, 55-64.*

Orlando, M., Sherbourve, C.D., & Thissen, D. (2001). Summed-score linking using Item Response Theory: Application to depression measured. Psychological Assessment, 12(3), 354-359.

Oruluwene, G.W., & Ukwuije, R.P.I. (2009). Application of the two-parameter latent trait model in the development and validation of chemistry achievement test *DELSU Journal of Educational Research and Development, 8(1), 1-4*

Oyedeji, O. A. (1998). Teaching for motivation. Ibadan: Ladeoge publishers.

Pilant, M.S. (2008) *Mathematics Microsoft student (DVD*). Microsoft Corporation.

Progar, S., & Socan, G. (2008). An empirical comparison of item response theory and classical test theory. *Horizons of Psychology, 17(3), 5-24*

Raju, N. S. (2004). *A FORTRAN program for calculating DIF (DTF) [computer software]* Chicago: Illinois institute of Technology.

Reever, B. B. (2012) *An introduction to modern measurement theory*.http.appliedresearchcancer.gov/areas/cogn,retreived April 2012

Robert-Okah, I. (2011). Gender Equity and women empowerment in Nigeria: A paradigm for natural development. *Journal of International Gender Studies (JIGS). 6, 167-179.*

Roever,C.(2005) *"that's not fair" fairness bias, and differential item functioning in language testing*. http:/ /www2.hawaii.edu/~roever/brownbag.pdf

Rogers, H. J., & Swaminathan, H.(1993). A comparison of logistic regression and mantel-Hanszel procedures for detecting differential item functioning. *Applied Psychological Measurements, 17, 105-116*

Roussos, L. A., & Stout, W.F.(1996). Simulation studies of the effect of small sample size and studied item parameters on SIBTEST and Mantel-Hanszel type/error performance. *Journal of Education, Measurement, 33(2), 215-230.*

Rudner,L.A,(2001)ItemResponsetheory.http://echo.edres.org:8080/irt/baker/chapter/.pdf .

Santor, D. A. & Ramsay, J. O. (1998). Progress in the technology of measurement: applications of item response model. *Psychological Assessment, 10, 345-359.*

Santrock, J. W. (2006). *Life span development*. New York: McGraw-Hill.

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement. 16(3), 143-157.*

Schmitt, W. H. (1983). Content Bias in Achievement test. *Journal of Educational measurement.* 16, 143-153.

Schumacher, R. (2005). Test bias and differential item functioning. http://www.applied-measurementassociation.com/white%20paper/TEST% 20BIAS%20AND%DIFFERENTIAL%20ITEM%20FUNCTIONING.

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

Smith, L. L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin, 28(0), 754-763.*

Suark, S. (1999). *EQUATE 99. Computer programme for equating two metrics in item response theory.* University of Illinois IRT Laboratory.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Gray, R. D. (1987). Empirical comparison of selected item detection procedures with bias manipulation. *Journal of Education Measurement, 21(1), 209-223.*

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using Logistic regression procedures. *Journal of Educational Measurement, 27, 361-370.*

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 model. Journal of Educational and Behavioural Statistics, 27(1), 53-75.

Teresi, J. (2004*). Differential item functioning and Health Assessment.* Columbia university stroud center and faculty of medicine, New York state Psychiatric Institute, Research division, Hebrew home for the aged at Riverdale.

Thelk, A. (2008). Detecting Items that function Differently for two – and four – year college students. *Research and Practice in Assessment.* 3, 23-27.

Thissen, D. (1991). *MULTILOGTM User's guide. Multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific software, Inc.

Thissen, D. & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104, 385-395.*

Toit, M. (2003). *IRT from SS1. Biglog. MG. multilog, parscale, test fact.* Scientific soft ware International, Inc.

Udo, E., Uyoata, U. E., Inyon, A. U., & Ekanem, I. E. (2011). Instructional strategies for enhancing gender equity in learning primary science: Cooperative small group instructional mode. *Journal of International Gender Studies (JIGS), 6, 1-10.*

Ugodulunwa, C. A. (2014). Quality Assurance in Research, Assessment and Evaluation in Nigeria. A lead paper presented at 16th National Conference of the Association of Educational Researchers and Evaluators of Nigeria (ASSEREN) at Calaber on 16th July.

Ukwuiji, R. R. I. (2003). *Peanut Educational statistics.* Port Harcourt: Celwil Nigeria Limited.

Umobong, M. E. (2005). Item Response Theory: Introducing objectivity into educational measurement. Afemikhe & G. T. Adewel (Eds*), Issues in Educational Measurement and Evaluation in Nigeria* in honour of Eole Falayajo, 385-397. Ibadan, Nigeria.

Umoinyang, I. E. (2011). Method of detecting Differential Item functioning (DIF) as a consistent error in achievement test. *Journal of the Science Teachers Association of Nigeria (JSTAN).* Stanonline.org/imo2011pdf.

Umoinyang, I. E. (1991). *Item Bias in Mathematics Achievement test*: Unpublished M.Ed Thesis, University of Calaber, Calaber.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18, 15-25.*

Uwadiae, I. (2008). WAEC released result. Saturday Punch September, 27:10.

Van der Flier, H. (1977). Environmental factors and deviant response patterns. In Y. H. Poortingan (Ed), *Basic problems of cross-cultural psychology.* Amsterdam: Swets of Linger, B. U.

Van der Linden, W., & Hambleton, R. K. (1997). Handbook of modern Item Response Theory. Heidelbery: Springer-Verlay.

Verstralen, H., Bechger, T., & Maris, G. (2001). *The combined use of classical test theory and item response theory*: Netherlands: Cito.

Warm, T. A. (1978). *A primer of Item response Theory*. Spingfield, YA: National technical information services, US department of commerce.

Weiss, D. T., & Davidson, M. L. (1981). Test Theory and Methods. In Rosenzweigh, M. H., & Porter, L. W. (Eds*). Annual review of Psychology, 32, 629-651.*

Wentzel, K. R. (1998). Parents' Aspirations for children's Educational Affainments: Relations to parental beliefs and social address variables, Merrill-Palmer Quarterly. Retrieved October 30, 2008 from http ://find. Yalegroup.com.

Wibery, M. (2007). Measuring and detecting Differential Item Functioning in criterion-referenced licensing test. (A theoretic comparison of methods. Educational Measurement series EMNO.60.

Wikipedia (2012). On line Encyclopedia.

Williams, N. J. (2003). *Item and person parameter estimation using hierarchical generalized linear models and polytomous item response theory models*. Unpublished dissertation. University of Texas, Austin.

Wozencraft (1963). Sex comparison of certain abilities. *Journal of Educational Research, 11, 21-27.*

Xitao, F. (1998). Item Response Theory and Classical Test Theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement Journal, 58(3), 357.*

Yen, W. N. (1992). Item Response Theory. In Akin, M. C. (Ed). *Encyclopedia of Educational Research*, 2, 657-667. NY: Maxwell Macmillan International

Yilwa, A. V., & Olarinoye, R. D. (2004) The influence of location, proprietorship, sex and grade level on Junior secondary schools students' performance in the skill of observing. *Nigerian Journal of Curriculum Studies, 11(1), 54-63.*

Young, D. J. (1998). Rural and urban differences in students achievement in science and mathematics: A multilevel Analysis. *School Effectiveness and School Improvement, 9(4). 386-412.*

Zieky, M. (2003). *A DIF Primer*. Educational Testing Service, USA.

Zumbo, B. D. (1999). *A handbook on the theory and methods of modeling as a unitary framework for binary and likert (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National defense. Available: http://www.educ.uba.ca/faculty/zumbo/DIF/index.html.

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in Educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and test bias. *Journal of Educational Research & Policy Studies, 5(1), 1-23.*

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for model-based approach for studying DIF.* Working paper of the Edgeworth laboratory for Quantitative Behavioural Science, University of Northern British Columbia: Prince George, B. C.

Zwick, R., Donogbue, J. R., & Grima, A. (1993). Assessing Differential Item Functioning for performance tests. Journal of Educational Measurement, 30(3), 233-251.

Zwick, R., & Thayer, D. T. (1994). *Evaluation of the magnitude of differential item functioning in polytomous items* (ETS RR-94-13). Princeton, NJ: Educational Testing Service.

APPENDIX A

Delta State University Abraka Delta State
Questionnaire on Students' Socio-Economic Status (SES)

Name____

School____ **Sex**_________________

Please put a tick (√) in the box beside the correct answer.

1. **Your parents have:**
   □not more than four children
   □not more than five children
   □more than six, seven… children
2. **You are your father's:**
   □first or second child
   □third or fourth child
   □fifth, sixth… or last child
3. **My family lives in:**
   □a two-room apartment or less
   □a flat of two/three bedroom
   □a whole house
4. **At home you:**
   □own a room to yourself
   □share a room with one or two others
   □share a room with more than two others
5. **Apart from your parents and their children, how many others live in your house?**
   □nobody
   □one or two other person's
   □three or more other person
6. **At home, you speak English:**

□all the time

□sometimes

□rarely/never

7. **In your home there is:**

□plasma TV and home theater

□a TV and a radio

□a radio set or none of these things

8. **Your parent's own:**

□a car

□more than one car

□do not own a car

9. **You attended:**

□a fee-paying private primary school

□a free UBE primary school

□an expensive fee-paying private primary school

10. **Your parents buy you books to read:**

□often

□sometimes

□rarely/never

11. **Apart from school textbooks there are:**

□About ten books in my home

□About twenty to fifty books in my home

□more than fifty books in my home

12. **At home you speak your native language**

□all the time

□sometimes

□rarely/never

13. **At home your parents buy:**

□no daily newspaper

□one daily newspaper

□more than one daily newspaper

14. **Do your parents encourage you to speak to them:**

□often

□sometimes

□rarely/never

15. **When you speak at home, do your parents/guardian insist that you speak English:**

□often

□sometimes

□rarely/never

16. In your free time, do your parents:

□encourage you to read as much as possible
□sometimes ask you to read
□never mind if you never read

17. Does your father/mother help with your homework?

□often
□sometimes
□rarely/never

18. Your father/guardian attended:

□no school at all
□primary/secondary school
□college of education/polytechnic/university

19. Your mother attended:

□no school at all
□primary/secondary school
□collage of education/polytechnic/university

20. Your father/guardian's pay your school fees and buy your school books promptly:

□Everytime
□Sometime
□Rarely/never

APPENDIX B

Manual Calculation of DIF Using Mental-Haenszel Method

Detection of DIF for Item 1

Matching level A

Group	1 (correct)	0 (wrong)	Total
Reference	1 (A_k)	10 (B_k)	11 (N_{rk})
Focal	3 (C_k)	5 (D_k)	8 (N_{fk})
Total	4 (N_{1k})	15 (N_{0k})	19 (N_k)

$E(A_k) = N_{rk}N_{1k}/N_k$; $E(B_k) = N_{rk}N_{0k}/N_k$; $E(C_k) = N_{fk}N_{1k}/N_k$; $E(D_k) = N_{fk}N_{0k}/N_k$

Group	1 (correct)	0 (wrong)	Total
Reference	$E(A_k)$=2.3	$E(B_k)$=8.7	11 (N_{rk})
Focal	$E(C_k)$=1.7	$E(D_k)$=6.3	8 (N_{fk})
Total	4 (N_{1k})	15 (N_{0k})	19 (N_k)

Matching level B

Group	1 (correct)	0 (wrong)	Total
Reference	1(A_k)	7(B_k)	8(N_{rk})
Focal	6(C_k)	3(D_k)	9(N_{fk})
Total	7(N_{1k})	10(N_{0k})	17(N_k)

Group	1 (correct)	0 (wrong)	Total
Reference	$E(A_k)$=3.3	$E(B_k)$=4.7	8(N_{rk})
Focal	$E(C_k)$=3.7	$E(D_k)$=5.3	9(N_{fk})
Total	7(N_{1k})	10(N_{0k})	17(N_k)

Matching level C

Group	1 (correct)	0 (wrong)	Total
Reference	4(A_k)	7(B_k)	11(N_{rk})
Focal	11(C_k)	2(D_k)	13(N_{fk})
Total	15(N_{1k})	9(N_{0k})	24(N_k)

Group	1 (correct)	0 (wrong)	Total

Reference	$E(A_k)=6.9$	$E(_{Bk})=4.1$	$11(N_{rk})$
Focal	$E(C_k)=8.1$	$E(_{Dk})=4.9$	$13(N_{fk})$
Total	$15(N_{1k})$	$9(N_{ok})$	$24(_{Nk})$

Ho: The odds of getting item correctly across all levels of the matching variable is the same for the focal group and the reference group.

Matching level		O	E	O-E	|O-E|
A	A_k	1	2.3	-1.3	1.3
	B_k	10	8.7	1.3	1.3
	C_k	3	1.7	1.3	1.3
	D_k	5	6.3	-1.3	1.3
B	A_k	1	3.3	-2.3	2.3
	B_k	6	3.7	2.3	2.3
	C_k	7	4.7	2.3	2.3
	D_k	3	5.3	-2.3	2.3
C	A_k	4	6.9	-2.9	2.9
	B_k	11	8.1	2.9	2.9
	C_k	7	4.1	2.9	2.9
	D_k	2	4.9	-2.9	2.9
					Σ=26

$MH\text{-}X^2 = (26-0.5)^2/\Sigma_k var(A_k)$

$Var(A_k) = (N_{rk}N_{fk}N_{1k}N_{0k})/N_k(N_k-1)$

For matching level A $\frac{11x8x4x15}{19(19-1)} = \frac{5280}{342} = 15.43$

For matching level B $\frac{8x9x7x10}{17(17-1)} = 18.53$

For matching level C =34.97

$\Sigma_k\ var(A_k) = 15.43+18.53+34.97 = 67.93$

$MH\text{-}X^2 = (26-0.5)2/68.93 = 9.43$

Degree of freedom (df) = 1 at .05 = 3.84

Since 9.43>3.84

We reject Ho and claimed that the odds of getting an item correct across levels of the matching variable is not the same for the focal group and the reference group.

The common odds-ratio $\alpha_{MH} = \frac{\left(\frac{1X5}{19}\right)+\left(\frac{1X3}{17}\right)+\left(\frac{4X2}{24}\right)}{\left(\frac{10x3}{19}\right)+\left(\frac{7X6}{17}\right)+\left(\frac{7x11}{24}\right)}$

= 0.11

0.11<1

This indicates possible bias against the focal group

$\Delta_{\alpha MH} = -2.35In\ 0.11 = 5.19$

|5.19| > 1.5

Therefore the DIF is large

An example of WINSTEPS DIF Analysis of M-H

Item	Chi^2	Probability	Cumlor
1	12.71	0.01	-.41
2	65.27	0.00	.95
3	3.39	0.07	-.19
4	2.14	0.14	.17

1. An item is said to reveal DIF if its probability is less than 0.05. Hence the items 1 and 2 revealed DIF.
2. Cumlor is the cumulative log-odds ratio in logits. When cumlor is negative, it indicates DIF in favour of focal group and when it is positive. It indicates DIF in favour of reference group. This holds if in the analysis, the focal group comes first before the reference group. The reverse is the case if the reference group come first before the focal group. In the example above the focal group comes first before the reference group. Hence the DIF in item 1 is in favour of the focal group while the DIF in item 2 is in favour of the reference group.
3. DIF categorization since cumlor is in logit are as follows:

$|\text{DIF}| \geq 0.65$ logit, DIF is large
0.42 logits $\leq |\text{DIF}| < 0.65$ logits, DIF is moderate
$|\text{DIF}| < 0.42$ logits, DIF is negligible
Hence the DIF in item 1 is negligible while the DIF in 2 is large.

APPENDIX C

Manual calculation of DIF using scheuneman chi-square procedure

Item 1

Ho: Item do not significantly function differently for reference group and focal group test takers

Matching level A

Group	Correct	
	Observed	Expected
Reference	1	2.3
Focal	3	1.7

$(2.3 – 1)^2/2.3 + (1.7 – 3)^2/1.7$

$0.73 + 0.99 = 1.72$

Matching level B

Group	Correct	
	Oberved	Expected
Reference	1	3.3
Focal	6	3.7

$(3.3 – 1)^2/3.3 + (3.7 – 6)^2/3.7$

$1.60 + 1.43 = 3.03$

Matching level C

Group	Correct	
	Observed	Expected
Reference	4	6.9
Focal	11	8.1

$(6.9 – 4)^2/6.9 + (8.1 – 11)^2/8.1$

$1.22 + 1.04 = 2.26$

$X^2 = 1.72 + 3.03 + 2.26$

$= 7.01$

df = (k-1) (r-1)

= (2-1) (3-1)

= 2

df=2 at 0.05 is 5.99

Since 7.01 > 5.99 we reject Ho and accept that the item significantly function differently for the reference group and focal group test takers.

An Example of Scheuneman's Software DIF Analysis

Item	Scheuneman's Signed χ^2-value	Favoured Group

	Male(M)	**Female(F)**	**Total (±)**	
1	2.426	85.696	88.122*	F
2	11.955	19.744	31.699*	F
3	116.124	0.58	116.704*	M
4	28.541	4.458	32.999*	M
5	17.289	5.086	22.375	-

*critical χ^2-value=27.69; df=13, p<.01

1. The total scheaneman's χ^2-value is equal to the sum of the χ^2-value of each group. Hence for item 1 the total χ^2-value is 2.426 (male) + 85.696 (female) = 88.122; for item 2 it is 11.955+19.744 = 31.699 and so on.
2. Degree of freedom (df) = (k-1) (r-1) where k is the number of groups and r is the number of matching levels. In this case K=2 and r=14. Hence the df=(2-1) (14-1) = 1X13 = 13.
3. The χ^2 table value for df=13 at $p < .01$ is 27.69, for $p <.05$ is 22.36
4. If χ2 –calculated > χ^2 – table, the item shows DIF.
5. For $p < .01$, items 1, 2, 3, and 4 are DIF items.
6. For $p < .05$, items 1, 2, 3, 4, and 5 are DIF items.

APPENDIX D

School of post graduate studies
Delta state university
Abraka
Date: 18th August 2012

The Registrar,
West African Examinations Council
P.M.B 1076
Yaba-Lagos.
Dear Sir,

REQUEST FOR PERMISSION TO USE PAST QUESTION FOR RESEARCH

I am a doctoral student of the above university. As a part of my requirement for the award of PhD in measurement and evaluation, I am carrying out a research on the comparison of index of DIF under the methods of IRT and CTT for mathematics multiple choice questions.

This is to help contribute to knowledge for development of test items that properly discriminate students.

I am requesting for approval to enable me use your mathematics multiple choice "dead" questions for the year SSCE JUNE 2012. The result will be used only for research purpose under supervision.

Thank you for the anticipated favourable response.

Yours faithfully,

Alordiah, Caroline Ochuko
REG. NO. PG/10/11/191906

APPENDIX E

Distribution of public secondary schools and population of SS3 students in Urban and Rural areas in Delta state and Edo state in 2012/2013 session

	State					
	Delta		**Edo**		**Total**	
	No. of sch.	No. of stud. in SS3	No. of sch.	No. of stud. in SS3	No of sch.	No of stud. in SS3
Urban	139	24489	96	17020	235	41509
Rural	310	15469	178	8983	488	24452
Total	449	39958	274	26003	723	65961

Source: Ministry of Education, Asaba, Delta State

Ministry of Education, Benin-city, Edo state

APPENDIX F

A Distribution of the population of SS3 students according to gender

State	Male	Female	Total
Delta	21542	18416	39958
Edo	12646	13357	26993
Total	34188	31773	65961

Source: Ministry of Education, Asaba, Delta State
Ministry of Education, Benin-city, Edo State

APPENDIX G

Sample distribution of SS3 students in the two states

States	**No. of Students**	**Proportion**	**Sample**
Delta	39958	0.606	1152
Edo	26003	0.394	748
Total	65961		1900

APPENDIX H

Sample distribution of SS3 students in the two states according to location.

	Delta (1152)		**Edo** (748)	
Proportion	0.629(urban)	0.371(rural)	0.629(urban)	0.371(rural)
Sample	693	459	451	297

APPENDIX I

Reliability Coefficient of Socio-Economic Status Questionnaire

	VAR00002	**VAR00001**
VAR00002 Pearson correlation	1	.702**
Sig(2-tailed)		.000
N	45	40
VAR00001 Pearson correlation	.702**	1
Sig (2-tailed)	.000	
N	45	40

** correlation is significant at the 0.01 level (2-tailed)

Reliability Coefficient of 2012 WASSCE Mathematics multiple-choice test

	VAR00002	**VAR00001**
VAR00002 Pearson correlation	1	.892**
Sig(2-tailed)		.000

N	45	45
VAR00001 Pearson correlation	.892**	1
Sig (2-tailed)	.000	
N	45	45

** correlation is significant at the 0.01 level (2-tailed)

APPENDIX J

West African Examination Council

West African Senior School Certificate Examination

General Mathematics multiple-choice test 2012 as will be used in this Study

NAME____

SEX____________________

SCHOOL____

INSTRUCTION: Answer all the questions. Circle the letter that bears the right answer

Time: 1½ hours

1. Express 302.10495 correct to five significant figures.
   A. 302.10
   B. 302.11
   C. 302.105
   D. 302.1049

2. Simplify $\frac{3\sqrt{5}\times 4\sqrt{6}}{2\sqrt{2}\times 3\sqrt{3}}$
   A. $\sqrt{2}$
   B. $\sqrt{5}$
   C. $2\sqrt{2}$
   D. $2\sqrt{5}$

3. In 1995, the enrolments of two schools X and Y were 1,050 and 1,190 respectively. Find the ratio of the enrolments of X to Y.
   A. 50: 11
   B. 15: 17
   C. 13: 55
   D. 12: 11

4. Convert 35_{10} to a number in base 2
   A. 1011
   B. 10011
   C. 100011
   D. 11001

5. The n^{th} term of a sequence is $T_n = 5 + (n - 1)2$. Evaluate $T_4 - T_6$.
   A. 30
   B. 16
   C. -16
   D. -30

6. Mr Manu travelled from Accra to Pamfokrom a distance of 720 km in 8 hours. What will be his speed in m/s?
   A. 25 m/s

B. 150 m/s
C. 250 m/s
D. 500 m/s

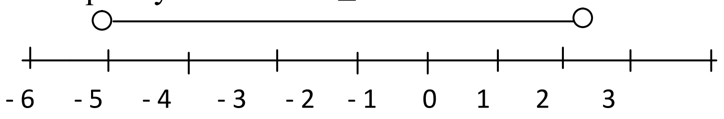
7. If #2,500.00 amounted to #3,500.00 in 4 years at simple interest, find the rate at which the interest was charged.
A. 5%
B. 7½%
C. 8%
D. 10%

8. Solve for x in the equation: $\frac{1}{x} + \frac{2}{3x} = \frac{1}{3}$
A. 5
B. 4
C. 3
D. 1

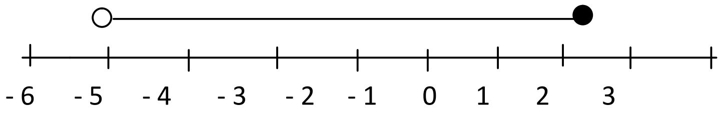
9. Simplify: $\frac{54k2 - 6}{3k + 1}$
A. $6(1- 3k^2)$
B. $6(3k^2 – 1)$
C. $6(3k – 1)$
D. $6(1\ 3k)$

10. Represent the inequality $-7 < 4x + 9 \leq 13$ on a number line.

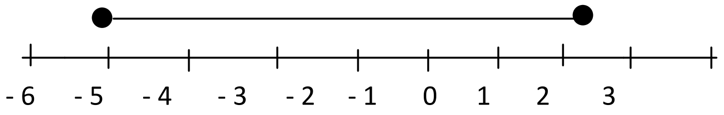
A.



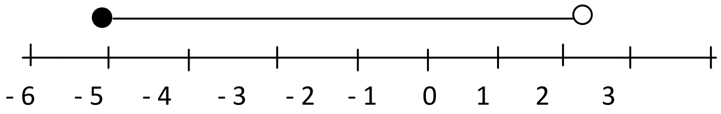
B.



C.



D.



11. Make p the subject of the relation: $q = \frac{3p}{r} + \frac{s}{2}$
A. $p = \frac{2q - rs}{6}$
B. p =2qr – sr -3
C. $p = \frac{2qr - s}{6}$
D. $p = \frac{2qr - rs}{6}$

12. If x + y = 2y – x + 1 = 5, find the value of x.
A. 3
B. 2
C. 1
D. -1

13. The sum of 12 and one third of n is 1 more than twice n. Express the statement in the form of an equation.
A. 12n – 6 = 0
B. 3n – 12 = 0
C. 2n – 35 = 0
D. 5n – 33 = 0

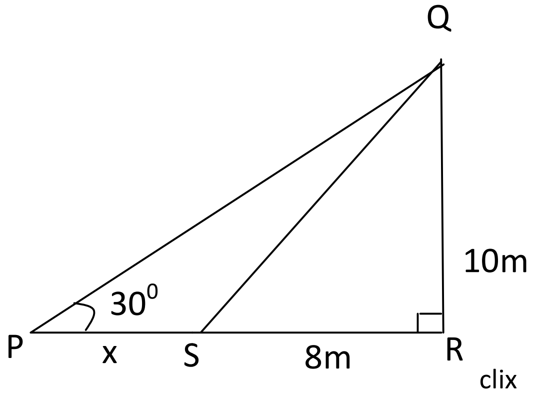
14. Solve the inequality: $\frac{-m}{2} - \frac{5}{4} \leq \frac{5m}{12} - \frac{7}{6}$
A. m ≥ 5/4
B. m ≤5/4
C. m ≥ -1/11
D. m ≤ -1/11

15. The curved surface area of a cylindrical tin is 704 cm^2. If the radius of its base is 8 cm, find the height. [Take π = 22/7].
A. 14 cm
B. 9 cm
C. 8 cm
D. 7 cm

16. The lengths of the minor and major arcs of a circle are 54 cm and 126 ca respectively. Calculate the angle of the major sector.
A. 306°
B. 252°
C. 246°
D. 234°

17. A sector of a circle which subtends 172° at the centre of a circle has a perimeter of 600 cm. Find, correct to the nearest cm, the radius of the circle. [Take π = 22/7].
A. 120 cm
B. 116 cm
C. 107 cm
D 100 cm

18.



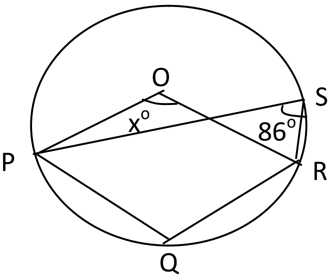
In the diagram, |QR| = 10 m, |SR| = 8 m, < QPS = 30˚, < QRP = 90˚ and |PS| = x.

A. 1.32 m
B. 6.32 m
C. 9.32 m
D. 17.32 m

19. In triangle XYZ, |XY| = 8 cm, |YZ| = 10 cm and |XZ| = 6 cm. Which of these relations is true?

A. |XY| + |YZ| = |XZ|
B. |XY| - |YZ| = |XZ|
C. |XZ|2 = |YZ|2 - |XY|2
D. |YZ|2 = |XZ|2 - |XY|2

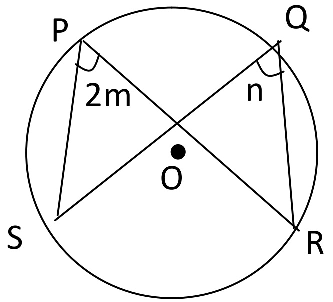
20.



In the diagram, O is the centre of the circle PQRS and <PSR = 86˚. If <POR = x˚, find x

A. 274
B. 172
C. 129
D. 86

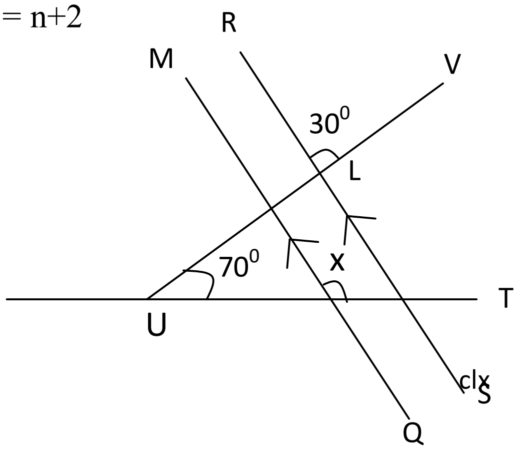
21.



The diagram is a circle centre O. If <SPR = 2m and <SQR = n, express m in terms of n.

A. m = n/2
B. m = 2n
C. m = n-2
D. m = n+2

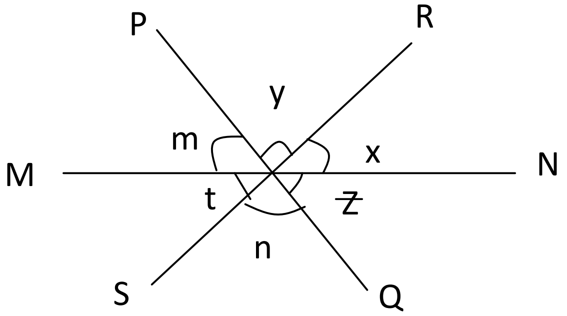
22.



In the diagram, MQ//RS, <TUV = 70° and <RLV = 30°. Find the value of x.

A. 150°
B. 110°
C. 100°
D. 95°

23.



In the diagram, MN, PQ, and RS are three intersecting straight lines. Which of the following statement(s) is/are true?

I. t = y
II. x + y + z + m = 180°
III. x + m + n = 180°
IV. x + n = m +z

A. I and IV
B. II only
C. III only
D. IV only

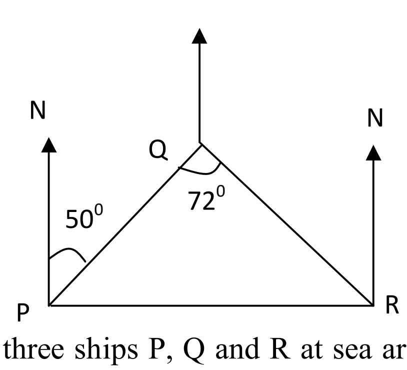
24. If cos (x + 40)° = 0,0872, what is the value of x?

A. 85°
B. 75°
C. 65°
D. 45°

25. A kite flies on a taut string of length 50 m inclined at an angle of 54° to the horizontal ground. The height of the kite above the ground is

A. 50 tan 36°
B. 50 sin 54°
C. 50 tan 54°
D. 50 sin 36°

26.



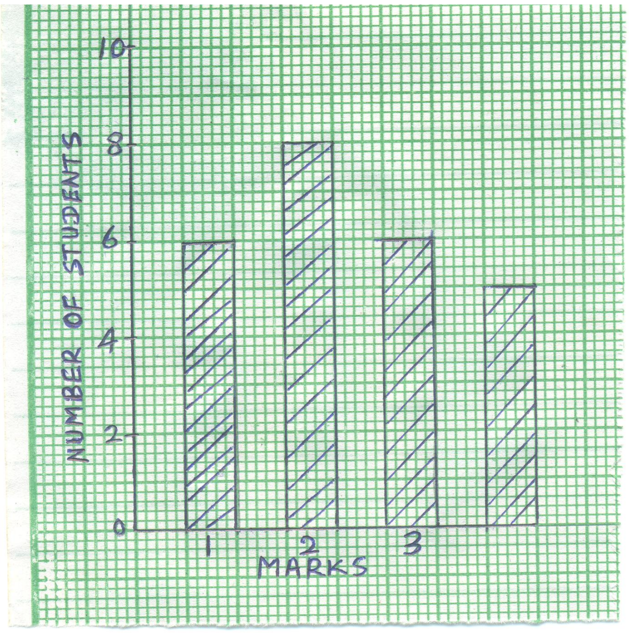
The positions of three ships P, Q and R at sea are illustrated in the diagram. The arrows indicate the North direction. The bearing of Q from P is 050° and <PQR = 72°. Calculate the bearing of R from Q.

A. 130°

B. 158°
C. 222°
D. 252°

27. Given that the mean of the scores 15, 21, 17, 26, 18 and 29 is 21, calculate the standard deviation of the scores.
A. √10
B. 4
C. 5
D. √30

28. A bag contains 4 red and 6 black balls of the same size. If the balls are shuffled briskly and two balls are drawn one after the other without replacement, find the probability of picking balls of different colours.
A. 8/15
B. 13/25
C. 11/15
D. 13/15



The bar chart shows the frequency distribution of marks scored by students in a class test. Use the bar chart to answer questions 29 to 31.

29. How many students are in the class?
A. 10
B. 24
C. 25
D. 30

30. Calculate the mean of the distribution.
A. 6.0
B. 3.0
C. 2.4

D. 1.8

31. What is the median of the distribution?

A. 2
B. 4
C. 6
D. 8

32. Which of these statements about y = 8√m is correct?

A. log y = log 8 x log √m
B. log y = 3 log 2 x ½ log m
C. log y= 3 log 2 – ½ log m
D. log y = 3 log 2 + ½ log m

33. If x+ 0.4y = 3 and y = ½ x, find the value of (x+y).

A. 1 ¼
B. 2 ½
C. 3 ¾
D. 5

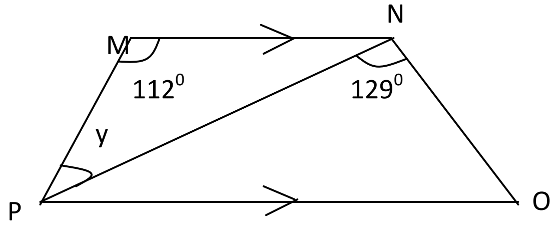
34. Express 3- ($\frac{x-y}{y}$) as a single fraction.

A. $\frac{3xy}{y}$
B. $\frac{x-4y}{y}$
C. $\frac{4y+x}{y}$
D. $\frac{4y-x}{y}$

35. Find the coefficient of m in the expansion of (m/2 – 1 ½) (m + 2/3).

A. – 1/6
B. -½
C. -1
D. -1 1/6

36.



In the diagram, MN//PO, <PMN = 112˚, <PNO = 129, <NOP = 37˚ and <MPN = y. Find the value of y.

A. 51˚
B. 54˚
C. 56˚

D. 68°

37. If P = {prime factors of 210} and Q = {prime numbers less than 10}, find P ∩ Q.

A. {1, 2, 3}
B. {2, 3. 5}
C. {1, 3, 5, 7}
D. {2, 3, 5, 7}

38. Alfred spent ¼ of his money on food, ⅓ on clothing and saved the rest. If he saved #72,000.00, how much did he spend on food?

A. #43,200.00
B. #43,000.00
C. #42,200.00
D. #40,000.00

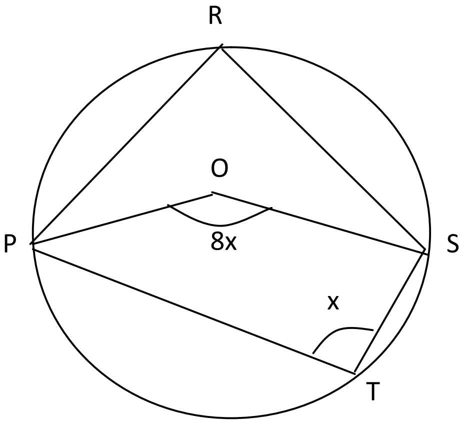
39. Solve: $(27/125)^{-1/3}$ x $(4/9)^{1/2}$

A. 10/9
B. 9/10
C. 2/5
D. 12/125

40. The sum of the interior angles of a regular polygon is 1800°. How many sides has the polygon?

A. 16
B. 12
C. 10
D. 8

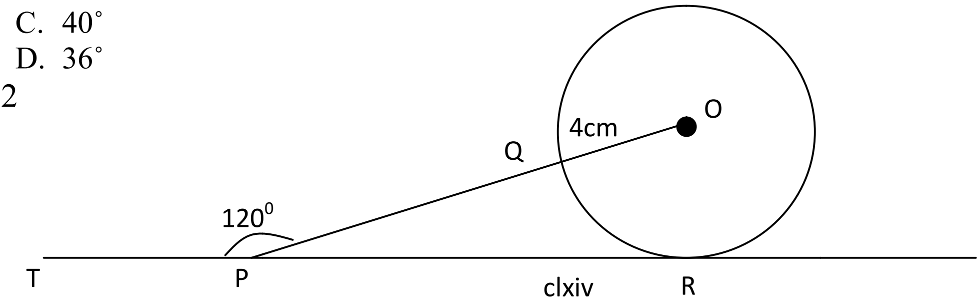
41.



The diagram is a circle with centre O, PRST are points on the circle. Find the value of <PRS.

A. 144°
B. 72°
C. 40°
D. 36°

42



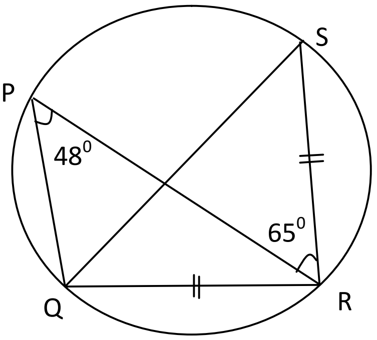
The diagram is a circle of radius |OQ| = $\overline{TR}$ is a tangent to the circle at R. If T$\hat{P}$O = 120°, find |PQ|.

A. 2.32 cm
B. 1.84 cm
C. 0.62 cm
D. 0.26 cm

43. If x and y are variables and k is a constant, which of the following describes an inverse relationship between x and y?

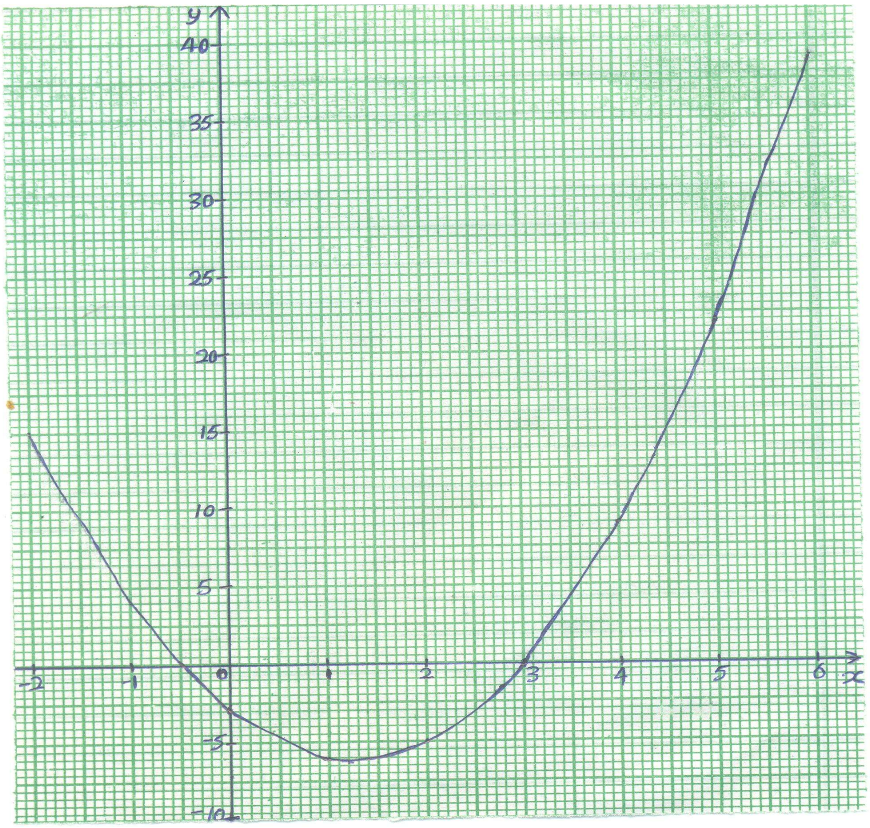
A. y = kx
B. y = k/x
C. y = k√x
D. y = x + k

44.



In the diagram, |SR| = |QR|, <SRP = 65° and <RPQ = 48°, find <PRQ.

A. 65°
B. 45°
C. 25°
D. 19°



The graph is that of y = $2x^2 - 5x - 3$. Use it to answer questions 45 and 46.

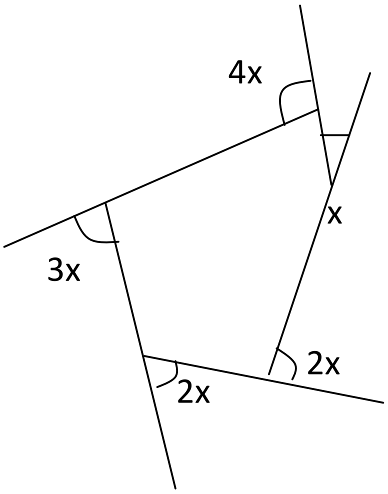
45. For what values of x will y be negative?

A. $-\frac{1}{2} \leq x < 3$
B. $-\frac{1}{2} < x \leq 3$
C. $-\frac{1}{2} < x < 3$
D. $-\frac{1}{2} \leq x \leq 3$

46. What is the gradient of $y = 2x^2 - 5x - 3$ at the point x = 4?

A. 11.1
B. 10.5
C. 10.3
D. 9.9

47.



The diagram is a polygon. Find the largest of its interior angles.

A. 30°
B. 100°
C. 120°
D. 150°

48. The volume of a cuboid is 54 cm ^{3}. If the length, width and height of the cuboid are in the ratio 2 : 1 : 1 respectively, find its total surface area.

A. 108 cm ^{2}
B. 90 cm ^{2}
C. 80 cm ^{2}
D. 75 cm ^{2}

49. A side and a diagonal of a rhombus are 10 cm and 12 cm respectively. Find its area.

A. 20 cm ^{2}
B. 24 cm ^{2}
C. 48 cm ^{2}
D. 96 cm ^{2}

50. Factorise completely: $32x^2y – 48x^3y^3$.

A. $16x^2y\ (2 - 3xy^2)$
B. $8xy\ (4x – 6x^2y^2)$
C. $8x^2y\ (4 – 6xy^2)$
D. $16xy\ (2x – 3x^2y^2)$

APPENDIX K
2012 WAEC WASSCE MATHEMATICS multiple-choice test
MODEL ANSWERS

1	A		26	B
2	D		27	C
3	B		28	A
4	C		29	C
5	C		30	C
6	A		31	A
7	D		32	D
8	A		33	C
9	C		34	D
10	B		35	D
11	D		36	B
12	B		37	D
13	D		38	A
14	C		39	A
15	A		40	B
16	B		41	A
17	A		42	C
18	C		43	B
19	C		44	D
20	B		45	C
21	A		46	A
22	C		47	D
23	C		48	B
24	D		49	D
25	B		50	A

APPENDIX L
LIST OF SCHOOLS USED FOR THE STUDY

DELTA STATE

S/N	NAME OF SCHOOL	LOCATION	L.G.A	MALE	FEMALE	TOTAL
1	Ethiope S/S sapele	urban	sapele	95	65	160
2	Marymount college Owa	urban	Ika north east	_	253	253
3	Osadenis H/S Asaba	urban	Oshimili south	280	_	280
4	Mixed secondary school Abavo	rural	Ika south	170	155	325
5	Elume S/S Elume	rural	Sapele	9	15	24
6	Owhe grammar school otor-owhe	rural	Isoko north	57	53	110
				611	541	1152

EDO STATE

S/N	NAME OF SCHOOL	LOCATION	L.G.A	MALE	FEMALE	TOTAL
1	Ogan M/S/S Ogan	rural	orihionwon	31	12	43
2	Evbotubu S/S Evbotubu	urban	Egor	260	191	451
3	Igbanke Grammar school Oligie Ottah	rural	Orihionwon	49	47	96
4	Ebele S/S Ebele	rural	Igueben	38	48	86
5	Ozalla S/S Ozalla	rural	Owen West	19	53	72
				397	351	748

Demographic Characteristics of the Respondents

VARIABLES		Number	Percentage
Gender	Male	1008	53.05
	Female	892	46.95
Location	Urban	1140	60

	Rural	756	40
SES	High	865	45.53
	Low	1035	54.47
Total		1900	100

As shown in table 8 the total number of respondents consisted of 1900. Out of this figure gender was categorized as male: n = 1008(53.05%), female: n = 892(46.95%); location was categorized as urban: n = 1140(60%) and rural: n = 756(40%); socio-economic status (SES) was categorized as High SES: n = 865(45.53%) and low SES: n = 1035(54.47%).

APPENDIX M

TARO YAMEN'S FORMULA FOR MINIMUM SAMPLE SIZE (Ukwuije, 2003)

$S = N/ (1+Na^2)$

Where S = Sample size
N = Population size
a = Level of significance

$s = 65961/ (1+65961x\ (0.05)^2)$

$=65961/ (1+65961 \times 0.0025)$

$=65961/165.9025$

$=397.589$

APPENDIX N

Preliminary Observation

In any analysis involving IRT, there are two basic assumptions that must be verified, the model fit and unidimensionality.

Unidimensionality

This assumption postulates that, only one ability is measured by the items that make up a test. The unidimensionality assumption is met, if there is a dominant factor that influences the test performance. In this study the confirmatory factor analysis was performed to determine whether or not a dominant factor exists among the 2012 WASSCE mathematics multiple-choice test. This method was also used by Guler, Uyanik & Teker (2013) and Adedoyin and Adedoyin (2013) when they were established unidimensionality assumption for IRT analysis. The factor would represent the construct underlining the mathematics skills measured by the examination. The confirmatory factor analysis yielded 11 eigenvalues greater than one, as shown in table 1

Table 1: Total Variance Explained by the result of Factor Analysis

component	Total	% of variance	Cumulation%
1	9.184	18.368	18.368
2	2.887	5.773	24.368
3	1.937	3.874	28.014
4	1.603	3.207	31.221
5	1.388	2.677	33.898
6	1.235	2.471	36.369
7	1.19	2.38	38.748
8	1.151	2.303	41.051
9	1.103	2.206	43.257
10	1.034	2.068	45.325
11	1.005	2.011	47.336

Extraction method: Principal Component Analysis.

The first eigenvalue was 9.184 greater than the next ten eigenvalue (2.887, 1,937, 1.603, 1.338, 1.235, 1.190, 1.151, 1.103, 1.034, and 1.005). The first factor explained 18.368% of the variance in the data set. The second factor explained 5.773% of the remaining variance. The rest of the variance was explained by the other 48 factors with 9 factors each having percentage of variance between 2 and 3 while 39 factors each have a percentage of variance of between 1 and 2.

The result of the eigenvalue test produced a scree plot as shown in figure 1. The eigenvalue of the first factor was large compared to the second factor, and eigenvalue of the remaining factors are all about the same

Factor Matrix of 2012 WASSCE mathematics multiple-choice test

ITEMS	FACTORS										
	1	2	3	4	5	6	7	8	9	10	11
11	0.632										

19	0.581										
34	0.567										
14	0.564										
13	0.548	-0.31									
9	0.546										
4	0.524	-0.337	-0.358								
6	0.519										
15	0.51										
8	0.51	-0.301									
32	0.51										
10	0.506										
21	0.505										
7	0.504	-0.313									
5	0.5	-0.378									
37	0.488										
39	0.474										
18	0.47										
29	0.466										
16	0.465		0.425								
44	0.461	0.331									
3	0.452		-0.416								
38	0.449	0.322									
12	0.437										
42	0.436										
23	0.431										
28	0.417					0.392					
33	0.412										
45	0.406										
43	0.406										
26	0.4										
27	0.394										
35	0.386										
2	0.348										
50	0.333										
36	0.32										
17	0.31	-0.401	0.43								
49				0.672							
47				0.636							
24				0.365	-0.494						
48					0.412						
46						0.535				0.354	
41						0.447			0.374		-0.343
22	0.3						0.566				
25	0.4							0.417			
20	0.367							0.411		0.35	
30	0.364							-0.351	-0.389		
31	0.342								-0.367		
40	0.328								0.349		
1											0.508

Table 2 shows that almost all the 50 items in the 2012 WASSCE mathematics multiple-choice test were loaded in the first factor. Eight items were loaded in factor 2, four in factor three and so on. Unidimensionality is indicated if the first factor loadings for all the items are significant and have the same sign + or – (McBride & Weiss, 1974 as cited by Ojerinde, 2013). Hence, unidimensionality is indicated. Also according to Orlando, Sherbouve and Thissen (2001) if the first eigenvalue is substantially greater than the next, the factor structure is deemed to have sufficiently satisfied the assumption of unidimentionality. In this study, the eigenvalue of the first factor is substantially greater than that of the other factors. Hence, the assumption of unidimentionality is sufficiently satisfied.

Test for Model Fit

Another assumption of IRT is the correct utilization of models that fits the data. The fitness of the data to the Rasch model and the IRT-three parameter model were examined.

Rasch Model

Level of Item data fit to the Rasch Model

```
|         TOTAL                         MODEL         INFIT        OUTFIT    |
|         SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN    620.7    1900.0        .00     .06       1.00    -.1   1.01    -.1 |
| S.D.    186.8        .0        .58     .01        .12    3.4    .16    2.7 |
| MAX.   1163.0    1900.0       1.48     .08       1.44    9.9   1.64    7.5 |
| MIN.    230.0    1900.0      -1.46     .05        .80   -5.8    .73   -4.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .06 TRUE SD     .58  SEPARATION 10.02  Item   RELIABILITY  .99 |
|MODEL RMSE    .06 TRUE SD     .58  SEPARATION 10.31  Item   RELIABILITY  .99 |
| S.E. OF Item MEAN = .08                                                    |
```

Based on the infit and outfit MNSQ statistics in table 3, both means of infit MNSQ (1.00) and outfit MNSQ (1.01) were almost equal to the value of 1.0 which is the value expected by the Rasch model. This suggests that the amount of distortion of the measurement was minimal. However, the standard deviation of the infit and outfit MNSQ (.12 and .16 respectively were slightly higher than the expected value of 0.1. This showed that the data demonstrated little variation from the Rasch model expectation. The individual items showed that infit MNSQ value ranged from .80 to 1.44 while outfit MNSQ values ranged from .85 to 1.64, which were within the accepted ranged of 0.7 – 1.1 as recommended by Ahmad (2012) for sample more than 1000. However, items 17, 41, 46, and 49 did not fit the model. Generally, the result shows that the scores demonstrated little variation from model expectation. There was evidence of consistency between the 1900 examinees response and 50 items on the scale and the model expectations; therefore, the 2012 WASSCE mathematics multiple-choice test fits the Rasch model.

IRT-Three Parameter Model

To determine whether the items fit the IRT-3P model a chi-square test was run on the data set using Bilog-MG. This is shown in table 4.

Results of chi-square Statistics for IRT-3P Model

Items	Chi-square	Prob.	df	Items	Chi-square	Prob.	df
1	7.9	0.4	7	26	11.4	0.18	7
2	12.4	0.08	7	27	11.3	0.18	7
3	11.3	0.18	7	28	13.5	0.06	7
4	13.5	0.06	7	29	8.3	0.31	7
5	4.2	0.75	7	30	10	0.19	7
6	7.8	0.37	7	31	4.2	0.75	7
7	11.4	0.18	7	32	11.7	0.11	7
8	5.2	0.66	7	33	7.7	0.36	7
9	6.3	0.43	7	34	5.2	0.66	7
10	10	0.19	7	35	10	0.19	7
11	13.6	0.06	7	36	13.6	0.06	7
12	12.4	0.08	7	37	11.3	0.18	7

13	11.4	0.18	7	**38**	11.4	0.18	7
14	7.9	0.4	7	**39**	7.9	0.4	7
15	13.6	0.06	7	**40**	8.2	0.42	7
16	5.2	0.66	7	**41**	5.2	0.66	7
17	4.2	0.75	7	**42**	11.6	0.12	7
18	7.9	0.4	7	**43**	6.3	0.43	7
19	11.6	0.12	7	**44**	5.2	0.66	7
20	10	0.19	7	**45**	13.6	0.06	7
21	6.3	0.43	7	**46**	32	0.01	7
22	11.6	0.12	7	**47**	12.4	0.08	7
23	7.7	0.36	7	**48**	6.3	0.43	7
24	5.2	0.66	7	**49**	31.8	0.01	7
25	6.3	0.43	7	**50**	4.2	0.75	7

Items whose probability is greater than 0.05 significantly fits the IRT-3P model.

The chi-square goodness of fit analysis showed that all the items except 46 and 49 fits the IRT-3P model.

APPENDIX O
TABLES IN CHAPTER TWO

Summary of Selected DIF Methods According To Their Characteristics

Methods	**U/N**	**D/P**	**T/M**	**PA/NPA**	**L/O**	**CTT/IRT**
Point Biserial	U	D	T	NPA	O	CTT
Rasch Model	U	D/P	T/M	PA	L/O	IRT

IRT-2P	U/N	D/P	T/M	PA	L/O	IRT
IRT-3P	U/N	D/P	T/M	PA	L/O	IRT
Discrimination	U	D	T	NPA	O	CTT
Factor Analysis	U	D	T	NPA	O	CTT
M-H	U	D	T/M	NPA	O	CTT
Mantel	U	P	T/M	NPA	O	CTT
GMH	U	P	T/M	NPA	O	CTT
Standardized	U/N	D	M	NPA	O	CTT
SMD	U/N	P	M	NPA	O	CTT
Logistic Regression	U/N	D/P	T/M	PA	O	CTT
SIBTEST	U/N	D	T/M	NPA	L	CTT
Poly- SIBTEST	U/N	P	T/M	NPA	L	CTT
TID	U	D	T/M	NPA	O	CTT
Scheuneman	U	D	T	NPA	O	CTT
ICCs	U/N	D/P	T/M	PA	L	IRT
IRT-Likelihood Test	U/N	D/P	T/M	PA	O/L	IRT
Comparison Method	U/N	D/P	T/M	PA	L	IRT
Lord's Chi-Square Test	U/N	D/P	T	PA	O	CTT
Log Linear Model	U/N	D/P	T/M	PA	O	CTT
Mixed Effect Models	U/N	D/P	T/M	PA	L	IRT
Kamata's Multilevel Rasch	U/N	P	T/M	PA	L	IRT
Parameter Index	U/N	D	M	PA	L	IRT

U – Uniform, N – Non-uniform, D – Dichotomous, P – Polytomous, T – Test DIF, M – Measure DIF, PA – Parametric, NPA – Non parametric, L – Latent, O – Observed, SIBTEST – Simultaneous Item Bias Test, TID – Transformation Item Difficulty, SMT – Standardized Mean Difference, GMH – Generalized Mantel Haenszel, MH – Mantel Haenszel, CTT – Classical Test Theory, IRT – Item Response Theory

Data layout for the M-H method

	Correct Response	**Incorrect Response**	**Total**
Reference Group	A_k	B_k	N_{rk}
Focal Group	C_k	D_k	N_{fk}
Total	N_{1k}	N_{0k}	N_k

Data layout for the Mantel Method

	Item Score					**Total**
	Y_1	Y_2	Y_3	…	Y_m	
Reference Group	N_{1rk}	N_{2rk}	N_{3rk}	…	Y_{mrk}	N_{rk}
Focal group	N_{1fk}	N_{2fk}	N_{3fk}	…	Y_{mfk}	N_{fk}
Total	N_{1k}	N_{2k}	N_{3k}	…	N_{mk}	N_k

Logistic Regression for Binary item

R-square values at each step in the sequential hierarchical regression	DIF X2(2) test	DIF R-square

	Step 1: Total score in the model	Step 2: Total score and Uniform DIF variable in the model	Step 3: Total score, uniform and non uniform DIF variable in the model	
Item 1	0.5625	0.5867	2.505 P=0.2858	0.024
Item 2	0.156	0.3677	69.06 P=0.0000	0.21

An Example of BILOG-MG DIF analysis

Items	b_r	b_f	b-dif	SEb-dif	Z-score
1	-0.47	-0.73	0.26	0.13	2
2	-0.49	-0.31	0.18	0.13	1.38
3	0.78	1.05	-0.27	0.14	-1.93
4	0.28	0.61	-0.33	0.14	-2.36

An Example of WINSTEPS DIF analysis

Item	b_r	b_f	Δ_b	Prob	t
1	-1.35	-1.59	0.22	0.03	2.14
2	-0.95	-0.19	-0.76	0.00	-4.62
3	-0.38	-0.28	-0.10	0.39	-0.89
4	0.14	0.01	0.13	0.25	1.15
5	0.90	-0.36	1.26	0.00	11.2

b_f = measure for focal group, b_r = measure for reference group, $\Delta_b = b_f - b_r$, t= 't' statistic which evaluate the significance of Δ_b.

APPENDIX P
COMPUTER PRINTS OUT

```
DIF-LOCATION 1-RURAL, 2-URBAN
TABLE 30.1 DIF 2                                   ZOU021WS.TXT  May 26 15:50 2014
INPUT: 1900 Person  50 Item  REPORTED: 1900 Person  50 Item  2 CATS WINSTEPS 3.75.0
--------------------------------------------------------------------------------
DIF class specification is: DIF=$S2W1
 ------------------------------------------------------------------------------------------------------------------------------
| Person Obs-Exp    DIF   DIF   Person Obs-Exp   DIF   DIF      DIF    JOINT      Welch      Mantel-Haenszel Size Item        |
| CLASS  Average MEASURE S.E.   CLASS  Average MEASURE S.E.  CONTRAST  S.E.    t  d.f. Prob. Chi-squ Prob. CUMLOR Number  Name |
|------------------------------------------------------------------------------------------------------------------------------|
|------------------------------------------------------------------------------------------------------------------------------|
| 2         .04   -1.68   .07  1        -.06   -1.20   .08      -.48   .10 -4.62  INF .0000 40.4686 .0000   -.69      1 I0001 |
| 2         .05   -1.53   .07  1        -.07    -.94   .08      -.58   .10 -5.65  INF .0000 30.1570 .0000   -.61      2 I0002 |
| 2         .02    -.39   .07  1        -.03    -.13   .09      -.25   .11 -2.30  INF .0215   .9761 .3232   -.13      3 I0003 |
| 2         .08   -1.05   .07  1         -.12    -.03   .09     -1.02   .11 -9.28  INF .0000 65.2687 .0000   -.95       4 I0004 |
| 2         .06    -.95   .07  1         -.09    -.19   .09      -.76   .11 -7.01  INF .0000 43.8729 .0000   -.78       5 I0005 |
| 2         .01    -.38   .07  1         -.01    -.28   .08      -.10   .11  -.89  INF .3728   .1286 .7199    .05       6 I0006 |
| 2         .02    -.30   .07  1         -.03    -.03   .09      -.27   .11 -2.42  INF .0155  4.0020 .0454   -.25       7 I0007 |
| 2         .01    -.13   .07  1         -.01     .01   .09      -.14   .11 -1.20  INF .2297   .2476 .6188    .07       8 I0008 |
| 2         .04    -.30   .07  1         -.06     .27   .10      -.57   .12 -4.90  INF .0000  5.2106 .0224   -.30       9 I0009 |
| 2         .02    -.90   .07  1         -.03    -.67   .08      -.22   .10 -2.16  INF .0306  1.4739 .2247   -.14      10 I0010 |
| 2         .04    -.23   .07  1         -.06     .38   .10      -.61   .12 -5.09  INF .0000  3.0417 .0812   -.25      11 I0011 |
| 2         .01    -.54   .07  1         -.01    -.45   .08      -.10   .11  -.91  INF .3646  4.3744 .0365   -.24      12 I0012 |
| 2         .02     .60   .07  1         -.02     .94   .12      -.35   .14 -2.44  INF .0146  2.4125 .1204   -.25      13 I0013 |
| 2         .00    -.16   .07  1          .00    -.16   .09       .00   .11   .00  INF 1.000  3.2905 .0697    .24      14 I0014 |
| 2         .03    -.92   .07  1         -.04    -.59   .08      -.33   .10 -3.20  INF .0014  4.3929 .0361   -.24      15 I0015 |
| 2         .00    -.22   .07  1          .01    -.27   .08       .05   .11   .48  INF .6287   .6154 .4328   -.10      16 I0016 |
| 2        -.08     .90   .08  1          .12    -.36   .08      1.26   .11 11.20  INF .0000 43.6756 .0000    .84      17 I0017 |
| 2        -.01     .14   .07  1          .01     .01   .09       .13   .11  1.15  INF .2519  4.1302 .0421    .27      18 I0018 |
| 2         .02     .01   .07  1         -.03     .29   .10      -.28   .12 -2.38  INF .0172   .1171 .7322   -.06      19 I0019 |
| 2        -.01    -.23   .07  1          .02    -.39   .08       .16   .11  1.46  INF .1444   .0162 .8988    .02      20 I0020 |
| 2         .02    -.13   .07  1         -.02     .11   .09      -.24   .11 -2.06  INF .0396   .8821 .3476    .13      21 I0021 |
| 2        -.06    -.10   .07  1          .08    -.79   .08       .69   .10  6.65  INF .0000 15.4379 .0001    .45      22 I0022 |
| 2         .01     .04   .07  1         -.01     .18   .09      -.14   .12 -1.20  INF .2306   .0977 .7546   -.05      23 I0023 |
| 2         .01     .17   .07  1         -.01     .31   .10      -.14   .12 -1.14  INF .2564  8.9520 .0028   -.39      24 I0024 |
| 2         .00    -.58   .07  1          .00    -.58   .08       .00   .10   .00  INF 1.000   .0008 .9769    .01      25 I0025 |
| 2        -.01     .04   .07  1          .02    -.13   .09       .18   .11  1.60  INF .1100  1.0320 .3097    .13      26 I0026 |
| 2         .01    -.47   .07  1         -.01    -.39   .08      -.08   .11  -.74  INF .4605  1.0005 .3172   -.12      27 I0027 |
| 2        -.02    -.16   .07  1          .04    -.47   .08       .32   .11  2.98  INF .0029 13.6944 .0002    .45      28 I0028 |
| 2         .01     .36   .07  1         -.01     .51   .10      -.15   .13 -1.21  INF .2256   .0496 .8238    .04      29 I0029 |
| 2        -.02     .20   .07  1          .04    -.15   .09       .35   .11  3.10  INF .0019  2.7128 .0995    .21      30 I0030 |
| 2        -.03     .26   .07  1          .04    -.13   .09       .39   .11  3.53  INF .0004  6.4584 .0110    .33      31 I0031 |
| 2         .00     .31   .07  1          .00     .34   .10      -.03   .12  -.26  INF .7924  4.7424 .0294    .32      32 I0032 |
| 2        -.01     .23   .07  1          .02     .01   .09       .22   .11  1.91  INF .0567  4.5374 .0332    .29      33 I0033 |
| 2         .02     .81   .08  1         -.03    1.24   .13      -.43   .16 -2.80  INF .0052  1.2083 .2717    .25      34 I0034 |
| 2        -.04     .45   .07  1          .06    -.16   .09       .61   .11  5.43  INF .0000 17.7244 .0000    .56      35 I0035 |
| 2        -.04     .17   .07  1          .05    -.32   .08       .49   .11  4.50  INF .0000  9.5848 .0020    .38      36 I0036 |
| 2         .03    -.10   .07  1         -.04     .34   .10      -.43   .12 -3.63  INF .0003  3.8082 .0510   -.27      37 I0037 |
| 2        -.01     .59   .07  1          .02     .38   .10       .21   .12  1.68  INF .0923  7.7093 .0055    .42      38 I0038 |
| 2         .03    -.22   .07  1         -.04     .16   .09      -.38   .12 -3.33  INF .0009  1.3531 .2447   -.16      39 I0039 |
| 2        -.02    -.07   .07  1          .03    -.38   .08       .31   .11  2.87  INF .0042  4.3170 .0377    .25      40 I0040 |
| 2        -.06    1.00   .08  1          .08     .01   .09       .99   .12  8.34  INF .0000   .2367 .6266    .07      41 I0041 |
| 2        -.01     .14   .07  1          .02    -.02   .09       .17   .11  1.48  INF .1389  2.1665 .1410    .20      42 I0042 |
| 2         .03    -.66   .07  1         -.04    -.34   .08      -.32   .11 -3.03  INF .0025  3.8014 .0512   -.23      43 I0043 |
| 2         .01     .64   .07  1         -.02     .87   .12      -.23   .14 -1.65  INF .0999   .4538 .5005   -.12      44 I0044 |
| 2        -.01     .57   .07  1          .02     .34   .10       .23   .12  1.85  INF .0651  2.9388 .0865    .25      45 I0045 |
| 2        -.06    2.07   .10  1          .08     .52   .10      1.54   .15 10.57  INF .0000 11.5426 .0007    .51      46 I0046 |
| 2        -.01    1.35   .09  1          .02     .99   .12       .36   .15  2.41  INF .0162   .5194 .4711   -.13      47 I0047 |
| 2        -.04     .33   .07  1          .05    -.20   .09       .53   .11  4.79  INF .0000  3.8186 .0507    .24      48 I0048 |
| 2        -.03    1.73   .10  1          .04     .95   .12       .78   .15  5.15  INF .0000  2.8796 .0897    .30      49 I0049 |
| 2        -.03     .03   .07  1          .04    -.37   .08       .40   .11  3.72  INF .0002  9.7667 .0018    .38      50 I0050 |
DIF-SEX 1-FEMALE, 2- MALE
TABLE 30.1 DIF                                        ZOU031WS.TXT  May 26 15:49 2014
INPUT: 1900 Person  50 Item  REPORTED: 1900 Person  50 Item  2 CATS WINSTEPS 3.75.0
--------------------------------------------------------------------------------

 DIF class specification is: DIF=$S1W1
 ----------------------------------------------------------------------------------------------------------------------------------
| Person Obs-Exp     DIF   DIF  Person Obs-Exp     DIF   DIF       DIF JOINT       Welch      Mantel-Haenszel Size    Item        |
| CLASS Average MEASURE  S.E.  CLASS  Average MEASURE  S.E.  CONTRAST  S.E.     t d.f. Prob. Chi-squ Prob. CUMLOR  Number  Name |
|---------------------------------------------------------------------------------------------------------------------------------|
| 2        -.02   -1.35   .07  1          .02   -1.57   .07       .22   .10  2.14  INF .0326  3.3890 .0656    .19       1 I0001 |
| 2         .00   -1.29   .07  1          .00   -1.25   .07      -.04   .10  -.41  INF .6854   .0014 .9705    .01       2 I0002 |
| 2         .02    -.38   .07  1         -.02    -.19   .08      -.19   .11 -1.85  INF .0644  1.8493 .1739   -.15       3 I0003 |
| 2         .02    -.77   .07  1         -.03    -.53   .07      -.24   .10 -2.35  INF .0190  3.0593 .0803   -.19       4 I0004 |
| 2         .00    -.65   .07  1          .00    -.65   .07       .00   .10   .00  INF 1.000   .0959 .7568    .04       5 I0005 |
| 2         .02    -.44   .07  1         -.02    -.23   .08      -.21   .10 -2.00  INF .0455  3.1542 .0757   -.20       6 I0006 |
| 2        -.02    -.08   .07  1          .03    -.34   .08       .26   .11  2.49  INF .0130  7.3572 .0067    .30       7 I0007 |
| 2         .01    -.13   .07  1         -.01    -.02   .08      -.11   .11 -1.05  INF .2928   .3600 .5485   -.08       8 I0008 |
| 2         .02    -.20   .07  1         -.02     .00   .08      -.20   .11 -1.82  INF .0682   .9825 .3216   -.12       9 I0009 |
| 2         .03    -.95   .07  1         -.03    -.65   .07      -.31   .10 -3.03  INF .0024  7.5855 .0059   -.29      10 I0010 |
| 2         .01    -.09   .07  1         -.01     .05   .08      -.13   .11 -1.23  INF .2192   .0114 .9151   -.02      11 I0011 |
| 2         .02    -.60   .07  1         -.02    -.40   .07      -.21   .10 -2.03  INF .0424  4.9578 .0260   -.23      12 I0012 |
| 2         .01     .61   .08  1         -.01     .81   .10      -.20   .13 -1.56  INF .1183  1.5000 .2207   -.17      13 I0013 |
| 2         .02    -.26   .07  1         -.02    -.04   .08      -.23   .11 -2.12  INF .0337  1.4844 .2231   -.14      14 I0014 |
| 2         .00    -.77   .07  1          .00    -.81   .07       .04   .10   .42  INF .6749   .4505 .5021    .08      15 I0015 |
| 2         .00    -.24   .07  1          .00    -.24   .08       .00   .11   .00  INF 1.000   .7276 .3937   -.10      16 I0016 |
| 2        -.01     .44   .08  1          .01     .34   .09       .10   .12   .87  INF .3866   .8427 .3586   -.11      17 I0017 |
| 2         .01     .02   .07  1         -.01     .17   .08      -.15   .11 -1.34  INF .1811  1.2578 .2621   -.13      18 I0018 |
| 2         .00     .10   .08  1          .00     .12   .08      -.02   .11  -.19  INF .8460   .0648 .7991    .04      19 I0019 |
| 2        -.01    -.26   .07  1          .01    -.33   .08       .07   .10   .66  INF .5079   .0871 .7679    .04      20 I0020 |
| 2         .00    -.04   .07  1          .00    -.04   .08       .00   .11   .00  INF 1.000  1.3375 .2475    .15      21 I0021 |
| 2        -.02    -.30   .07  1          .02    -.47   .07       .17   .10  1.62  INF .1056   .6521 .4194    .09      22 I0022 |
| 2         .01     .04   .08  1         -.01     .14   .08      -.09   .11  -.84  INF .4037   .2835 .5944   -.07      23 I0023 |
| 2        -.01     .28   .08  1          .01     .15   .08       .13   .11  1.13  INF .2590   .1637 .6858    .05      24 I0024 |
| 2         .01    -.62   .07  1         -.01    -.54   .07      -.08   .10  -.79  INF .4303   .6509 .4198   -.09      25 I0025 |
| 2         .01    -.09   .07  1         -.01     .06   .08      -.15   .11 -1.40  INF .1629  2.1407 .1434   -.17      26 I0026 |
| 2         .01    -.50   .07  1         -.01    -.36   .08      -.14   .10 -1.33  INF .1832  1.9425 .1634   -.15      27 I0027 |
| 2         .01    -.32   .07  1         -.01    -.25   .08      -.07   .10  -.66  INF .5125   .1124 .7375   -.04      28 I0028 |
| 2         .02     .29   .08  1         -.02     .56   .09      -.27   .12 -2.27  INF .0231  3.4177 .0645   -.24      29 I0029 |
| 2        -.02     .16   .08  1          .02    -.04   .08       .20   .11  1.78  INF .0745  2.2107 .1371    .17      30 I0030 |
| 2        -.03     .26   .08  1          .03    -.05   .08       .31   .11  2.80  INF .0052  5.6620 .0173    .27      31 I0031 |
| 2         .01     .26   .08  1         -.01     .36   .09      -.10   .12  -.87  INF .3835   .0026 .9595    .00      32 I0032 |
| 2         .00     .15   .08  1          .00     .15   .08       .00   .11   .00  INF 1.000   .0018 .9661    .01      33 I0033 |
| 2         .00     .92   .09  1          .00     .92   .10       .00   .13   .00  INF 1.000  2.9143 .0878    .28      34 I0034 |
| 2        -.03     .43   .08  1          .04     .00   .08       .43   .11  3.79  INF .0002 12.7056 .0004    .42      35 I0035 |
| 2        -.01     .04   .08  1          .01    -.09   .08       .13   .11  1.21  INF .2278   .3776 .5389    .07      36 I0036 |
| 2         .00     .05   .08  1          .00     .05   .08       .00   .11   .00  INF 1.000   .3735 .5411    .08      37 I0037 |
| 2         .01     .47   .08  1         -.01     .57   .09      -.10   .12  -.80  INF .4251   .1963 .6578   -.07      38 I0038 |
| 2         .02    -.22   .07  1         -.03     .07   .08      -.28   .11 -2.60  INF .0094  4.2589 .0390   -.24      39 I0039 |
| 2         .00    -.19   .07  1          .00    -.19   .08       .00   .11   .00  INF 1.000   .1231 .7257   -.04      40 I0040 |
| 2        -.02     .80   .09  1          .03     .44   .09       .36   .12  2.96  INF .0031   .0009 .9759   -.01      41 I0041 |
| 2        -.02     .20   .08  1          .02    -.05   .08       .25   .11  2.27  INF .0234  5.9604 .0146    .29      42 I0042 |
| 2         .00    -.53   .07  1          .00    -.53   .07       .00   .10   .00  INF 1.000   .5577 .4552    .09      43 I0043 |
| 2        -.01     .80   .09  1          .01     .61   .09       .19   .13  1.50  INF .1342  3.9104 .0480    .28      44 I0044 |
| 2        -.01     .56   .08  1          .01     .40   .09       .17   .12  1.40  INF .1607  1.3575 .2440    .15      45 I0045 |
| 2        -.02    1.74   .11  1          .03    1.19   .11       .55   .15  3.65  INF .0003   .1847 .6673    .07      46 I0046 |
| 2        -.01    1.34   .10  1          .01    1.12   .10       .22   .14  1.52  INF .1277   .0081 .9284    .02      47 I0047 |
| 2        -.01     .21   .08  1          .02     .04   .08       .17   .11  1.51  INF .1325   .4536 .5006    .08      48 I0048 |
| 2        -.01    1.63   .11  1          .01    1.31   .11       .32   .15  2.09  INF .0366   .3593 .5489    .10      49 I0049 |
| 2         .00    -.12   .07  1          .00    -.12   .08       .00   .11   .00  INF 1.000   .0001 .9932    .01      50 I0050 |

MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
 ADJUSTED THRESHOLD VALUES

   ITEM            GROUP          |   ITEM            GROUP
                1          2      |                1          2
----------------------------------+-------------------------------
   ITEM0001 |  -0.493  |  -0.623  |   ITEM0026 |   1.288  |   1.257
            |   0.085* |   0.096* |            |   0.095* |   0.106*
            |          |          |            |          |
   ITEM0002 |  -0.324  |  -0.282  |   ITEM0027 |   0.750  |   0.750
            |   0.086* |   0.100* |            |   0.089* |   0.102*
            |          |          |            |          |
   ITEM0003 |   1.105  |   0.727  |   ITEM0028 |   0.744  |   1.193
            |   0.095* |   0.101* |            |   0.091* |   0.106*
            |          |          |            |          |
   ITEM0004 |   0.482  |   0.445  |   ITEM0029 |   1.938  |   1.708
            |   0.090* |   0.102* |            |   0.108* |   0.117*
            |          |          |            |          |
   ITEM0005 |   0.630  |   0.283  |   ITEM0030 |   1.519  |   1.241
```

```
            |   0.091* |   0.097* |            |   0.097* |   0.101*
            |          |          |            |          |
   ITEM0006 |   0.824  |   0.925  |   ITEM0031 |   1.380  |   1.524
            |   0.092* |   0.104* |            |   0.095* |   0.105*
            |          |          |            |          |
   ITEM0007 |   1.013  |   1.081  |   ITEM0032 |   1.869  |   1.507
            |   0.093* |   0.104* |            |   0.108* |   0.117*
            |          |          |            |          |
   ITEM0008 |   1.260  |   1.137  |   ITEM0033 |   1.557  |   1.414
            |   0.097* |   0.106* |            |   0.099* |   0.108*
            |          |          |            |          |
   ITEM0009 |   1.337  |   0.979  |   ITEM0034 |   2.755  |   2.234
            |   0.102* |   0.107* |            |   0.139* |   0.132*
            |          |          |            |          |
   ITEM0010 |   0.315  |   0.232  |   ITEM0035 |   1.549  |   1.628
            |   0.088* |   0.101* |            |   0.099* |   0.107*
            |          |          |            |          |
   ITEM0011 |   1.482  |   1.034  |   ITEM0036 |   1.138  |   1.456
            |   0.103* |   0.112* |            |   0.092* |   0.104*
            |          |          |            |          |
   ITEM0012 |   0.630  |   0.697  |   ITEM0037 |   1.474  |   1.249
            |   0.088* |   0.096* |            |   0.101* |   0.108*
            |          |          |            |          |
   ITEM0013 |   2.195  |   2.193  |   ITEM0038 |   1.973  |   1.947
            |   0.117* |   0.121* |            |   0.109* |   0.120*
            |          |          |            |          |
   ITEM0014 |   1.138  |   1.058  |   ITEM0039 |   1.233  |   1.145
            |   0.097* |   0.105* |            |   0.098* |   0.109*
            |          |          |            |          |
   ITEM0015 |   0.264  |   0.349  |   ITEM0040 |   1.065  |   1.050
            |   0.089* |   0.101* |            |   0.091* |   0.100*
            |          |          |            |          |
   ITEM0016 |   1.000  |   0.987  |   ITEM0041 |   1.895  |   2.339
            |   0.091* |   0.100* |            |   0.096* |   0.094*
            |          |          |            |          |
   ITEM0017 |   1.416  |   2.296  |   ITEM0042 |   1.534  |   1.265
            |   0.092* |   0.103* |            |   0.100* |   0.107*
            |          |          |            |          |
   ITEM0018 |   1.519  |   1.306  |   ITEM0043 |   0.672  |   0.563
            |   0.101* |   0.104* |            |   0.091* |   0.100*
            |          |          |            |          |
   ITEM0019 |   1.452  |   1.414  |   ITEM0044 |   2.315  |   2.103
            |   0.102* |   0.112* |            |   0.119* |   0.123*
            |          |          |            |          |
   ITEM0020 |   1.046  |   0.795  |   ITEM0045 |   1.956  |   1.891
            |   0.092* |   0.097* |            |   0.105* |   0.114*
            |          |          |            |          |
   ITEM0021 |   1.387  |   1.089  |   ITEM0046 |   2.657  |   3.835
            |   0.101* |   0.111* |            |   0.112* |   0.137*
            |          |          |            |          |
   ITEM0022 |   0.732  |   0.925  |   ITEM0047 |   2.693  |   3.083
            |   0.086* |   0.097* |            |   0.123* |   0.121*
            |          |          |            |          |
   ITEM0023 |   1.445  |   1.381  |   ITEM0048 |   1.380  |   1.567
            |   0.100* |   0.108* |            |   0.094* |   0.103*
            |          |          |            |          |
   ITEM0024 |   1.482  |   1.699  |   ITEM0049 |   2.980  |   3.406
            |   0.098* |   0.098* |            |   0.132* |   0.129*
            |          |          |            |          |
   ITEM0025 |   0.612  |   0.504  |   ITEM0050 |   1.026  |   1.298
            |   0.089* |   0.101* |            |   0.092* |   0.102*
----------------------------------------------------------------------
 *STANDARD ERROR
 MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
 GROUP THRESHOLD DIFFERENCES

   ITEM        GROUP   |   ITEM        GROUP   |   ITEM        GROUP
                2 - 1  |                2 - 1  |                2 - 1
-----------------------+-----------------------+------------------------
   ITEM0001 |  -0.130  |   ITEM0018 |  -0.213  |   ITEM0035 |   0.079
            |   0.129* |            |   0.145* |            |   0.146*
            |          |            |          |            |
   ITEM0002 |   0.043  |   ITEM0019 |  -0.038  |   ITEM0036 |   0.318
            |   0.132* |            |   0.151* |            |   0.138*
            |          |            |          |            |
   ITEM0003 |  -0.377  |   ITEM0020 |  -0.250  |   ITEM0037 |  -0.225
            |   0.139* |            |   0.134* |            |   0.148*
            |          |            |          |            |
   ITEM0004 |  -0.038  |   ITEM0021 |  -0.298  |   ITEM0038 |  -0.026
            |   0.136* |            |   0.150* |            |   0.162*
            |          |            |          |            |
   ITEM0005 |  -0.346  |   ITEM0022 |   0.193  |   ITEM0039 |  -0.088
            |   0.133* |            |   0.129* |            |   0.147*
            |          |            |          |            |
   ITEM0006 |   0.101  |   ITEM0023 |  -0.064  |   ITEM0040 |  -0.016
            |   0.139* |            |   0.147* |            |   0.135*
            |          |            |          |            |
   ITEM0007 |   0.068  |   ITEM0024 |   0.217  |   ITEM0041 |   0.444
            |   0.139* |            |   0.138* |            |   0.135*
            |          |            |          |            |
   ITEM0008 |  -0.124  |   ITEM0025 |  -0.108  |   ITEM0042 |  -0.269
            |   0.144* |            |   0.134* |            |   0.147*
            |          |            |          |            |
   ITEM0009 |  -0.358  |   ITEM0026 |  -0.031  |   ITEM0043 |  -0.109
            |   0.148* |            |   0.142* |            |   0.135*
            |          |            |          |            |
   ITEM0010 |  -0.083  |   ITEM0027 |   0.000  |   ITEM0044 |  -0.212
            |   0.134* |            |   0.136* |            |   0.171*
            |          |            |          |            |
   ITEM0011 |  -0.448  |   ITEM0028 |   0.448  |   ITEM0045 |  -0.065
            |   0.152* |            |   0.139* |            |   0.155*
            |          |            |          |            |
   ITEM0012 |   0.068  |   ITEM0029 |  -0.230  |   ITEM0046 |   1.178
            |   0.131* |            |   0.159* |            |   0.177*
            |          |            |          |            |
   ITEM0013 |  -0.002  |   ITEM0030 |  -0.278  |   ITEM0047 |   0.390
            |   0.168* |            |   0.140* |            |   0.173*
            |          |            |          |            |
   ITEM0014 |  -0.080  |   ITEM0031 |   0.144  |   ITEM0048 |   0.187
            |   0.143* |            |   0.142* |            |   0.140*
            |          |            |          |            |
   ITEM0015 |   0.085  |   ITEM0032 |  -0.362  |   ITEM0049 |   0.426
            |   0.135* |            |   0.159* |            |   0.184*
            |          |            |          |            |
   ITEM0016 |  -0.013  |   ITEM0033 |  -0.143  |   ITEM0050 |   0.272
            |   0.135* |            |   0.146* |            |   0.137*
            |          |            |          |            |
   ITEM0017 |   0.881  |   ITEM0034 |  -0.521  |            |
            |   0.139* |            |   0.192* |            |
------------------------------------------------------------------------
 *STANDARD ERROR
MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
 ADJUSTED THRESHOLD VALUES

   ITEM             GROUP        |   ITEM             GROUP
                 1          2    |                1          2
```

```
----------------------------------+----------------------------------
   ITEM0001 |   0.317  |  -0.688  |   ITEM0026 |   2.346  |   2.622
            |   0.145* |   0.132* |            |   0.169* |   0.138*
            |          |          |            |          |
   ITEM0002 |   0.805  |  -0.399  |   ITEM0027 |   1.863  |   1.640
            |   0.149* |   0.134* |            |   0.161* |   0.131*
            |          |          |            |          |
   ITEM0003 |   2.346  |   1.792  |   ITEM0028 |   1.695  |   2.230
            |   0.172* |   0.132* |            |   0.161* |   0.135*
            |          |          |            |          |
   ITEM0004 |   2.556  |   0.514  |   ITEM0029 |   3.585  |   3.237
            |   0.177* |   0.130* |            |   0.202* |   0.148*
            |          |          |            |          |
   ITEM0005 |   2.245  |   0.715  |   ITEM0030 |   2.317  |   2.918
            |   0.170* |   0.127* |            |   0.168* |   0.136*
            |          |          |            |          |
   ITEM0006 |   2.064  |   1.809  |   ITEM0031 |   2.346  |   3.042
            |   0.166* |   0.134* |            |   0.169* |   0.137*
            |          |          |            |          |
   ITEM0007 |   2.541  |   1.953  |   ITEM0032 |   3.249  |   3.101
            |   0.175* |   0.131* |            |   0.193* |   0.152*
            |          |          |            |          |
   ITEM0008 |   2.618  |   2.292  |   ITEM0033 |   2.618  |   2.975
            |   0.177* |   0.137* |            |   0.176* |   0.141*
            |          |          |            |          |
   ITEM0009 |   3.120  |   1.955  |   ITEM0034 |   4.964  |   4.105
            |   0.186* |   0.142* |            |   0.270* |   0.171*
            |          |          |            |          |
   ITEM0010 |   1.314  |   0.807  |   ITEM0035 |   2.303  |   3.417
            |   0.155* |   0.132* |            |   0.169* |   0.145*
            |          |          |            |          |
   ITEM0011 |   3.325  |   2.092  |   ITEM0036 |   1.996  |   2.871
            |   0.195* |   0.144* |            |   0.162* |   0.135*
            |          |          |            |          |
   ITEM0012 |   1.747  |   1.488  |   ITEM0037 |   3.249  |   2.354
            |   0.158* |   0.126* |            |   0.195* |   0.137*
            |          |          |            |          |
   ITEM0013 |   4.403  |   3.697  |   ITEM0038 |   3.325  |   3.676
            |   0.240* |   0.150* |            |   0.197* |   0.155*
            |          |          |            |          |
   ITEM0014 |   2.332  |   2.213  |   ITEM0039 |   2.911  |   2.109
            |   0.172* |   0.139* |            |   0.184* |   0.138*
            |          |          |            |          |
   ITEM0015 |   1.471  |   0.757  |   ITEM0040 |   1.877  |   2.397
            |   0.158* |   0.131* |            |   0.162* |   0.130*
            |          |          |            |          |
   ITEM0016 |   2.078  |   2.108  |   ITEM0041 |   2.634  |   4.470
            |   0.165* |   0.129* |            |   0.175* |   0.127*
            |          |          |            |          |
   ITEM0017 |   1.929  |   4.278  |   ITEM0042 |   2.556  |   2.816
            |   0.164* |   0.139* |            |   0.173* |   0.143*
            |          |          |            |          |
   ITEM0018 |   2.618  |   2.806  |   ITEM0043 |   1.956  |   1.264
            |   0.176* |   0.139* |            |   0.163* |   0.131*
            |          |          |            |          |
   ITEM0019 |   3.156  |   2.550  |   ITEM0044 |   4.266  |   3.782
            |   0.193* |   0.141* |            |   0.234* |   0.155*
            |          |          |            |          |
   ITEM0020 |   1.863  |   2.091  |   ITEM0045 |   3.249  |   3.633
            |   0.162* |   0.129* |            |   0.192* |   0.147*
            |          |          |            |          |
   ITEM0021 |   2.810  |   2.293  |   ITEM0046 |   3.606  |   6.519
            |   0.181* |   0.144* |            |   0.208* |   0.177*
            |          |          |            |          |
   ITEM0022 |   1.102  |   2.344  |   ITEM0047 |   4.488  |   5.148
            |   0.150* |   0.129* |            |   0.239* |   0.155*
            |          |          |            |          |
   ITEM0023 |   2.945  |   2.613  |   ITEM0048 |   2.217  |   3.177
            |   0.186* |   0.138* |            |   0.167* |   0.137*
            |          |          |            |          |
   ITEM0024 |   3.193  |   2.870  |   ITEM0049 |   4.431  |   5.873
            |   0.193* |   0.124* |            |   0.238* |   0.172*
            |          |          |            |          |
   ITEM0025 |   1.496  |   1.414  |   ITEM0050 |   1.903  |   2.604
            |   0.157* |   0.131* |            |   0.161* |   0.133*
---------------------------------------------------------------------
 *STANDARD ERROR
 MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
 GROUP THRESHOLD DIFFERENCES

   ITEM        GROUP    |   ITEM        GROUP    |   ITEM        GROUP
                2 - 1   |                2 - 1   |                2 - 1
-----------------------+-----------------------+------------------------
   ITEM0001 |  -1.005  |   ITEM0018 |   0.188  |   ITEM0035 |   1.114
            |   0.196* |            |   0.225* |            |   0.223*
            |          |            |          |            |
   ITEM0002 |  -1.204  |   ITEM0019 |  -0.606  |   ITEM0036 |   0.875
            |   0.200* |            |   0.239* |            |   0.211*
            |          |            |          |            |
   ITEM0003 |  -0.555  |   ITEM0020 |   0.227  |   ITEM0037 |  -0.895
            |   0.216* |            |   0.207* |            |   0.238*
            |          |            |          |            |
   ITEM0004 |  -2.042  |   ITEM0021 |  -0.518  |   ITEM0038 |   0.351
            |   0.219* |            |   0.231* |            |   0.250*
            |          |            |          |            |
   ITEM0005 |  -1.531  |   ITEM0022 |   1.242  |   ITEM0039 |  -0.801
            |   0.212* |            |   0.198* |            |   0.230*
            |          |            |          |            |
   ITEM0006 |  -0.255  |   ITEM0023 |  -0.331  |   ITEM0040 |   0.521
            |   0.214* |            |   0.232* |            |   0.208*
            |          |            |          |            |
   ITEM0007 |  -0.587  |   ITEM0024 |  -0.323  |   ITEM0041 |   1.837
            |   0.219* |            |   0.229* |            |   0.217*
            |          |            |          |            |
   ITEM0008 |  -0.326  |   ITEM0025 |  -0.082  |   ITEM0042 |   0.260
            |   0.224* |            |   0.204* |            |   0.225*
            |          |            |          |            |
   ITEM0009 |  -1.166  |   ITEM0026 |   0.276  |   ITEM0043 |  -0.692
            |   0.233* |            |   0.218* |            |   0.209*
            |          |            |          |            |
   ITEM0010 |  -0.507  |   ITEM0027 |  -0.224  |   ITEM0044 |  -0.484
            |   0.203* |            |   0.207* |            |   0.280*
            |          |            |          |            |
   ITEM0011 |  -1.232  |   ITEM0028 |   0.535  |   ITEM0045 |   0.384
            |   0.242* |            |   0.211* |            |   0.241*
            |          |            |          |            |
   ITEM0012 |  -0.258  |   ITEM0029 |  -0.348  |   ITEM0046 |   2.913
            |   0.202* |            |   0.250* |            |   0.273*
            |          |            |          |            |
   ITEM0013 |  -0.706  |   ITEM0030 |   0.601  |   ITEM0047 |   0.660
            |   0.283* |            |   0.217* |            |   0.285*
            |          |            |          |            |
   ITEM0014 |  -0.119  |   ITEM0031 |   0.695  |   ITEM0048 |   0.960
            |   0.221* |            |   0.217* |            |   0.216*
            |          |            |          |            |
   ITEM0015 |  -0.714  |   ITEM0032 |  -0.148  |   ITEM0049 |   1.441
```

```
            |   0.205* |             |   0.246* |             |   0.294*
            |          |             |          |             |
   ITEM0016 |   0.030  |   ITEM0033  |   0.358  |   ITEM0050  |   0.701
            |   0.210* |             |   0.226* |             |   0.209*
            |          |             |          |             |
   ITEM0017 |   2.348  |   ITEM0034  |  -0.859  |             |
            |   0.215* |             |   0.320* |             |
----------------------------------------------------------------------
 *STANDARD ERROR
MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
 ADJUSTED THRESHOLD VALUES

   ITEM              GROUP          |   ITEM             GROUP
                1          2        |                1          2
------------------------------------+---------------------------------
   ITEM0001 |  -0.727  |  -0.470  |   ITEM0026 |  1.369  |  1.152
            |   0.095* |   0.089* |            |  0.107* |  0.097*
            |          |          |            |         |
   ITEM0002 |  -0.312  |  -0.387  |   ITEM0027 |  0.826  |  0.627
            |   0.094* |   0.093* |            |  0.098* |  0.095*
            |          |          |            |         |
   ITEM0003 |   1.050  |   0.778  |   ITEM0028 |  0.974  |  0.864
            |   0.103* |   0.096* |            |  0.100* |  0.098*
            |          |          |            |         |
   ITEM0004 |   0.605  |   0.276  |   ITEM0029 |  2.018  |  1.650
            |   0.099* |   0.095* |            |  0.120* |  0.108*
            |          |          |            |         |
   ITEM0005 |   0.459  |   0.428  |   ITEM0030 |  1.246  |  1.480
            |   0.098* |   0.092* |            |  0.099* |  0.103*
            |          |          |            |         |
   ITEM0006 |   0.996  |   0.706  |   ITEM0031 |  1.230  |  1.611
            |   0.104* |   0.095* |            |  0.101* |  0.101*
            |          |          |            |         |
   ITEM0007 |   0.855  |   1.174  |   ITEM0032 |  1.759  |  1.611
            |   0.101* |   0.098* |            |  0.117* |  0.111*
            |          |          |            |         |
   ITEM0008 |   1.270  |   1.103  |   ITEM0033 |  1.488  |  1.457
            |   0.108* |   0.098* |            |  0.106* |  0.103*
            |          |          |            |         |
   ITEM0009 |   1.295  |   1.020  |   ITEM0034 |  2.474  |  2.484
            |   0.112* |   0.101* |            |  0.141* |  0.133*
            |          |          |            |         |
   ITEM0010 |   0.459  |   0.043  |   ITEM0035 |  1.295  |  1.827
            |   0.097* |   0.093* |            |  0.103* |  0.106*
            |          |          |            |         |
   ITEM0011 |   1.352  |   1.160  |   ITEM0036 |  1.182  |  1.332
            |   0.113* |   0.104* |            |  0.101* |  0.096*
            |          |          |            |         |
   ITEM0012 |   0.782  |   0.492  |   ITEM0037 |  1.352  |  1.346
            |   0.096* |   0.091* |            |  0.107* |  0.105*
            |          |          |            |         |
   ITEM0013 |   2.340  |   2.067  |   ITEM0038 |  2.028  |  1.886
            |   0.131* |   0.112* |            |  0.123* |  0.110*
            |          |          |            |         |
   ITEM0014 |   1.246  |   0.932  |   ITEM0039 |  1.377  |  0.993
            |   0.107* |   0.099* |            |  0.108* |  0.102*
            |          |          |            |         |
   ITEM0015 |   0.250  |   0.282  |   ITEM0040 |  1.057  |  1.014
            |   0.096* |   0.095* |            |  0.099* |  0.094*
            |          |          |            |         |
   ITEM0016 |   1.004  |   0.939  |   ITEM0041 |  1.856  |  2.306
            |   0.100* |   0.093* |            |  0.101* |  0.093*
            |          |          |            |         |
   ITEM0017 |   1.731  |   1.844  |   ITEM0042 |  1.230  |  1.533
            |   0.102* |   0.094* |            |  0.105* |  0.105*
            |          |          |            |         |
   ITEM0018 |   1.514  |   1.302  |   ITEM0043 |  0.584  |  0.601
            |   0.109* |   0.100* |            |  0.099* |  0.094*
            |          |          |            |         |
   ITEM0019 |   1.454  |   1.383  |   ITEM0044 |  2.081  |  2.306
            |   0.109* |   0.107* |            |  0.124* |  0.121*
            |          |          |            |         |
   ITEM0020 |   0.870  |   0.939  |   ITEM0045 |  1.807  |  2.005
            |   0.100* |   0.092* |            |  0.115* |  0.107*
            |          |          |            |         |
   ITEM0021 |   1.238  |   1.216  |   ITEM0046 |  2.829  |  3.523
            |   0.109* |   0.105* |            |  0.126* |  0.120*
            |          |          |            |         |
   ITEM0022 |   0.689  |   0.885  |   ITEM0047 |  2.742  |  3.006
            |   0.094* |   0.091* |            |  0.130* |  0.118*
            |          |          |            |         |
   ITEM0023 |   1.471  |   1.332  |   ITEM0048 |  1.344  |  1.541
            |   0.111* |   0.100* |            |  0.101* |  0.099*
            |          |          |            |         |
   ITEM0024 |   1.488  |   1.634  |   ITEM0049 |  2.982  |  3.376
            |   0.101* |   0.097* |            |  0.139* |  0.125*
            |          |          |            |         |
   ITEM0025 |   0.598  |   0.473  |   ITEM0050 |  1.111  |  1.138
            |   0.098* |   0.093* |            |  0.100* |  0.095*
*STANDARD ERROR
```

TRANSFORMED ITEM DIFFICULTY (TID) - (SES)							
	P-Value		Z-Value		Delta		
ITEMS	H	L	H	L	H	L	Di
1	0.65	0.58	0.39	0.21	14.56	13.84	-1.12
2	0.60	0.55	0.26	0.13	14.04	13.52	-0.87
3	0.44	0.31	-0.15	-0.49	12.40	11.04	-1.09
4	0.49	0.41	-0.01	-0.22	12.96	12.12	-1.45
5	0.51	0.39	0.03	-0.28	13.12	11.88	-1.96
6	0.41	0.36	-0.22	-0.35	12.12	11.60	-0.95
7	0.39	0.33	-0.27	-0.44	11.92	11.24	-0.67
8	0.38	0.29	-0.30	-0.55	11.80	10.80	-1.10
9	0.40	0.28	-0.25	-0.58	12.00	10.68	-1.24
10	0.52	0.44	0.06	-0.15	13.24	12.40	-1.01
11	0.40	0.26	-0.25	-0.64	12.00	10.44	-1.31
12	0.45	0.39	-0.12	-0.28	12.52	11.88	-0.95

13	0.24	0.18	-0.70	-0.91	10.20	9.36	-1.01
14	0.39	0.31	-0.28	-0.49	11.88	11.04	-0.97
15	0.50	0.45	0.00	-0.12	13.00	12.52	-0.79
16	0.40	0.33	-0.25	-0.44	12.00	11.24	-0.96
17	0.23	0.27	-0.73	-0.61	10.08	10.56	0.72
18	0.36	0.26	-0.35	-0.64	11.60	10.44	-1.11
19	0.34	0.26	-0.41	-0.64	11.36	10.44	-1.08
20	0.43	0.32	-0.17	-0.47	12.32	11.12	-1.29
21	0.39	0.27	-0.28	-0.61	11.88	10.56	-1.24
22	0.41	0.37	-0.22	-0.33	12.12	11.68	0.95
23	0.35	0.27	-0.38	-0.61	11.48	10.56	-1.02
24	0.30	0.26	-0.52	-0.64	10.92	10.44	-0.17
25	0.48	0.39	-0.05	-0.27	12.80	11.92	-1.01
26	0.36	0.29	-0.35	-0.55	11.60	10.80	-0.94
27	0.44	0.37	-0.15	-0.33	12.40	11.68	-0.91
28	0.37	0.37	-0.33	-0.33	11.68	11.68	0.04
29	0.30	0.22	-0.52	-0.77	10.92	9.92	-1.03
30	0.37	0.26	-0.33	-0.64	11.68	10.44	-1.03
31	0.33	0.27	-0.44	-0.61	11.24	10.56	-1.01
32	0.33	0.21	-0.44	-0.80	11.24	9.80	-1.65
33	0.34	0.25	-0.41	-0.67	11.36	10.32	-1.29
34	0.24	0.13	-0.70	-1.12	10.20	8.52	-1.86
35	0.31	0.25	-0.49	-0.67	11.04	10.32	-1.08
36	0.34	0.31	-0.41	-0.67	11.36	10.32	-1.17
37	0.37	0.26	-0.33	-0.64	11.68	10.44	-1.17
38	0.27	0.20	-0.61	-0.84	10.56	9.64	-1.10
39	0.38	0.29	-0.30	-0.55	11.80	10.80	-1.13
40	0.39	0.32	-0.28	-0.47	11.88	11.12	-0.69
41	0.23	0.21	-0.73	-0.80	10.08	9.80	-0.19
42	0.36	0.25	-0.35	-0.67	11.60	10.32	-1.31
43	0.47	0.38	-0.07	-0.30	12.72	11.80	-1.14
44	0.25	0.17	-0.61	-0.95	10.56	9.20	-1.23
45	0.28	0.20	-0.58	-0.84	10.68	9.64	-1.07
46	0.10	0.14	-1.28	-1.08	7.88	8.68	1.20
47	0.16	0.13	-0.99	-1.12	9.04	8.52	-0.31
48	0.32	0.27	-0.36	-0.61	11.56	10.56	-1.05
49	0.13	0.11	-1.12	-1.22	8.52	8.12	-0.47
50	0.36	0.32	-0.35	-0.47	11.60	11.12	-0.22

	TRANSFORMED ITEM DIFFICULTY(TID)- (LOCATION)						
	P-Value		Z-Value		Delta		
ITEMS	U	R	U	R	U	R	Di
1	0.71	0.46	0.56	-0.10	15.24	12.60	-1.92
2	0.68	0.41	0.46	-0.22	14.84	12.12	-1.86
3	0.46	0.24	-0.10	-0.70	12.60	10.20	-1.73
4	0.60	0.22	0.26	-0.77	14.04	9.92	-1.98
5	0.57	0.25	0.18	-0.67	13.72	10.32	-1.86
6	0.46	0.27	-0.10	-0.61	12.60	10.56	-1.89
7	0.44	0.23	-0.15	-0.73	12.40	10.08	-1.91
8	0.41	0.22	-0.22	-0.77	12.12	9.92	-1.98
9	0.44	0.18	-0.15	-0.91	12.40	9.36	-1.99
10	0.56	0.35	0.16	-0.36	13.64	11.56	-1.94
11	0.43	0.17	-0.17	-0.95	12.32	9.20	-1.91
12	0.49	0.30	-0.01	-0.52	12.96	10.92	-1.72

13	0.28	0.10	-0.58	-1.28	10.68	7.88	-1.93
14	0.41	0.24	-0.22	-0.70	12.12	10.20	-1.74
15	0.57	0.33	0.18	-0.44	13.72	11.24	-1.82
16	0.42	0.27	-0.20	-0.61	12.20	10.56	-1.87
17	0.23	0.28	-0.73	-0.58	10.08	10.68	0.67
18	0.36	0.22	-0.35	-0.77	11.60	9.92	-1.89
19	0.38	0.18	-0.30	-0.91	11.80	9.36	-1.80
20	0.43	0.29	-0.17	-0.55	12.32	10.80	-1.74
21	0.41	0.20	-0.22	-0.84	12.12	9.64	-1.93
22	0.40	0.37	-0.25	-0.33	12.00	11.68	-1.70
23	0.38	0.19	-0.30	-0.87	11.80	9.52	-1.96
24	0.35	0.17	-0.36	-0.95	11.56	9.20	-1.86
25	0.50	0.33	0.00	-0.44	13.00	11.24	-1.65
26	0.37	0.26	-0.33	-0.70	11.68	10.20	-1.35
27	0.47	0.28	-0.07	-0.58	12.72	10.68	-1.79
28	0.41	0.31	-0.22	-0.49	12.12	11.04	-1.42
29	0.32	0.15	-0.28	-1.03	11.88	8.88	-1.67
30	0.35	0.25	-0.36	-0.67	11.56	10.32	-1.21
31	0.33	0.24	-0.44	-0.70	11.24	10.20	-1.19
32	0.33	0.17	-0.44	-0.95	11.24	9.20	-1.58
33	0.34	0.22	-0.41	-0.77	11.36	9.92	-1.73
34	0.25	0.08	-0.67	-1.40	10.32	7.40	-1.86
35	0.30	0.25	-0.52	-0.67	10.92	10.32	-0.46
36	0.35	0.28	-0.36	-0.58	11.56	10.68	-1.16
37	0.61	0.17	-0.28	-0.95	11.88	9.20	-1.63
38	0.28	0.17	-0.58	-0.95	10.68	9.20	-1.35
39	0.42	0.20	-0.20	-0.84	12.20	9.64	-1.94
40	0.40	0.29	-0.25	-0.55	12.00	10.80	-1.49
41	0.22	0.22	-0.77	-0.77	9.92	9.92	0.45
42	0.36	0.22	-0.35	-0.77	11.60	9.92	-1.77
43	0.51	0.28	0.03	-0.58	13.12	10.68	-1.67
44	0.27	0.11	-0.61	-1.22	10.56	8.12	-1.59
45	0.28	0.17	-0.58	-0.95	10.68	9.20	-1.23
46	0.10	0.45	-1.28	-0.12	7.88	12.52	1.97
47	0.17	0.10	-0.95	-1.28	9.20	7.88	-1.45
48	0.32	0.26	-0.28	-0.64	11.88	10.44	-1.48
49	0.13	0.10	-1.12	-1.28	8.52	7.88	0.98
50	0.38	0.29	-0.30	-0.55	11.80	10.80	-1.03

TRANSFORMED ITEM DIFFICULTY (GENDER)							
	P-Value		Z-Value		Delta		
ITEMS	M	F	M	F	M	F	Di
1	0.61	0.62	0.28	0.31	14.12	14.24	0.09
2	0.60	0.55	0.26	0.13	14.04	13.52	-1.26
3	0.41	0.32	-0.22	-0.47	12.12	11.12	-1.09
4	0.49	0.39	-0.02	-0.28	12.92	11.88	-1.08
5	0.47	0.42	-0.07	-0.20	12.72	12.20	-1.01
6	0.43	0.33	-0.17	-0.44	12.32	11.24	-1.07
7	0.36	0.35	-0.35	-0.36	11.60	11.56	-0.61
8	0.37	0.29	-0.33	-0.55	11.68	10.80	-1.12

9	0.38	0.29	−0.30	−0.55	11.80	10.80	−1.13
10	0.53	0.42	−0.08	−0.20	12.68	12.20	−1.09
11	0.36	0.28	−0.35	−0.58	11.60	10.68	−1.05
12	0.46	0.36	−0.10	−0.35	12.60	11.60	−1.13
13	0.24	0.17	−0.70	−0.95	10.20	9.20	−1.14
14	0.39	0.29	−0.28	−0.56	11.88	10.76	−1.20
15	0.49	0.45	−0.02	−0.12	12.92	12.52	−0.76
16	0.39	0.33	−0.28	−0.44	11.88	11.24	−1.22
17	0.27	0.23	−0.61	−0.73	10.56	10.08	−0.70
18	0.34	0.26	−0.41	−0.64	11.36	10.44	−1.02
19	0.33	0.27	−0.44	−0.61	11.24	10.56	−1.02
20	0.39	0.35	−0.28	−0.36	11.88	11.56	−0.44
21	0.35	0.30	−0.36	−0.52	11.56	10.92	−1.03
22	0.40	0.38	−0.25	−0.30	12.00	11.80	−0.43
23	0.34	0.27	−0.41	−0.61	11.36	10.56	−1.02
24	0.30	0.26	−0.52	−0.64	10.92	10.44	−0.58
25	0.46	0.39	−0.58	−0.28	10.68	11.88	−1.10
26	0.36	0.28	−0.35	−0.58	11.60	10.68	−1.13
27	0.44	0.36	−0.15	−0.35	12.40	11.60	−1.15
28	0.40	0.34	−0.25	−0.41	12.00	11.36	−1.01
29	0.29	0.20	−0.55	−0.84	10.80	9.64	−1.21
30	0.32	0.30	−0.47	−0.52	11.12	10.92	−0.32
31	0.30	0.30	−0.52	−0.52	10.92	10.92	0.45
32	0.30	0.23	−0.52	−0.73	10.92	10.08	−1.01
33	0.32	0.26	−0.47	−0.64	11.12	10.44	−1.07
34	0.20	0.16	−0.84	−0.99	9.64	9.04	−0.63
35	0.27	0.29	−0.61	−0.55	10.56	10.80	0.12
36	0.34	0.30	−0.41	−0.52	11.36	10.92	−0.71
37	0.33	0.28	−0.44	−0.58	11.24	10.68	−1.02
38	0.26	0.20	−0.64	−0.84	10.44	9.64	−1.10
39	0.38	0.28	−0.30	−0.58	11.80	10.68	−0.32
40	0.38	0.32	−0.30	−0.47	11.80	11.12	1.04
41	0.22	0.22	−0.77	−0.77	9.92	9.92	0.45
42	0.31	0.30	−0.49	−0.52	11.04	10.92	0.32
43	0.44	0.40	−0.15	−0.25	12.40	12.00	0.65
44	0.25	0.19	−0.61	−0.87	10.56	9.52	0.75
45	0.25	0.22	−0.61	−0.77	10.56	9.92	0.86
46	0.12	0.13	−1.17	−1.12	8.32	8.52	0.51
47	0.15	0.13	−1.03	−1.12	8.88	8.52	0.69
48	0.31	0.28	−0.49	−0.58	11.04	10.68	0.72

49	0.13	0.12	-1.12	-1.17	8.52	8.32	0.68
50	0.36	0.32	-0.35	-0.47	11.60	11.12	0.67

Crosstabs
MHL * TIDL

Crosstab

			TIDL		
			NO DIF	DIF	Total
MHL	NO DIF	Count	2	23	25
		Expected Count	2.0	23.0	25.0
		% within MHL	8.0%	92.0%	100.0%
		Residual	.0	.0	
	DIF	Count	2	23	25
		Expected Count	2.0	23.0	25.0
		% within MHL	8.0%	92.0%	100.0%
		Residual	.0	.0	
Total		Count	4	46	50
		Expected Count	4.0	46.0	50.0
		% within MHL	8.0%	92.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.000[a]	1	1.000		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.000	1	1.000		
Fisher's Exact Test				1.000	.695
Linear-by-Linear Association	.000	1	1.000		
N of Valid Cases	50				

MHL * IRT3PL

Crosstab

			IRT3PL		
			NO DIF	DIF	Total
MHL	NO DIF	Count	13	12	25
		Expected Count	10.0	15.0	25.0
		% within MHL	52.0%	48.0%	100.0%
		Residual	3.0	-3.0	
	DIF	Count	7	18	25
		Expected Count	10.0	15.0	25.0
		% within MHL	28.0%	72.0%	100.0%
		Residual	-3.0	3.0	
Total		Count	20	30	50
		Expected Count	20.0	30.0	50.0
		% within MHL	40.0%	60.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.000[a]	1	.083		
Continuity Correction[b]	2.083	1	.149		
Likelihood Ratio	3.036	1	.081		
Fisher's Exact Test				.148	.074
Linear-by-Linear Association	2.940	1	.086		
N of Valid Cases	50				

RASCHL * TIDL

Crosstab

			TIDL		Total
			NO DIF	DIF	
RASCHL	NO DIF	Count	0	19	19
		Expected Count	1.5	17.5	19.0
		% within RASCHL	.0%	100.0%	100.0%
		Residual	-1.5	1.5	
	DIF	Count	4	27	31
		Expected Count	2.5	28.5	31.0
		% within RASCHL	12.9%	87.1%	100.0%
		Residual	1.5	-1.5	
Total		Count	4	46	50
		Expected Count	4.0	46.0	50.0
		% within RASCHL	8.0%	92.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.665[a]	1	.103		
Continuity Correction[b]	1.200	1	.273		
Likelihood Ratio	4.035	1	.045		
Fisher's Exact Test				.284	.137
Linear-by-Linear Association	2.612	1	.106		
N of Valid Cases	50				

RASCHL * IRT3PL

Crosstab

			IRT3PL		Total
			NO DIF	DIF	
RASCHL	NO DIF	Count	18	1	19
		Expected Count	7.6	11.4	19.0
		% within RASCHL	94.7%	5.3%	100.0%
		Residual	10.4	-10.4	
	DIF	Count	2	29	31
		Expected Count	12.4	18.6	31.0
		% within RASCHL	6.5%	93.5%	100.0%

	Residual	-10.4	10.4	
Total	Count	20	30	50
	Expected Count	20.0	30.0	50.0
	% within RASCHL	40.0%	60.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	38.257[a]	1	.000		
Continuity Correction[b]	34.667	1	.000		
Likelihood Ratio	44.634	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	37.492	1	.000		
N of Valid Cases	50				

RASCHG * MHG

Crosstab

			MHG		
			NO DIF	DIF	Total
RASCHG	NO DIF	Count	34	2	36
		Expected Count	30.2	5.8	36.0
		% within RASCHG	94.4%	5.6%	100.0%
		Residual	3.8	-3.8	
	DIF	Count	8	6	14
		Expected Count	11.8	2.2	14.0
		% within RASCHG	57.1%	42.9%	100.0%
		Residual	-3.8	3.8	
Total		Count	42	8	50
		Expected Count	42.0	8.0	50.0
		% within RASCHG	84.0%	16.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	10.436[a]	1	.001		
Continuity Correction[b]	7.845	1	.005		
Likelihood Ratio	9.397	1	.002		
Fisher's Exact Test				.004	.004
Linear-by-Linear Association	10.227	1	.001		
N of Valid Cases	50				

RASCHG * TIDG

Crosstab

			TIDG		
			NO DIF	DIF	Total
RASCHG	NO DIF	Count	15	21	36
		Expected Count	16.6	19.4	36.0

		% within RASCHG	41.7%	58.3%	100.0%
		Residual	-1.6	1.6	
	DIF	Count	8	6	14
		Expected Count	6.4	7.6	14.0
		% within RASCHG	57.1%	42.9%	100.0%
		Residual	1.6	-1.6	
Total		Count	23	27	50
		Expected Count	23.0	27.0	50.0
		% within RASCHG	46.0%	54.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.972[a]	1	.324		
Continuity Correction[b]	.449	1	.503		
Likelihood Ratio	.971	1	.324		
Fisher's Exact Test				.361	.251
Linear-by-Linear Association	.952	1	.329		
N of Valid Cases	50				

RASCHL * MHL

Crosstab

			MHL		Total
			NO DIF	DIF	
RASCHL	NO DIF	Count	13	6	19
		Expected Count	9.5	9.5	19.0
		% within RASCHL	68.4%	31.6%	100.0%
		Residual	3.5	-3.5	
	DIF	Count	12	19	31
		Expected Count	15.5	15.5	31.0
		% within RASCHL	38.7%	61.3%	100.0%
		Residual	-3.5	3.5	
Total		Count	25	25	50
		Expected Count	25.0	25.0	50.0
		% within RASCHL	50.0%	50.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.160[a]	1	.041		
Continuity Correction[b]	3.056	1	.080		
Likelihood Ratio	4.235	1	.040		
Fisher's Exact Test				.079	.040
Linear-by-Linear Association	4.076	1	.043		
N of Valid Cases	50				

RASCHL * TIDL

Crosstab

			TIDL		
			NO DIF	DIF	Total
RASCHL	NO DIF	Count	0	19	19
		Expected Count	1.5	17.5	19.0
		% within RASCHL	.0%	100.0%	100.0%
		Residual	-1.5	1.5	
	DIF	Count	4	27	31
		Expected Count	2.5	28.5	31.0
		% within RASCHL	12.9%	87.1%	100.0%
		Residual	1.5	-1.5	
Total		Count	4	46	50
		Expected Count	4.0	46.0	50.0
		% within RASCHL	8.0%	92.0%	100.0%