# EFFECTIVENESS OF MANTEL-HAENSZEL AND LOGISTIC REGRESSION STATISTICS IN DETECTING DIFFERENTIAL ITEM FUNCTIONING UNDER DIFFERENT CONDITIONS

**BY**

**UKANDA, FERDINAND INGUBU**

**A THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN EDUCATIONAL PSYCHOLOGY**

**DEPARTMENT OF EDUCATIONAL PSYCHOLOGY**

**MASENO UNIVERSITY**

## DECLARATION

This thesis is my original work and has not been presented for a degree in any other University

Signature…… Date …….

Ferdinand Ingubu Ukanda
PG/PHD/093/2010

This thesis has been submitted for examination with our approval as University Supervisors

Signature…… Date …….

Prof Lucas Othuon
Department of Educational Psychology
Maseno University

Signature…… Date …….

Prof John Agak
Department of Educational Psychology
Maseno University

Signature…… Date …….

Prof Paul Oleche
Department of Pure and Applied Mathematics
Maseno University

## ACKNOWLEDGEMENT

I am grateful to the Almighty God for His favors that enabled me undertake this doctoral program. I am indebted to the mentoring and continued assistance from my supervisors, Professor Lucas A Othuon, Professor John Agak and Professor Paul Oleche who not only introduced me to psychometrics, but also literally guided my steps in the area of Differential Item Functioning analysis and research. Their support enabled me to generate data using statistical software and draw statistical graphs. Much gratitude to my colleague Julius Okoth who guided me in the writing of a routine for Mantel-Haenszel analysis and Logistic Regression using SPSS. I owe special thanks to Maseno University for admitting me to the doctoral program and for all the support along the way.

I am grateful to my family, first to my wife Christine for her perseverance and unwavering emotional support throughout my pursuit of a doctoral degree, and for all the encouragement when the days seemed dark and the completion of the task was out of imaginable time. I owe special thanks to my children: Ian Shitsimi and Nigel Winjila .Thank you all for your love and understanding. You all inspired me throughout my academic undertakings. I extend much appreciation to my other family members including my brother Ponventra Anjimbi and my sister in law Jane. My other brothers; Constantine, Billy, Brian, and Pascal and my sisters Isabellah and Mildred and my mothers Clare and Joyce. Your prayers gave me a push each day to “strive towards excellence”. Last but not least, I would like to thank Bro Denis Abok and Mr Peter Obwogo for the support they accorded me as I undertook my doctoral program. To all I say thank you and God bless you abundantly.

## DEDICATION

This thesis is dedicated to my late sister Doreen Lihabi Mwashi for her encouragement in the course of my study but who did not see me achieve this meticulous dream. We loved you but God loved you more.

## ABSTRACT

Differential Item Functioning (DIF) is the different probability of responding to a test item by individuals with the same ability level, but from two different groups. The groups may be based on gender, race or disability. DIF can be detected by methods such as Mantel-Haenszel (MH) and Logistic Regression (LR) which classify DIF items into negligible, moderate, and large DIF. Conditions such as Sample size, Ability distribution and Test length may have a significant effect on DIF detection. A conceptual measurement model indicated that person achievement is made observable through a set of items and the items vary in their locations on the latent variable. The purpose of this study was to determine the effect of different conditions on the detection of DIF using MH and LR statistics. The objectives of the study were; to determine the effect of different conditions on the Effect size and the number of DIF detections; and to compare the effect of different conditions on the number of DIF detections using MH and LR statistics. A Factorial research design was used in the study. The independent variables were Sample size, Ability distribution and Test length. The dependent variables were the Effect sizes and the number of DIF detections. The population of the study was 2000. A stratified random sampling technique was used with the stratifying criteria as the reference and focal groups with sample sizes 20, 60 and 1000. This was based on the examinee numbers in a classroom or in a school. WinGen3 software was used to generate dichotomous data with 1000 replications so as to reduce the sampling variance. Two Ability distribution conditions were established with tests of 10, 30 and 50 items selected according to the number of items often observed on personality inventories and achievement tests. A pilot study was conducted. Face validity was obtained by experts and a reliability coefficient of 0.75 was obtained using Kuder-Richardson method. ANOVA was used for analysis at a level of significance of .05. Line graphs also aided interpretation. The findings of the study showed that sample size had a significant effect on the Effect size for Type B items using MH and Type A items using LR statistic. Ability distribution had a significant effect on Type C items using MH but no effect using LR statistic. Test Length had no effect on all DIF types. Ability distribution contributed to the number of DIF items of all kinds detected using both statistics. MH statistic detected more Type C items than LR statistic. The LR statistic detected more Type A and B items than the MH statistic. It was therefore concluded that the effect of Sample size and Ability distribution depended on the DIF statistic used. Test length had no significant effect on Effect size using both statistics. Also the number of DIF items detected depended on the Ability distribution. MH detected more Type C items than Type A and B items while LR detected more Type A and B items than Type C items. It was recommended that test developers use MH statistic when detecting Type C items, and LR for Type A and B items. The findings may be used by test developers to determine the items to be included in a test or those to be omitted to ensure that a test presented to the examinees is free of bias.

## TABLE OF CONTENTS

## LIST OF ABBREVIATIONS

| | |
|---|---|
| ANOVA | Analysis of Variance |
| CSV | Comma Separated Values |
| CTT | Classical Test Theory |
| DIF | Differential Item Functioning |
| ES | Effect Size |
| ICC | Item Characteristic Curve |
| IRT | Item Response Theory |
| LR | Logistic Regression |
| LRT | Likelihood Ratio Test |
| MH | Mantel-Haenszel |
| SIBTEST | Simultaneous Item Bias Test |
| SPSS | Statistical Package for Social Sciences |
| WLS | Weighted Least Squares |

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER ONE

# INTRODUCTION

## 1.1 Introduction

This chapter provides an introduction to the current study. The chapter is organized into nine sections. These include the background of the study; the statement of the problem; the purpose of the study; the research hypotheses; the assumptions of the study; the limitations of the study; the significance of the study; the conceptual framework and definitions of terms used in the study.

## 1.2 Background to the Study

Standardized tests can be used to determine skills or ability levels of various examinees. The skills may include personality characteristics, vocational tasks, or academic ability. However test scores can be affected by other sources of variation that may not be completely controlled. This may result in an unfair advantage to a particular sub-population of examinees, especially if the sub-populations are matched on the ability of interest (Cromwell, 2006). For instance the failure to understand and account for group differences on any test may lead to interpretations and consequently actions that are invalid. To ensure that tests are fair for all examinees, items should be screened for text that may be inappropriate to various sub-groups which would include female examinees, minority group examinees or disabled examinees.

Differential Item Functioning (DIF) is a statistical method for determining if item bias creates an unfair advantage to a sub population. DIF is defined as the different probability of giving the right answer to a test item by two individuals with the same Ability level, but from different groups (MaCarthy, Oshima & Raju, 2007). To determine the validity of tests, DIF

analysis can be employed to investigate whether educational and psychological measurement of structures differs in terms of groups (Crane, Gibsons, Narasimhalu, Lai & Cella, 2007). The difference in item performance can be due to the content of the item or to how the question in the item is posed (Wiberg, 2009). DIF can be determined by comparing two subpopulations' outcome on an item and determining the presence of DIF. DIF also involves a decision of whether there is a large enough difference between subpopulations to eliminate or change the item of interest.

A test item is considered to be biased when a dimension on the examination is deemed to be irrelevant to the construct that is being measured, placing one group of examinees at a disadvantage in taking the test (Roever, 2005). Thus, if DIF is not evident for an item, then there is no item bias. DIF is required but is not sufficient for item bias. That is, the presence of DIF is not sufficient to declare item bias. An item might show DIF, but not be considered biased if the difference is a result of the actual difference in the groups' ability to respond to the item. If one group of test-takers is at a high level and the other group of test-takers is at a low level, the lower group would perform significantly lower (Roever, 2005). If test-takers differed in knowledge, a difference in item responses would be expected. Consequently, a difference in the performance of groups of examinees with different abilities on specific items is not indicative of test bias, but rather of item impact (Schumacher, 2005)

To detect DIF, test of statistical significance may be required. The accuracy of statistical significance tests may therefore be influenced by large sample sizes which may cause a false positive or a Type I error for an unbiased item (Wang & Su, 2004). Statistical significance tests use probability ($p$) values and chi square ($\chi^2$) tests that are not robust to varying sample sizes that are not comparable across different methods. Statistical tests that use an Effect size

measure are preferred in order to control for false positives and quantify the amount of DIF when detected (Thompson, 2002). This takes into account not only the magnitude of DIF, but also replicability and generalizability (Huberty, 2002).

Possible influential outliers should be identified and their specific influence on the results examined. This can be done by using *Cook's D* to identify points that are suspicious from a statistical perspective. Removing the data points may be dangerous as it may end up destroying some of the most important information in the data. Removal can only be done if there exists some substantive information for eliminating outliers about these points and also whether they involve special properties or circumstances not relevant to the situation under investigation. If no such distinguishing features can be found then there are no clear grounds for eliminating outliers (Sarkar, et al 2011). A regression can be performed with and without the outliers and their specific influence on the results examined. If the influence is minor then it may not matter whether or not the outliers are eliminated. If the influence is substantial, then it is probably best to present the results of both analyses and simply alert the reader to the fact that these points may be questionable (Sarkar, et al 2011).

There are various DIF detection methods which include parametric methods such as IRT methods and non-parametric methods such as Simultaneous Item Bias Test (SIBTEST) (Shealy and Stout, 1993), Mantel-Haenszel (MH) and Logistic Regression (LR). IRT methods require data that is consistent with the assumption of normality and also require a lot of theoretical knowledge. The data used in this study violates the assumption of normality and this makes IRT an unsuitable method for use in this study. IRT is only robust in detecting DIF when the sample size is large. This makes it unsuitable for use in the current study since the data is simulated for large and small sample sizes. SIBTEST (Shealy & Stout, 1993) is a

non-parametric method which may be good in detecting DIF items but it has a main disadvantage in that it can only detect DIF for large sample sizes. This method is not recommended for use in a study where small sample sizes are required. Also SIBTEST works iteratively until all suspected items are removed from the valid subset. The final subsets of items that are DIF free are used as the matching criterion. This method is therefore not suitable for use in a study where all items are considered to be DIF items with different DIF magnitudes.

The Mantel-Haenszel (MH) statistic provides an Effect size measure that quantifies the magnitude of DIF when detected. The Effect size measure given by MH is referred to as the Odds Ratio (Mantel & Haenszel, 1959). The Odds Ratio can be used to quantify the magnitude of DIF into three categories namely negligible (Type A), moderate (Type B) and large (Type C) DIF. An item displaying negligible DIF is regarded as a good item. That displaying negligible DIF requires refinement while that displaying large DIF should not be included in the main test. Non-parametric statistics such as MH and LR violate the assumption of normality and can be used with small sample sizes.

Logistic Regression is another statistic that also provides an Effect size measure known as the Weighted Least Squares Estimate ($R^2$) which also quantifies the amount of DIF into three categories namely negligible (Type A), moderate (Type B) and large (Type C) DIF (Zumbo, 1999). Just like Mantel-Haenszel, items with negligible DIF are good items; those with moderate DIF need refinement while those with large DIF should not be included in the test (Swaminathan, 1993). An advantage of using logistic regression is that estimates of the regression coefficients can be plotted. This plot can then be used to detect where along the scale the DIF is becoming problematic (Miller et al., 1993). The LR procedure might give

clear perspective on the possible causes of DIF by inclusion of other relevant examinee characteristics. The effect of characteristics such as Sample size and Ability distribution and test characteristics such as Test length can be investigated using the LR statistic. LR procedures also use total score as a proxy for latent trait and this feature is important when considering all items to be DIF items. LR statistic can also be used with small sample sizes. Another important feature of the LR method is its ability to detect both uniform and non-uniform DIF. However since MH was weak in detecting non-uniform DIF, only uniform DIF was considered in the current study where it was compared with the LR statistic when detecting DIF items.

MH and LR statistics may have similar categories of quantifying DIF as Type A,B and C, but may differ in DIF detection under certain conditions some of which may cause false positives or Type I errors for unbiased items (Wang & Su, 2004). Some of the conditions may include examinee conditions such as Sample size and Ability distribution; and test conditions such as Test length. Several Monte Carlo DIF detection studies have noted that DIF items can be detected for large sample sizes. This may result in minimal variance and least error rates with DIF detection procedures (Gonzalez-Roma, Hernandez and Gomez-Binto, 2006). It was noted that for small sample sizes no DIF items were detected using the SIBTEST method (Salubayba, 2013). The effect of small sample sizes and large sample sizes on the detection of DIF items using MH and LR methods, which are robust to sample size conditions, can therefore be compared. The study also determined whether the effect of sample size depended on the type of DIF detected. Gonzalez- Roma Hernandez and Gomez-Binto (2006) perceived Sample sizes 100, 200, 400, and 800 to be large but were unable to distinguish between large and small Sample sizes when determining the effect of Sample size on DIF detection. Simulation studies can therefore be performed to determine effect of small sample sizes and

large sample sizes, on the Effect sizes using both Mantel-Haenszel and Logistic Regression statistics. Clerk (2010) used equal and unequal Sample sizes with large samples for both the focal and reference groups, but failed to investigate the effect of small sample sizes. These studies indicate that the influence of various sample sizes on the effect size while comparing MH and LR statistics has been inadequately studied especially when considering the effect of small sample sizes.

Ability distribution is another condition that is common in DIF detection. It is given in terms of mean and standard deviation. Some distributions are assumed to be normal with mean 0 and standard deviation 1. Some DIF statistics such as MH and LR are robust to various Ability distributions. Others such as IRT methods can only be used when the data have a normal ability distribution. The effect of Ability distribution on the Effect sizes of both MH an LR statistics can be of major concern to DIF researchers. Several studies have reported that a difference in mean ability of 1 standard deviation between certain reference and focal groups occurs frequently in real testing situations (French & Maller, 2007). Jodoin and Gierl (2002) simulated data for equal Ability distributions and unequal ability distributions under small sample sizes (250/250) to larger sample sizes (1000/1000) and set the unequal Ability distributions with a difference of .50 for the means of the reference and focal group with the same standard deviation. There were no differences in Type I error and power rates for the larger sample sizes. The findings of these studies indicate that the effect of ability distribution on the effect size and the number of DIF detections has been inadequately studied.

The number of test items that is used for DIF detection can be of great concern to DIF researchers. No standard exists as to how many items should be used for DIF detection as this change from one study to another. Some DIF detection methods may require use of all items

in a test, while others may require only those items suspected as DIF items (Guler & Penfield, 2009). Moreover, since total test score is used as the criterion variable for grouping examinees, a more reliable test score (longer test length) may result in improved performance of the MH procedure. However, Rogers and Swaminathan (1993) showed that Test length had no significant influence on the power of the MH procedure for DIF detection, but only long tests were used (40- and 80-items). Uttaro and Millsap (1994) used both short (20 items) and moderate (40 items) test lengths, but DIF was presented only in the studied item. For the 20-item test, the MH procedure gave inflated Type I error rate when the groups differed in ability distributions. However, the inflation in the Type I error rate disappeared entirely in the 40-item test. Moreover, test length generally had little effect on the detection rates in both the 20 and 40 item tests. The studies quoted earlier indicate that the influence of Test length on the power of the MH and LR statistics for DIF detection has rarely been studied using varied test lengths. They tested the influence of test length using statistical significant tests and found no effect. It was important to undertake a study that determines the effect of test length on the effect sizes using MH and LR DIF statistics.

MH and the LR DIF methods are currently seen as practical means of determining DIF because of their simplicity and ease of use, at the same time providing an effect size statistic to determine if the DIF found is damaging (Wang & Su, 2004; Swaminathan & Rogers, 1990). Simulation studies demonstrated that an Effect size could be incorporated with the MH (Roussos & Stout, 1996) and Logistic Regression DIF methods (Jodoin & Gierl, 2002). Logistic Regression analysis for DIF as a viable procedure for detecting DIF can be considered for two conditions: uniform and non-uniform. Uniform DIF occurs when there is no interaction between the Ability level and group membership. Non-uniform DIF occurs when there is an interaction between Ability level and group membership (Swaminathan &

Rogers, 1990). Logistic Regression DIF when compared to MH was found to be a better detector of both uniform and non-uniform DIF (Swaminathan & Rogers, 1990).

Zumbo and Thomas (1996) introduced the use of an Effect size; weighted-least-squares $R^2$ and were empirically tested through simulations by Jodoin and Gierl (2002). They focused on testing the Effect size for Logistic Regression and generating a classification guideline for negligible, moderate, and large DIF effect sizes. The MH DIF method has been compared to other methods like SIBTEST, when testing for Type I error with the use of an Effect size. The Logistic Regression DIF method and the inclusion of the Weighted Least Squares $R^2$ have been empirically compared to the MH DIF method by Hidalgo and Lopez-Pina (2004). Hidalgo and Lopez-Pina (2004) study did not however compare the DIF magnitudes of MH and LR under simulated conditions of sample size, ability distribution and test length. Simulations have an advantage in that they can allow one to manipulate data to suite various conditions by using computer software.

The data can also be replicated a number of times in order to reduce the variance of estimated parameters (Kristjansson et al, 2005). Studies have also reported that more replications produce parameter estimate with less sampling variance. While estimated other parameters may require large number of replications, Kristjansson et al (2005) proposed that when comparing the number of DIF items correctly detected, a small number of replications such as 10 may be sufficient. Studies have shown that using no replications or a very small number of replications may result in sampling variance that is large enough to seriously bias the parameters being estimated (Hambleton, Jones & Rodgers, 1993). These studies do not suggest the number of replications that are sufficient for DIF detection. It would be more informative to use one thousand replications for each condition to improve the accuracy of

the empirical estimations of the sampling distribution characteristics and to produce parameter estimates with less sampling variance.

### 1.3 Statement of the Problem

Differential item functioning DIF analysis is typically used to identify test items that are differentially difficult for respondents who have the same Ability level of knowledge or skill but differ in ways that should be irrelevant to their performance on a test. Conclusions drawn about group differences among examinee groups should therefore be accurate. A study by Salubayba (2013) noted that different conditions such as sample size affected the accuracy of some DIF detection methods. The study noted that for small sample sizes, DIF items were not detected using SIBTEST method. Hernandez and Gomez-Bento (2006) used Sample sizes of 100, 200, 400 and 800 which were perceived to be large enough to determine the effect of sample size on DIF detection. They found out that DIF items were detected only when the sample size was large. Studies have not determined the effect of small sample sizes on DIF detection. To detect DIF for small sample sizes DIF methods that are robust to small sample sizes can be identified and used to determine the effect of sample size on the effect size and the number of DIF detections.

The MH and LR DIF methods can be used both with small sample sizes and also non-normal data. The comparison between the two statistics was to find out whether the effect of sample size was dependent on the DIF detection procedure. The number of items on the entire test or measure can also affect the detection of DIF. Zumbo (1999) recommended typically to have a minimum of 20 items in a test. This recommendation has enabled most studies to be done on long test lengths as opposed to short test lengths. DIF researchers therefore may find it necessary to determine the effect of short tests, such as those often observed on personality

inventories, on DIF detection by comparing different DIF methods. This is to determine if the effect of test length was dependent on the DIF detection procedure such as Mantel-Haenszel and Logistic Regression. A study by Khalid (2011) used items of varied Test lengths (40-80 items) and noted its effect on the accuracy of DIF detection. It was found that the influence of Test Length was rather modest and that the number of items did not greatly affect the detection of DIF of any kind. Studies have found that differences in Ability Distributions, assessed in terms of mean and standard deviation of the data, affected DIF detection rates (French & Maller, 2007: Wang & Su 2004). They simulated data generated from a normal ability distribution for both focal and reference groups. Studies that have simulated non-normal ability distribution and compared the DIF detection using different DIF statistics are still limited.

## 1.4 Purpose of the Study

The main purpose of this study was to determine how different conditions such as Sample size, Test Length and Ability Distribution, affected the detection of Differential Item Functioning (DIF) using Mantel-Haenszel and Logistic Regression statistics.

The main objectives of the study were specifically:

1) To determine the effect of Sample size, Ability distribution and Test Length on the Effect size and the number of detections of DIF items across the DIF types using Mantel Haenszel Statistic;

2) To determine the effect of Sample size, Ability distribution and Test Length on the Effect size and the number of detections of DIF items across the DIF types using Logistic Regression Statistic;

3) To compare the effect of Sample size, Ability distribution and Test Length on the number of detections of DIF items across the DIF types using Mantel Haenszel and Logistic Regression statistics.

## 1.5 Research Hypotheses

Specifically, the following research hypotheses were addressed:

1) There is a significant effect of Sample size, Ability distribution and Test Length on the Effect size and the number of detections of DIF items across the DIF types using Mantel Haenszel Statistic.
2) There is a significant effect of Sample size, Ability distribution and Test Length on the Effect size and the number of detections of DIF items across the DIF types using Logistic Regression Statistic.
3) There is a significant effect of Sample size, Ability distribution and Test Length on number of detections of DIF items across the DIF types using Logistic Regression and Mantel-Haenszel statistics.

## 1.6 Assumptions of the Study

The following assumptions were made:-

1. The data generated was representing item responses on a dichotomously scored test by two sub populations representing the reference and focal group test takers.
2. The test-taker's answer to one item was independent of the test-taker's answer to any of the other items and that the independent variables were measured without an error.

## 1.7 Limitation of the Study

The study was limited to only two DIF detection procedures MH and LR DIF methods. The results would not therefore be generalized to other detection methods. The study was also limited to dichotomously scored items. Polytomous items did not form part of this study. The study was also limited only to simulated data using statistical computer software. While the results reveal significant findings and draw important implications in the field of DIF, simulation is prone to misspecification errors. Therefore generalization based on simulation studies must be treated with caution beyond the parameter range specified in the model. Real data did not form part of the study. The study was also limited to the use of an Effect size statistic. Statistical significance tests did not form part of the study because they inflate Type I error rates, when DIF analyses inherently uses large Sample sizes.

## 1.8 Significance of the Study

To ensure that accurate assessment occurs for all examinees, assessment instruments should be unbiased (Awuor, 2008). For example, items intended to measure reading proficiency must be valid for use with examinees from diverse groups for instance ethnicity, gender, and special education status for meaningful score interpretation (Finch & French, 2007). Continuing to evaluate the accuracy of detection methods is an essential step in gathering score validity evidence. However, studies that are focused on the effect of Sample size and Test length at different Ability distributions are still limited. This study was designed to determine the statistical power of DIF detection using LR and MH procedures under different conditions. The findings of this study would be useful to test developers who should ensure that the tests administered to examinees are free of factors not relevant to the concepts being tested. Also educational policy makers who should make sure that item analysis is done before the items are presented to groups of examinees, and also to make informed decisions

related to test development. The present study can give many researchers a place to begin researching with prior knowledge of MH DIF analyses. This simulation study shows that researchers have to worry about varying Sample sizes and Ability distributions because they seem to affect the statistics for DIF analyses. The present study was a systematic study that offers a baseline for many types of studies DIF to be researched on. The findings of this study can also contribute to research and practice in schools and institutions' testing program, the formulation and implementation of educational policies and decisions related to test development. Test developers and test users can use the findings to make informed decisions regarding the selection of test item evaluation procedures in the area of Differential Item Functioning under different examinee conditions. Perfectly good items should not be discarded from a test because of inaccurate functioning of a DIF detection procedure.

## 1.9 Conceptual Framework

The present study adopted the measurement model for the relationship between different conditions and observed responses by Engelhard (2016) as shown in Figure 1.1.

The model is written as follows;

$$\ln\left[P_{nijk1}\Big/P_{nijk0}\right] = \theta_n - \delta_i - \alpha_j - \lambda_k$$

Where $P_{nijk1}$ = the probability of person n succeeding on an item i for group j and condition k,

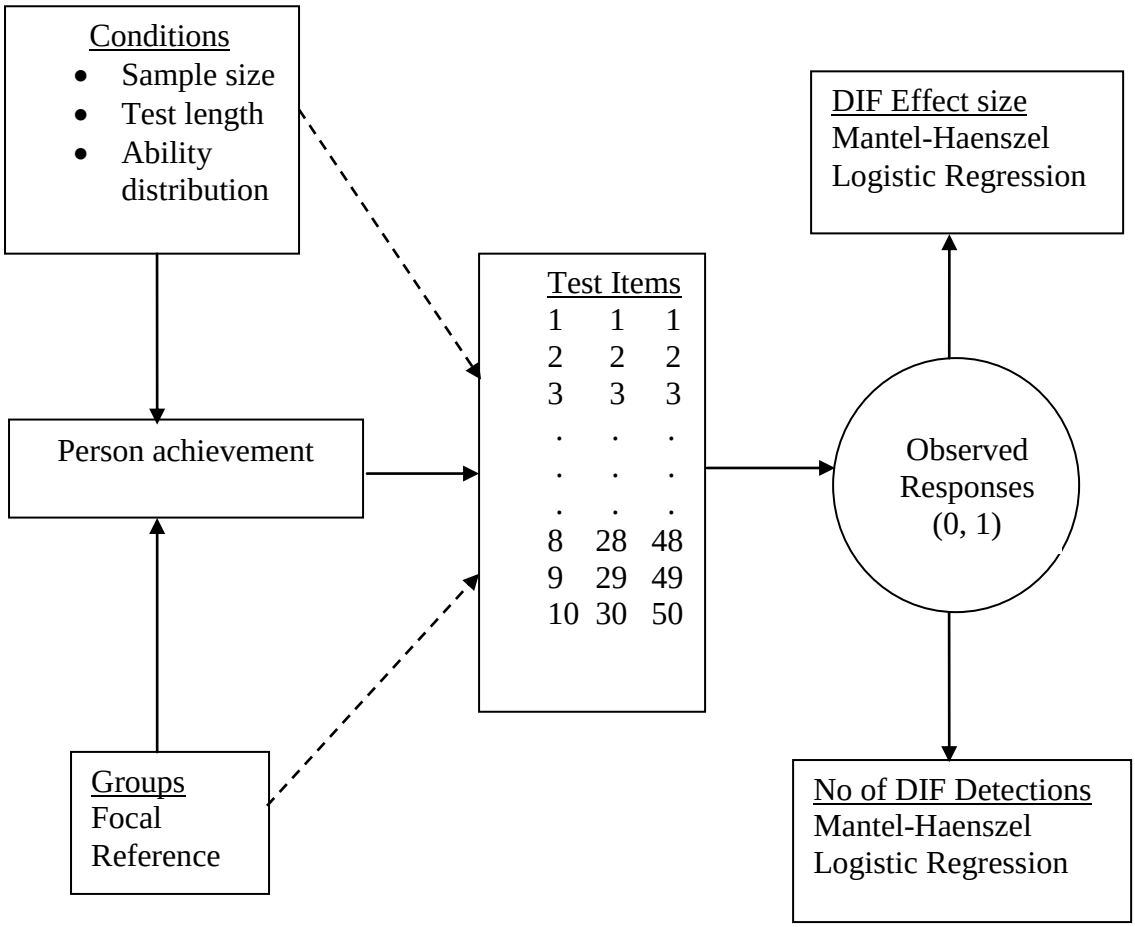



***Figure 1.1 Measurement model for the relationship between different conditions and observed responses***.

$P_{nijk0}$ = the probability of person n failing on an item i for group j and condition k,

$\theta_n =$ location of person n on the latent variable,

$\delta_i$ = difficulty of item *i*,

$\alpha_j$= location of group j, and

$\lambda_k$= location of condition k.

$\alpha_j$ = location of group j, and

$\lambda_k$ = location of condition k.

The idea is that person achievement is the latent variable that is made observable through a set of 10, 30 and 50 items and that the items vary on their locations on the latent variable.

The observed responses are dichotomous, and they are a function of both person achievement and item difficulty (Engelhard, 2016). Group and administration condition may influence person achievement levels. The broken lines from groups and conditions are modeled as interaction effects between groups, conditions and items and are considered as a construct-irrelevant source of variance. In other words, the item locations are not invariant over groups or conditions. This model can be written in exponential form as follows:

$$P_{nijk1} = exp[\theta_i - \delta_i - \alpha_i - \lambda_k]/\gamma$$

Where $P_{nijk1}$ = the probability of person n succeeding on an item *i* for group *j* and condition *k*, and

$\gamma$ = sum of all possible numerators.

Once estimates of the main effect parameters are obtained this equation can be used to define a residual. The unstandardized residual reflects the difference between the observed and expected responses. The observed responses can then be transformed to the dependent variables which are the effect sizes and the number of DIF detections by DIF detection methods such as MH and LR.

## 1.10 Definition of Terms

**Ability distribution:** The ability level of individuals given in terms of mean and standard deviation.

**Academic Performance:** This is described as the scholastic standing of a learner at a given moment. It refers to how an individual is able to demonstrate his or her intellectual abilities.

**Accuracy:** The degree to which the measurement procedure represents the concept under study. It is concerned with the validity of the measure.

**Adverse Impact:** Adverse impact is a term describing the situation in which group differences in test performance result in disproportionate examinee selection or related decisions (e.g., promotion). This is *not* evidence for test bias.

**Coefficient of determination $R^2$:** This refers to the proportion of variance in one variable which is shared by the other. Or the amount of knowledge one variable provides in determining the values of the others.

**Dichotomously scored items:** Items scored only on the basis of 0 and 1 for instance some multiple choice items.

**Differential Item Functioning (DIF):** This occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item *after matching on the underlying ability* that the item is intended to measure.

**Focal Group:** A group of examinees with conceptual or policy reasons for concern with its performance.

**Item analysis:** A set of statistical techniques to examine the performance of individual items. This is important when developing a test or when adopting a known measure.

**Item bias:** Item bias occurs when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. DIF is required, but not sufficient, for item bias.

**Item impact:** Item impact is evident when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item because there are true differences between the groups in the underlying ability being measured by the item.

**Item Response Theory (IRT):** It is a modern test theory that describes the interaction between ICC and person abilities. Because ability is not manifested directly, it is also referred to as Latent Trait Theory.

**Monte Carlo Method:** A technique that involves using random numbers and probability to generate data.

**Outlier:** An observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.

**Polytomously scored items:** Items scored on a scale other than 0 or 1.

**Reference Group:** A group of examinees without conceptual or policy reasons for concern with its performance.

**Sample Size:** Number of respondents that take part in a particular study.

**Simulation Study:** A computer based imitation of the operation of a real process using a model that represents the behaviors and functions selected.

**Test Length:** The number of items that constitute the entire test.

**Type A DIF items:** Test items displaying negligible DIF and are retained in a test.

**Type B DIF items:** Test items displaying moderate DIF and can be retained in a test.

**Type C DIF items:** Test items displaying large DIF and are removed from the test.

**Type I error rate:** The proportion of items with DIF falsely identified.

**Type II error rate:** The proportion of items with DIF not identified.

**Uniform DIF:** It exists when there is no interaction between ability level and group membership. The probability of answering the item correctly is greater for one group than the other uniformly over all levels of ability.

**Non-uniform DIF**: It exists when there is interaction between ability level and group membership, that is, when the difference in the probabilities of a correct answer for the two groups is not the same at all ability levels.

**Validity:** Validity as used here is in reference to construct validity focusing on the degree to which true differences between groups in the underlying ability provide evidence that supports that the interpretations of the scores are correct and that the manner in which the interpretations are used is appropriate.

# CHAPTER TWO

# LITERATURE REVIEW

## 2.1 Introduction

This chapter reviews the related literature on Differential Item Functioning under varied conditions. As noted little research has been done in Kenya and therefore the bulk of the critical review comes from outside Kenya sources. The chapter is organized into five sections. The first section highlights the accuracy of DIF detection methods. The second explains DIF detection across Sample size conditions. The third looks at DIF detection across Ability Distribution conditions. The forth section highlights DIF detection and Test length. The last section looks at DIF detection and Item bias.

## 2.2 Mantel-Haenszel Procedure and Different conditions

Differential item functioning (DIF) refers to differences in the functioning of items across groups, often times demographic, which are matched on the latent trait or more generally the attribute being measured by the items or test. In studies of differential item functioning (DIF), researchers study whether the probability of correct response to an item is the same for two groups after controlling for differences in ability (DeMars, 2015). One group is termed the reference group and one is termed the focal group. If there are conceptual or policy reasons for concern with one group's performance, that group is labeled the focal group. Otherwise, the distinction is arbitrary. Simulations are frequently used when developing new models or estimators for DIF analyses. A simulation is a computer based imitation of the operation of a real process using a model that represents the behaviors and functions selected. Simulation can be used to show eventual real effects of alternative conditions and courses of action. It involves the acquisition of a valid source of information, relevant selection of key characteristics and behavior, the use of assumptions and validity of the outcomes. Monte-

Carlo techniques using pseudo random numbers so selected and runs for some boundary conditions. Computer simulation models real life situation on a computer that can be studied to see how a system works or a pattern of behavior manifests. By changing variables in the simulation, predictions can be made about the behavior of a system or trait (De Boeck, 2008).

With real data, researchers can only determine whether two estimators are different, but with simulations they can determine which is more accurate because the true values are known. For simulation studies, DIF is often modeled in an IRT framework as a difference in item parameters. The secondary trait underlying the DIF is thus categorical, applying to all group members. The shift in the item response function is the same for every member within the focal group or within the reference group (DeMars, 2015). For example, if the discrimination parameter is the same for both groups but the item difficulty is higher for the focal group, the item will be harder for all focal group members than for matched reference group members. However, it might seem more reasonable to conceptualize the item parameter difference as representing the *average* DIF, not a constant that applies to each group member. The response probabilities of some focal group members may be more like the response probabilities of reference group members, and the other way round. Some have proposed applying IRT mixture models to account for this (Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002). In mixture models, the reference and focal group may be disproportionally distributed across latent classes, but some members of each group may have a high posterior probability of membership in the class disadvantaged by the DIF item. However the use of latent class models implies that the DIF changes the response function in the same way for all members of a class. DIF might be more realistically conceptualized as a continuous secondary trait that is distributed differently in the reference and focal groups (Ackerman, 1992; Camilli, 1992; Roussos & Stout, 1996). Another equivalent way of

phrasing this is that DIF could be considered a random effect. In the DIF literature, the most common indices, such as the Mantel-Haenszel (Dorans, 1989; Holland & Thayer, 1988), and SIBTEST (Shealy & Stout, 1993), do not use IRT. Instead, they condition on observed score or on a true score based on classical test theory (CTT). However, simulated data for studying observed-score indices is typically generated with IRT models, and the choice of a continuous or categorical secondary trait for the data generation may have implications for the findings for indices conditioned on observed score.

Most simulation studies have generated the DIF as a categorical trait, but some scholars have argued that theoretically a continuous secondary trait causes DIF. Ackerman (1992) provided a detailed explanation of how group differences in the distribution of a secondary trait, termed a nuisance trait, would lead to DIF. Roussos and Stout (1996) and Camilli (1992) each described further mathematical details. Other researchers have advocated this conceptualization (Bolt & Stout, 1996; Douglas, Roussos, & Stout, 1996). Additionally, Ackerman (1992), Jiang and Stout (1998), and Shealy and Stout (1993) simulated DIF using a continuous secondary trait and studied power for the MH and SIBTEST procedures.. If the research question to be explored was: Does simulating DIF with an IRT model, compared to a unidimensional model with group-specific item difficulties, yield Mantel-Haenszel DIF effect size estimates that differ in bias or standard error?, then there was no reason to think that the bias would depend on the method of simulating DIF. It seemed plausible that the estimates might be more stable across replications when the DIF was simulated with a unidimensional model because the difference in item difficulties had the same effect on all group or class members. If the IRT conceptualization of DIF better represents real-life cognitive processes, the stability of the uni-dimensional method of simulation would be inaccurate (DeMars, 2015).

When examining items for DIF, the groups are matched on the measured attribute, otherwise this may result in inaccurate detection of DIF (Osterlind & Everson, 2009). Common procedures for assessing DIF are item response theory (IRT) based methods, Simultaneous Item Bias Test (SIBTEST) and Mantel-Haenszel method. The MH, the SIBTEST non-parametric methods and the IRT is a parametric method (Tan & Gierl**,** 2005). The parametric method when used in a simulation study, assumes a specific item response model especially the assumption of normality and the large sample size condition.

The non-parametric approach, and does not assume a specific item response model. Finch (2005) observed that for detecting DIF in dichotomous items with the nonparametric approach, the Mantel-Haenszel procedure (Holland & Thayer, 1988; Mantel & Haenszel, 1959) was among the most popular ones. Swaminathan and Rogers (1990) and Rogers and Swaminathan (1993) have classified DIF into two kinds: uniform and non-uniform DIF. In their classification they suggested that uniform DIF exists when there is no interaction between ability level and group membership. The existence of uniform DIF, therefore, suggests that the probability of answering the item correctly is greater for one group than the other uniformly over all levels of ability. When there is interaction between ability level and group membership, that is, when the difference in the probabilities of a correct answer for the two groups is not the same at all ability levels, then non uniform DIF exists (Finch, 2007). The accuracy of a DIF statistic is determined by its ability to detect uniform and non- uniform DIF under various conditions and also its ability to control for Type one error rates. Several DIF researchers have reported that DIF detection by either M-H or an IRT based procedure resulted in inflated Type I error. The studies indicated that once the percentage of DIF items in a test increased to 10% or 15%, the MH method began to lose control over the Type I error (Fidalgo, Mellenberg, & Muñiz, 2000).

However, study results have shown that high percentages of DIF items do not necessarily lead to inflated Type I error for the M-H and IRT- based DIF detection methods (DeMars, 2010). Awuor (2008) stated that SIBTEST DIF methodology controls for ability while detecting items that exaggerate the ability difference across groups of examinees. Thus a large DIF value obtained with SIBTEST or MH suggests that the item is more likely to be measuring additional constructs that function differently from one group to another. Gierl et al., (2004) used SIBTEST to detect the presence of DIF and to quantify the size of DIF. They divided the items in their study into the studied or suspect subtest and the matching or valid subtest as normally required to operationalize SIBTEST. The studied subtest contained the items believed to measure the primary and secondary dimensions based on the substantive analysis whereas the matching subtest contained the items believed to measure only the primary dimension. The matching subtest was intended to place the reference and focal group examinees into subgroups at each score level so their performance on items from the studied subtest could be compared. The study was unable to detect DIF where it was highly suspected, in order to determine the effect of variation on the power of a procedure to detect DIF of moderate and large magnitudes. The inability to detect DIF where it is was highly suspected or where it was modeled in the study and the mixed results in studies suggested that there is a gap in the DIF studies that needed to be bridged.

### 2.2.1 The Mantel-Haenszel (MH) Procedure

This is a non-parametric approach for identifying DIF (Mantel & Haenszel, 1959). MH yields a Chi-square test with one degree of freedom to test the null hypothesis that there is no relation between group membership and test performance on one item after controlling for ability. MH is computed by matching examinees in each group on total test score and then forming a 2 (group) × 2 (item response) × K (score level), contingency table for each item

where K is the score level on the matching variable of the total test score. At each score level j, a 2×2 contingency table is created for each item. The MH-$\chi^2$ statistic tests the null hypothesis that there is no relationship between group membership and test performance on one item after controlling for overall test performance.

The MH statistical procedure (Mantel & Haenszel, 1959) 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 the contingency table analysis. For dichotomous items, K contingency tables (2×2) are constructed for each item, where K is the number of test score levels into which the matching variable has been divided.

Table 2.1 shows the 2×2 table for calculating the MH statistic for item *i* on a *j* score level in the test. In typical applications of the MH procedure an item shows uniform DIF if the odds of correctly answering the analyzed item at a given score level *j* is different for the two groups at some level *j* of the matching variable.

The odds ratio ($\alpha$) is given by:

$$\alpha = ( P_{Rj}/1\text{-}P_{Rj} ) / (P_{Fj}/1\text{-}P_{Fj})$$

in which $P_{Rj}$ and $P_{Fj}$ are the correct item response probabilities for the reference group and the focal group respectively. The test score level *j* is calculated as follows:

$$P_{Fj} = \frac{C_j}{N_{Fj}} \quad \text{and} \quad P_{Rj} = \frac{A_j}{N_{Rj}}$$

**Table 2.1 Score on $i^{th}$ item and *j* score**

| Group | 1 | 0 | Total |
|---|---|---|---|
| Reference | $A_j$ | $B_j$ | $N_{R.j}$ |
| Focal | $C_j$ | $D_j$ | $N_{F.j}$ |
| Total | $N_{1j}$ | $N_{0j}$ | $N_{..j}$ |

The MH statistic for detecting DIF in an item is expressed as:

$$MH = \frac{\left[ \left| \sum_{J=1}^{K} A_j - \sum_{j=1}^{k} E(A_j) \right| - 0.5 \right]^2}{\sum_{j=1}^{k} Var\ (A_j)}$$

in which E ($A_j$) = ( $N_{Rj}N_{1.j}$)/$N_{..j}$ and Var ($A_j$) = $N_{Rj}N_{Fj}N_{1.j}N_{0.j}$/($N_{..j}$) 2($N_{..j}$ – 1). The MH statistic, under the null hypothesis, is distributed as a $\chi^2$ distribution with one degree of freedom. Under the MH procedure an effect size estimate based on the common odds ratio α is expressed as

$$\alpha_{MH} = \frac{\sum_{j=1}^{K} A_j D_{j/} \ / \ N_{..j}}{\sum_{j=1}^{K} B_j C_{j/} \ / \ N_{..j}}$$

Holland and Thayer (1988) proposed a logarithmic transformation of α for interpretive purposes, with the aim of obtaining a symmetrical scale in which a zero value indicates an absence of DIF, a negative value indicates that the item favours the reference group over the focal group and a positive value indicates DIF in the opposite direction. This transformation is expressed as

$$\Delta\alpha_{MH} = -2.35\text{In}\ (\alpha_{MH})$$

Based on this transformation, Zwick and Ercikan (1989) proposed the following interpretation guidelines to evaluate the DIF Effect size. Type A items – negligible DIF: items with $|\Delta\alpha_{MH}| < 1$, Type B items – moderate DIF: items with $1 \leq |\Delta\alpha_{MH}| \leq 1.5$ and the MH test statistically significant, Type C items – large DIF: items with $|\Delta\alpha_{MH}| > 1.5$ and the MH test statistically significant. Zwick and Ercikan (1989) pointed out that Type B items could be used in the test if there are no others to replace them, and that Type C items can be selected only if they are necessary to meet test specifications.

In polytomous items the data is organized into K two dimensional 2×c tables, where c is the number of response categories in the item. Table 2 shows the contingency table for item *i* with level *j*.

**Table 2.2 Scores on $i^{th}$ item**

| **Group** | **$R_1$** | **$R_2$** | **$R_3$** | | **Rc** | **Total** |
|---|---|---|---|---|---|---|
| Reference | $N_{R1j}$ | $N_{R2j}$ | $N_{R3j}$ | ..... | $N_{Rcj}$ | $N_{R.j}$ |
| Focal | $N_{F1j}$ | $N_{F2j}$ | $N_{F3j}$ | ..... | $N_{Fcj}$ | $N_{F.j}$ |
| Total | $N_{.1j}$ | $N_{.2j}$ | $N_{.3j}$ | ..... | $N_{.cj}$ | $N_{...j}$ |

Mantel (1963) proposed a statistic which is an extension of the standard Mantel-Haenszel procedure. The Mantel statistic is computed by means of the following expression:

$$MANTEL = \frac{\left[\sum_{j=1}^{K} F_j - \sum_{J=1}^{K} E(F_j)\right]^2}{\sum_{j=1}^{k} Var(F_j)}$$

in which F*j* is expressed as

$$F_j = \sum_{c=1}^{c} R_c N_{Fcj}$$

Or the total score for the focal group at K ability level.

Based on the general characteristics of the MH procedure, new statistics have been developed, for example, the Breslow-Day Chi square (Breslow & Day, 1980) and new procedures for DIF detection such as the combined decision rule (Penfied, 2003). The MH procedure is used to estimate the constant odds ratio that yields a measure of effect size for evaluating the amount of DIF that is present. The proposed values for the constant odds ratio called the $\Delta$-*MH* when transformed onto the delta scale which serves as the effect size measure for classifying the DIF at item level (Hidalgo & Lopez-Pina, 2004). DIF is considered negligible when $\Delta$-*MH* is not significantly different from 0 and the magnitude is $|\Delta\text{-}MH| < 1.5$. DIF is considered moderate when $\Delta$-*MH* is significantly different from 0 and has either (a) $1 \leq |\Delta\text{-}MH| < 1.5$ or (b) $|\Delta\text{-}MH|$ is at least 1 but not significantly greater than 1. DIF is considered large when $\Delta$-*MH* is significantly greater than 1 and $|\Delta\text{-}MH| \geq 1.5$ (Zieky, 1993). These ratings are referred to as A, B, C level of DIF to denote negligible, moderate and large amounts of DIF. The main limitation of the MH method is its inability to detect non-uniform DIF.

Chi-square statistics are affected by sample size; therefore, testing for both statistical significance and effect size might be useful to avoid detecting items with small practical

significance erroneously, such as DIF items (Clauser & Mazor, 1998; Millsap & Everson, 1993). Another issue that pertains to sample size directly relates to the statistical procedure being used to detect DIF. Aside from sample size considerations of the reference and focal groups, certain characteristics of the sample itself must be met to comply with assumptions of each statistical test utilized in DIF detection. For instance, using IRT approaches may require larger samples than required for the Mantel-Haenszel procedure. This is important, as investigation of group size may direct one toward using one procedure over another. The MH method was selected for this study because it can be used with small sample sizes and also non-normal data. The MH method has been rarely used to detect DIF items where small sample sizes and non-normal data were required. Non parametric statistics such as MH violate the assumption of normality that is required for parametric methods such as IRT.

Although the MH procedure is one of the most utilized DIF methods due to its simplicity and practicality, it also has some major drawbacks. The MH procedures are successful in detecting uniform DIF but it might yield misleading results in non-uniform DIF or when using more complex models (DeMars, 2009; Güler & Penfield, 2009; Millsap & Everson, 1993; Narayan & Swaminathan, 1996). The MH method can only be considered in a study such as the current one, where uniform DIF is required.

### 2.2.2 Mantel-Haenszel and Simultaneous Item Bias Test (SIBTEST)

The Simultaneous item bias test (SIBTEST), generated by Shealy and Stout (1993), provides a DIF procedure that can do a set of DIF analyses at the same time. In SIBTEST, items suspected to be functioning differentially are called “suspected subsets” and remaining items are called “valid item subsets”. The SIBTEST matches reference and focal group according to their estimated latent ability based upon the observed score on what the practitioner

considers to be the valid items. First, examinee scores are calculated on a valid subset, and then the proportion of correct responses is calculated for suspected items. The SIBTEST works iteratively until all suspected items are removed from the valid subset. The final subsets of items that are DIF free are used as the matching criterion (Clauser & Mazor, 1998). The SIBTEST can detect both uniform and non-uniform DIF. The hypotheses for testing uniform and non-uniform DIF are

$$H0: \beta \text{ uni} = 0 \text{ vs. } H1: \beta \text{ uni} \neq 0,$$

where β uni is the parameter specifying the amount of DIF for an item can be rejected (Shealy & Stout, 1993). β uni is defined as:

$$\beta \text{ uni} = \int B(\Theta)\, fF\,(\Theta) d\,\Theta,$$

where d Θ is the differential of theta, B (Θ) is integrated over Θ to produce β uni, a weighted expected mean difference in the probability of a correct response on an item between reference and focal group examinees who have the same ability. The difference in the probabilities of correct response for examinees from the reference and focal groups can be expressed as:

$$B(\Theta) = P(\Theta, R) - P(\Theta, F)$$

SIBTEST detects bias by comparing the responses of examinees in the reference and focal groups that have been allocated to bins using their scores on a "matching subtest" (Stout & Roussos, 1996). The matching subtest is a subset of items that, ideally, are known to be unbiased. In most practical applications, the user does not have accurate a priori knowledge regarding bias. Fortunately, simulation studies have shown that the SIBTEST procedure is tolerant of small to moderate amounts of contamination of the matching criterion (Shealy & Stout, 1993). These studies have found that the Type I error rates are not inflated

substantially when the matching subtest contains relatively few biased items, however, the Type II errors are more likely because the power to detect DIF is reduced by contamination. A newer version of SIBTEST (Shealy & Stout, 1993) can be used to compute a weighted mean difference between the reference and focal groups. The means in this procedure are adjusted to correct for any differences in the ability distributions of the reference and focal groups using a regression correction procedure, and in effect, creates a matching subtest free from statistical bias (Jiang & Stout, 1998). Results from simulation studies reveal that the regression correction procedure reduces Type I error under many testing conditions (e.g., Bolt & Gierl, 2004; Roussos & Stout, 1996b; Shealy & Stout, 1993). DIF items are interpreted in terms of beta estimates, where categories A ($|\beta| < .059$), B ($059 \leq |\beta| < .088$), and C ($|\beta| \geq .088$) represent small, median and large DIF, respectively, (Roussos & Stout, 1996).

The SIBTEST method may be good in detecting DIF items but it has a main disadvantage in that it can only detect DIF for large sample sizes. This method is therefore not suitable for use in a study where small sample sizes are considered. Also SIBTEST works iteratively until all suspected items are removed from the valid subset. The final subsets of items that are DIF free are used as the matching criterion. This method was unsuitable for use in a study where all items were considered as DIF items but of different magnitudes, and where no purification was done to get rid of items suspected as non DIF items. A disadvantage of the SIBTEST method is that it is not entirely robust to between-group differences in the unconditional distribution of the ability. This is a problem which the SIBTEST shares with other methods which use the proportion difference, if between group effects are present in the unconditional distribution of the examinees ability (Penfield & Camilli, 2007). SIBTEST differs with MH method in that it can perform both uniform and non-uniform DIF analysis, MH can only

perform uniform DIF analysis. MH method does not also work iteratively to remove all the suspected items from the valid test.

### 2.2.3 Mantel-Haenszel DIF Detection and Sample Size

A major consideration in DIF detection pertains to issues of sample size, specifically with regard to the reference and focal groups. Prior to any analyses, information about the amount of people in each group should be typically known such as the number of males/females or members of ethnic/racial groups (Özlem & Özbek, 2016). However, the issue more closely revolves around whether the amount of people per group is sufficient for there to be enough statistical power to identify DIF. In some instances such as ethnicity there may be evidence of unequal group sizes such that Whites represent a far larger group sample than each individual ethnic group being represented. Therefore, in such instances, it may be appropriate to modify or adjust data so that the groups being compared for DIF are in fact equal or closer in size (Finch, 2016). Dummy coding or recoding is a common practice employed to adjust for disparities in the size of the reference and focal group. In this case, all Non-White ethnic groups can be grouped together in order to have a relatively equal sample size for the reference and focal groups. This would allow for a "majority/minority" comparison of item functioning. If modifications are not made and DIF procedures are carried out, there may not be enough statistical power to identify DIF even if DIF exists between groups (Özlem & Özbek 2016). Equal sample sizes can be manipulated for both the reference and the focal groups in a simulation study.

Another issue that pertains to sample size directly relates to the statistical procedure being used to detect DIF. Aside from sample size considerations of the reference and focal groups, certain characteristics of the sample itself must be met to comply with assumptions of each

statistical test utilized in DIF detection (Erdem, 2014). For instance, using IRT approaches may require larger samples than required for the Mantel-Haenszel procedure. This is important, as investigation of group size may direct one toward using one procedure over another. Additionally, as with all analyses, statistical test assumptions must be met. Some procedures are more robust to minor violations while others less so. Thus, the distributional nature of sample responses should be investigated prior to implementing any DIF procedures. For instance in a study that has small sample sizes, the MH statistic may be considered over SIBTEST and IRT based DIF methods that can detect DIF items for large sample sizes. This was the basis on which the MH procedure was selected for use in the current study because it is robust to both small and large sample sizes unlike SIBTEST and IRT methods which can only detect DIF with large sample sizes.

A study by Salubayba (2013) noted that different conditions such as sample size affected the accuracy of some DIF detection methods. The study used the SIBTEST method to determine the effect of large and small sample sizes DIF detection. The study determined whether DIF items could be detected when using small or large sample sizes. It was noted that below sample size 100, DIF items were not detected using SIBTEST method. This finding indicated that SIBTEST method was not suitable for detecting DIF items with small sample sizes. The study did not determine the significant effect of sample size on the DIF detection method. When selecting a DIF detection method to use for DIF detection, small sample sizes are of major concern especially when the reference and focal groups are considered equal in size. Methods that are robust to small sample sizes such as Mantel-Haenszel can be selected for DIF analysis and also studying the effect of small sample sizes on the effect sizes and the number of DIF detections. Some classroom situations have students as low as 20 which informed the selection of the small sample size.

A study by Hernandez and Gomez- Bento (2006) used sample sizes of 100, 200, 400 and 800 which were perceived to be large enough to determine the effect of sample size on DIF detection also using SIBTEST method. It was found out that in all the sample sizes studied, DIF items were detected. The study did not however determine the effect of sample size on the procedure of DIF detection but determined if DIF items were detected using different sample sizes. The study also used only one DIF detection method which was robust to large sample sizes. The study did not use small sample sizes and also did not compare different DIF methods to determine their statistical power under different sample size conditions. The current study used the MH statistic to detect DIF items for both small and large sample sizes. The current study also determined the effect of sample size on the effect size of different DIF types using MH statistic. The current study also compared the number of DIF detections of different types using MH method using statistical graphs. The comparison was to determine if the effect of sample size was dependent on the DIF type.

Too many outcome measures may decrease efficiency of a study and increase the occurrence of chance differences. Gonzalez-Romá, Hernandez and Gomez-Benito (2006) conducted a simulation study to investigate statistical power and Type I error rate of a procedure based on the mean and covariance structure analysis model to detect DIF. They manipulated the type of DIF (uniform and non-uniform), DIF magnitude, (low, medium and large), equality or inequality of latent trait distribution and equality and inequality of sample sizes (100, 200, 400, and 800) across groups. They chose these sample sizes because they perceived that these were the sample sizes that were representative of those available to most researchers and practitioners and that did provide a wide range for testing the influence of sample size. In the study four, conditions showed equal sample sizes for both the focal and reference groups

(800-800, 400-400, 200-200, and 100-100). In six conditions both groups showed unequal sample sizes (800-400, 800-200, 800-100, 400-200, 400-100, and 200-100) with the reference group being the largest group because in empirical DIF studies the reference group is usually the larger group. The test that was simulated had 10 items with one item manipulated to demonstrate DIF. Results of Gonzalez- Romá et al., (2006) study indicated that when both groups' sample sizes were as low as 200/200 and 400/200 respectively, the mean and covariance analysis procedure showed acceptable power level to detect medium-sized uniform and non-uniform DIF. However, the results also indicated that power increased as sample sizes and DIF magnitude increased and that the control for Type I error was better when sample sizes were large (Romá, 2003). This study used many large sample size conditions which differed between the reference and the focal groups. This also raises the concern that studies using small sample sizes have rarely been done. The current study used small and large sample size conditions for both the focal and reference groups to determine the effect of sample size on DIF detection using MH method. The test that was simulated in the current study had 10, 30 and 50 items and none was manipulated to demonstrate DIF unlike the previous study which had only 10 items.

A study by Gierl, Gotzmann, and Boughton (2004) used the SIBTEST procedure to determine the effect of DIF conditions and DIF percentages on the DIF detection rates .The DIF conditions were balanced and unbalanced DIF conditions when DIF percentages were manipulated. They observed that when the DIF percentage and sample size were small adverse effects in DIF detection rates were not experienced. However, with large DIF percentage of 40% and 60% in the studied and matching subtests respectively the proportions of incorrect decisions increased as sample size increased for most conditions. On the basis of the study results they concluded that SIBTEST provided adequate DIF detection because

incorrect item rejections were less than 5% and the correct rejections were greater than 80% when DIF was balanced and sample sizes were at least 1,000 examinees per group. In the Gierl et al., study SIBTEST had inadequate DIF detection in all 40% and 60% unbalanced DIF conditions. This study manipulated variables with a large sample size such as 1000. The Gierl et al., study results were sited because they were considered useful guide in the interpretation of the results of a study on the effect of unequal sample sizes when DIF percentage was fixed and when purification was not needed because the DIF items were known a priori. The study simulated DIF conditions to determine DIF detection. This was consistent with the current study which manipulated sample size conditions using statistical software. This study did not however compare SIBTEST with another DIF detection procedure with similar statistical properties. A study that compares DIF methods is likely to give consistent results in DIF detection. The current study further determined the effect of sample size on the statistical power of the MH statistic.

### 2.2.4 Mantel-Haenszel DIF Detection and Ability Distribution

Several studies have found that differences in Ability Distributions, sometimes referred to as *impact*, affect DIF detection rates (French & Maller, 2007: Wang & Su, 2004). Ability Distribution was assessed in terms of mean and standard deviation. A study by Fidalgo and Laura (2018) was designed to determine whether the capability of DIF detection is affected by ability distribution using generalized MH statistics for polytomous items. The results showed that there was little impact of ability distribution on the performance of DIF statistic. The study used the reference and the focal group with the same ability distribution, mean 0 and standard deviation 1 and a non-normal distribution with mean -1 and standard deviation 1 to generate the data. Data was generated from the partial credit model (PCM) for the four point polytomous items. The software used for data generation was however not stated. This

study did not determine the effect of ability distribution using dichotomous items by comparing normal and non-normal ability distributions. Also no comparison was made with other DIF methods to determine whether the effect of ability distribution was dependent on the DIF statistic used. The current study generated dichotomous data with normal ability distribution mean 0, standard deviation 1 and non-normal, mean 1 and standard deviation 2 using WINGEN 3 computer software. The study also determined whether the effect of ability distribution was dependent on the DIF procedure selected, using the MH statistic.

A study by Swaminathan and Narayanan (1994) was carried out to determine the performance of Mantel-Haenszel and SIBTEST procedures for detecting Differential item functioning. The study simulated data to reflect conditions varying in sample size and ability distribution differences between the focal and reference groups, proportion of DIF items in a test, DIF effect sizes and type of item. The study manipulated five factors: sample size, ability distribution, differences, proportion of items containing DIF, DIF effect size and type of item. The ability distribution of the focal and reference groups was set as equal with mean 0 and standard deviation 1 and taken as equal ability distribution. The second condition was unequal with mean 0.0 and standard deviation 0.5 for the reference and mean 0.0 and standard deviation 1 with a difference in SD of 0.5. The third was mean 0 and -1 for the reference and focal groups with standard deviation of 1. 1296 conditions were studied The SIBTEST and MH procedures were used in DIF detection. ANOVA was performed to determine the effects of five conditions on the performance of MH and SIBTEST statistics. The data was replicated 100 times.

The findings of the study showed that the SIBTEST and MH procedures were equally powerful in detecting uniform DIF for equal ability distributions. The SIBTEST procedure was more powerful than MH in detecting DIF for unequal ability distributions. Both

procedures had sufficient power to detect DIF for a sample size of 300 in each group. Ability distribution did not have a significant effect on the SIBTEST procedure but it did affect the MH procedure. The study did not manipulate the number of items although a test length of 40 items was used to test the power of SIBTEST and MH. The independent variables were the five different conditions that were manipulated in the study while the dependent variables were the number of DIF items detected. The study did not look at the effect of ability distribution on the type of DIF detected. It did not consider the effect of ability distribution on the effect size using ANOVA. The current study set two ability distributions as equal with mean 0 and standard deviation 1 and mean 1 and standard deviation 2 for both the reference and focal groups. The independent variables in the current study were the ability distribution conditions and the dependent variables were the effect sizes and the number of DIF detections of each type. The current study also considered the effect of ability distribution for each DIF type.

### 2.2.5 Mantel-Haenszel DIF Detection and Test Length

The number of items that is used for DIF detection is of major concern in DIF research. No standard exists as to how many items should be used for DIF detection as these changes from study-to-study. In some cases it may be appropriate to test all items for DIF, whereas in others it may not be necessary (Güler & Penfield, 2009). If only certain items are suspected of DIF with adequate reasoning, then it may be more appropriate to test those items and not the entire set. However, often times it is difficult to simply assume which items may be problematic. For this reason, it is often recommended to simultaneously examine all test items for DIF. SIBTEST method tests items that are suspected to be DIF items while MH and considers all items to be DIF items simultaneously. This provides information about all items, shedding light on problematic items as well as those that function similarly for both the

reference and focal groups. With regard to statistical tests, some procedures such as IRT-Likelihood Ratio testing require the use of anchor items (Güler & Penfield, 2009). Some items are constrained to be equal across groups while items suspected of DIF are allowed to freely vary. In this instance, only a subset would be identified as DIF items while the rest would serve as a comparison group for DIF detection. Once DIF items are identified, the anchor items can also be analyzed by then constraining the original DIF items and allowing the original anchor items to freely vary. Thus it seems that testing all items simultaneously may be a more efficient procedure. However, as noted, depending on the procedure implemented different methods for selecting DIF items are used. The current study tested all items for DIF because the criterion for identifying anchor items or those suspected to be DIF items was not clear. The study also used the MH DIF method because it does not require anchor items that are common with the SIBTEST method but tests all items for DIF.

Apart from identifying the number of items being used in DIF detection, of additional importance is determining the number of items on the entire test or measure itself. The typical recommendation as noted by Zumbo (1999) is to have a minimum of 20 items. The reasoning for a minimum of 20 items directly relates to the formation of matching criteria. As noted earlier, a total test score is typically used as a method for matching individuals on ability. The total test score is divided up into normally 3-5 ability levels which are then used to match individuals on ability prior to DIF analysis procedures. Using a minimum of 20 items allows for greater variance in the score distribution which results in more meaningful ability level groups (Güler & Penfield, 2009). The recommendation of 20 items enabled the present study to use a test with items fewer than 20 considered as short tests and items more than 20 considered as long tests to determine the effect of test length on the DIF detection procedure. The short test lengths are based on the number of items that are often observed on personality

inventories (10-15 items). Although the psychometric properties of the instrument should have been assessed prior to being utilized, it is important that the validity and reliability of an instrument be adequate. Test items need to accurately tap into the construct of interest in order to derive meaningful ability level groups. Of course, one does not want to inflate reliability coefficients by simply adding redundant items. The key is to have a valid and reliable measure with sufficient items to develop meaningful matching groups. Gadermann et al. (2012), and John and Soto (2007) offer more information on modern approaches to structural validation and more precise and appropriate methods for assessing reliability.

A study by Khalid (2011) examined the power of MH procedure by varying the magnitude of DIF, Test length (40-80 items) and Sample size. It was found that the influence of test length was rather modest. The findings showed that the number of items do not greatly affect the detection of DIF of any kind by MH method. This study only used one DIF detection method using a test length of 40 and 80 items. This study did not consider using a test with few items and also did not compare the effect of test length using other DIF detection methods. The current study used the MH method which had long tests such as 50 items and short tests such as those with 10 and 20 items. It was not clear how the influence of test length was determined in this study. The current study used Analysis of variance to determine the effect of test length on the effect size of different DIF types and also statistical graphs to aid interpretation.

Fidalgo and Mellendam (2000) studied the effects of three amounts of DIF (10%, 15% and 30% of DIF-items), three test lengths (20, 40, and 60 items), and three test score (matching criterion) purification types (single-stage, two-stage, and iterative) on robustness and power of Mantel-Haenszel (MH) DIF detection procedures. Item response data were generated

under the three parameter logistic model (3PLM) for focal and reference group subjects, where the ability distributions of the two groups were equal. In the 10% DIF item conditions the three MH procedures are robust and have sufficient power, but in the 15% and 30% DIF item conditions robustness violation and insufficient powers occur. The influence of Test length on power of MH statistic was rather modest. On the other hand, test score purification improved power, but the size of their effects was much larger in the 15% and 30% DIF item conditions than in the 10% DIF item conditions. This study used three test lengths whose influence was rather modest. The study did not consider a test with few items probably less than 20 and it did not determine the effect of test length on a variable of interest. The present study determined the effect of test length (long and short tests) on the effect size and the number of DIF detections using ANOVA and statistical graphs by comparing DIF types using MH method.. Data was also generated under the two parameter logistic model (2PLM) unlike this study which considered the three parameter logistic model (3PLM) for focal and reference group subjects.

A study by Elusua and Wells (2013) was designed to detect DIF items using SIBTEST method. Polytomous item response data was generated using Samejima's GRM to represent a test with 15 items; each comprised of five categories eg likert-type items. The Test length of 15 items was selected according to the number of items that are often observed on personality inventories. The data was generated for two groups; a reference and a focal group. The findings indicated that no items displayed DIF when using the SIBTEST method. The study did not compare SIBTEST with other DIF detecting methods for detecting polytomous DIF items. The study did not compare different test lengths to determine the effect of test length on the DIF detection procedure. The findings of no DIF items for polytomous items was similar to those of dichotomous items when using short tests This data was for only 15

polytomous items unlike the current study that had 10, 30 and 50 dichotomous items and determined the effect of test length on the effect size and the number of DIF items detected. One test length could not enable one determine the influence of test length on the effect size and the number of DIF items detected. The study considered non cognitive tests commonly used in questionnaires which were not scored dichotomously. The current study compared dichotomously scored items with short and long tests using MH DIF method.

A study by Chahine and Childs (2010) used the grade 9 Assessment of mathematics student questionnaire data (n=153688) and corresponding Teacher questionnaire (n=4919) for the 2005/2006 school year obtained from EQAO. The teacher questionnaire contained 109 items exploring teachers' classroom practices. The student file contained 36 items 6 of which were scored dichotomously and 6 on a scale of 1-4. This study did not use an achievement test but items on student questionnaire and a corresponding Teacher Questionnaire which previous studies had been based on. DIF detection was done using SIBTEST method with gender being the basis for detecting DIF items. It was found that many DIF items were detected using both the teacher and the student questionnaires. The sample size was large and SIBTEST could detect DIF items for large sample sizes. The study did not consider DIF detection for small sample sizes and also for a test with few items. The study did not also compare various test lengths to determine the effect of test length on the procedure of DIF detection using non-cognitive tests. The current study compared the statistical power of MH in detecting DIF items under different test lengths.

### 2.3 Logistic Regression (LR) Procedure and Different conditions

Logistic regression is another approach commonly used to identify DIF (Swaminathan & Rogers, 1990). The Logistic regression procedure uses the item response (0 or 1) as the

dependent variable, with grouping variable (dummy coded as 1=reference, 2=focal), total scale score for each subject (characterized as variable TOT) and a group by TOT interaction as independent variables. This method provides a test of DIF conditionally on the relationship between the item response and the total scale score, testing the effects of group for uniform DIF, and the interaction of group and TOT to assess non-uniform DIF. Uniform DIF exists when there is no interaction between ability level and group membership. The presence of DIF in the LR approach is determined by testing the improvement in model fit that occurs when a term for group membership and a term for the interaction between test score and group membership are successively added to the regression model. A chi-square test is then used to evaluate the presence of uniform and non-uniform DIF on the item of interest by testing each term included in the model. The general model for logistic regression takes the form:

$$p(u = 1) = \frac{e^z}{1 + e^z}$$

where u is the score on the studied item. Performance on the studied item is first conditioned on the total test score. In this step, $z = \beta_0 + \beta_1 X$ where $X$ is the total test score (Model 1). This serves as the baseline model. The presence of uniform DIF is then tested by examining the improvement in chi-square model fit associated with adding a term for group membership ($G$) against the baseline model. That is, Model 2 (i.e. $z = \beta_0 + \beta_1 X + \beta_2 G$) subtracted from Model 1. The presence of non-uniform DIF is tested by examining the improvement in chi-square model fit associated with adding a term for group membership ($G$ ) and a term for the interaction between test score and group membership ( $XG$ ) against model 2. In other words, Model 3 (i.e. $z = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG$) subtracted from Model 2. Zumbo and Thomas (1996) developed an index to quantify the magnitude of DIF for the LR procedure based on partitioning a weighted least-squares estimate of $R^2$ that yields an Effect size measure. This

index is obtained, first, by computing the $R^2$ measure of fit DIF for each term in the LR model (i.e., test score, group membership, test score-by-group membership interaction) and then by partitioning the $R^2$ for each of the terms. A DIF Effect size for the group membership term is produced by subtracting the $R^2$ for the group membership term (Model 2) from the $R^2$ for the total test score term (Model 1).

The result is an Effect size measure associated with group membership that quantifies the magnitude of uniform DIF (herein called $R^2\Delta$ - U). A second DIF Effect size is produced for the total score-by-group membership term by subtracting the $R^2$ for the group membership interaction that quantifies the magnitude of non-uniform DIF (herein called $R^2\Delta$ - N). As with the MH Effect size measures, $R^2\Delta$ can be used with the LR significance test to identify items with DIF. Jodoin (1999) empirically established guidelines for interpreting $R^2\Delta$. An item has negligible or A-level DIF when the chi-square test for model fit is not statistically significant or when $R^2\Delta < 0.035$. An item has moderate or B-level DIF when the chi square test is statistically significant and when $0.035 \leq R^2\Delta < 0.070$. An item has large or C-level DIF when the chi-square test is statistically significant and when $R^2\Delta \geq 0.070$. Items with A level statistical ratings are considered unbiased, while, items falling into category C are inferred to have large DIF and therefore biased. These guidelines are applicable to both uniform and non-uniform DIF, and were used to classify DIF items in the current study.

The logistic regression procedure can be used with multiple examinee groups but not with polytomous item scores (Elosua & Wells, 2013). The LR procedure when used has not given clear perspective on the possible causes of DIF by inclusion of other relevant examinee characteristics such as sample size, ability distribution and test length. Just like the MH procedure, LR can be used in studies that require small sample sizes and non-normal data and also tests of different test lengths. LR procedures also use the total score as a proxy for latent

trait and this feature can be used to categorize the ability levels of the examinees. Within the logistic regression approach, leveraged values and outliers are of particular concern and must be examined prior to DIF detection. There was no substantial information about these points that suggested that they be removed. They did not involve special properties or circumstances not relevant to the situation under investigation. They did not involve possible measurement errors. Therefore there were no clear grounds for eliminating the outliers. The logistic regression was performed both with and without the outliers and their specific influence on the results was examined. This influence was minor and therefore it did not matter whether or not the outliers were omitted.

The current study used the total test score as a proxy for the latent trait. The total test score is a requirement for determining ability levels using the LR method. This method was therefore suitable for use in the present study whose effect size could be obtained using the statistical methods. The effect size was also quantified into type A, B and C. This method can also be used with non-normal data. Statistics such as LR violate the assumption of normality and can be used with both small and large sample sizes. The LR method was therefore suitable for use in the present study.

### 2.3.1 Logistic Regression and the Item Response Theory (IRT) Procedures

Although there is no single IRT method that can be used to detect DIF, all IRT procedures compare item characteristic curves (ICC) that are assumed to be invariant across groups after they have been rescaled. A general framework includes: (a) matching examines, (b) selecting an appropriate IRT model, (c) estimating item and examinee parameters for each group, (d) transforming estimates to a common scale, and (e) finding the DIF area by subtracting the reference and focal group's ICC from each other (Camilli & Shephard, 1994 or Clauser &

Mazor, 1988). Because item parameters are estimated separately for the focal and reference groups, they share different scales and cannot be compared directly. A common scale is needed. Scaling is possible on both item and ability parameters. Scaling is performed on the item difficulty parameter by constraining the mean and standard deviation to 0 and 1, respectively. This methodology is convenient for three unidimensional logistic models and the normal ogive model (Camilli, 2006). This process puts estimates on a common scale; however, they are constrained separately. Scaling on ability parameters by constraining the mean and variance to 0 and 1, respectively, does not provide a common scale for comparison and an additional transformation is required (Angoff, 1993).

Among the various IRT procedures, the area method is perhaps the easiest and also provides a test of significance. Raju's (1990) procedure was examined as an example of the AREA method. According to Raju (1990), the area between two ICCs can be found by subtracting the two ICCs from each other. Also, item mean and variance can be calculated for each item and later they can be used in hypothesis testing. Raju formulated mean and variance for both signed and unsigned areas for the one, two and three parameter logistic models. Although IRT provides a general framework for DIF analyses, it has some major drawbacks. All IRT methods require a large sample size and this increases the number of parameters that have to be estimated. IRT procedures, unlike the LR method, also require a considerable knowledge of IRT theory. Compared to LR, IRT is less practical and much more complex (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985). IRT methods do not quantify the amount of DIF into DIF types and only determines DIF in terms of area between ICCs of the focal and reference groups. The area between the ICCs may not be a good measure of the magnitude of DIF and also it does not show whether an item shows negligible moderate or large DIF. DIF methods such as LR that quantify the amount of DIF were therefore selected

for use in the current study. IRT methods are parametric methods and therefore cannot be used in a study which violates basic assumptions especially that of normality. The IRT method also assumes a certain latent trait or ability of the examinees. It can therefore not be used in studies such as LR that use the total test score as a criterion of determining the ability levels of the examinees. IRT and LR are both parametric tests but differ on the conditions under which they can be used.

### 2.3.2 Logistic Regression DIF Detection and Sample Size

An issue that pertains to sample size directly relates to the statistical procedure being used to detect DIF. Aside from sample size considerations of the reference and focal groups, certain characteristics of the sample itself must be met to comply with assumptions of each statistical test utilized in DIF detection (Erdem, 2014). For instance, using IRT approaches may require larger samples than required for the Logistic Regression procedure. This is important, as investigation of group size may direct one toward using one procedure over another. Additionally, as with all analyses, statistical test assumptions must be met. Some procedures are more robust to minor violations while others less so. Thus, the distributional nature of sample responses should be investigated prior to implementing any DIF procedures. For instance in a study that has small sample sizes, the LR statistic may be considered over SIBTEST and IRT based DIF methods that can detect DIF items for large sample sizes. This was the basis on which the LR procedure was selected for use in the current study because it is robust to both small and large sample sizes unlike SIBTEST and IRT methods which can only detect DIF with large sample sizes.

Simulation studies have demonstrated the utility of using the Likelihood Ratio Test for DIF. The LRT method is a parametric method and can be compared to the LR method in DIF detection under different sample size conditions. For example, a study by Clerk (2010)

simulated data for 30 items and had three conditions: a sample size of 300 in both the reference and focal group, 1000 in both groups, and 1000 in the reference group and 300 in the focal group. They also had a matched (reference and focal groups with the same mean ability level) and an unmatched (reference group has a higher mean ability level) condition for each of the above conditions. Their results indicated acceptable Type I error rates for all six combinations of sample sizes and ability matching conditions. They didn't assess the power of the Likelihood Ratio Test (LRT).

This study used a large sample size of 300 and 1000 in both the reference and focal groups. This therefore indicated that studies using small sample sizes are still limited. The study did not indicate the ability level that was being investigated. The DIF method that was used was parametric therefore the basic assumptions of parametric methods were not violated. The LR method was selected in the current study because it can detect DIF with small sample sizes. The study did not indicate the whether the effect of sample size depended on the DIF type. Studies that use small sample sizes are still limited. The current study therefore used small sample sizes and a large sample size and determined their effect on DIF detection of different DIF types using LR statistic.

A study by Ankenmann, Witt, and Dunbar (1999) simulated a 26-item mixed format test (20 items were dichotomous, six were polytomous) with three conditions: a sample size of 2000 in both the reference and focal group, 500 in each group, and 2000 in the reference group and 500 in the focal group. Like Clerk (2010), Ankenmann et al. had a matched and an unmatched condition for each sample size pairing. In addition, Ankenmann et al. manipulated the a and b parameters for one of the groups to introduce DIF. Ankenmann et al. found the

LRT had acceptable Type I error rates across most of the simulated conditions but that the test may lack power when sample sizes are around 500. This study compared polytomous and dichotomous items under different sample size conditions, but it did not compare different DIF detection methods. It also used the 2PL model to introduce DIF. The current study used only dichotomous items that are required for LR DIF procedures under large and small sample sizes, for both the reference and focal groups. Dichotomous items were generated by the statistical software used. As noted earlier LRT could not be used in the current study because it is a parametric method and can only detect DIF for large sample sizes.

Among the advantages of the LR procedure is the sample size requirement. In a study by Kubiak and Colwell (1990) for the ETS testing programs, a sample size of 500 for the groups combined and a minimum of 100 for the focal group were considered adequate for purposes of test assembly. Hills (1989) suggested that samples as small as 200 for the combined group, with a minimum of 100 in each group, are adequate for screening purposes. Although the most popular method in recent literature are those based on item response theory (IRT) and chi-square distributions, most notably the LR statistic, the use of these procedures generally has failed to produce meaningful interpretations of bias ((Zwick & Ercikan, 1989). Mazor, Clauser, and Hambleton (1992) reported relatively good DIF detection results using the LR procedure with sample sizes as small as 100 in the reference and focal groups. Similar results were obtained by Parshall and Miller (1995); Fidalgo, Mellenbergh, and Muñiz (1998); and Muñiz, Hambleton, and Xing (2001). However, the sample size requirements of the LR procedure are far lower than those of the item response theory (IRT)–based methods for DIF detection (Camilli & Shepard, 1994; Clauser & Mazor, 1998; Millsap & Everson, 1993; Penfield &Lam, 2000). Since LR method can be used by both small and large sample sizes, it was therefore considered for use in the current study.

### 2.3.3 Logistic Regression DIF Detection and Ability Distribution

Studies have found that differences in Ability Distributions, sometimes referred to as *impact*, affect DIF detection rates (French & Maller, 2007: Wang & Su, 2004). Ability Distribution was assessed in terms of mean and standard deviation.

In their study, French & Maller 2007 simulated mean latent trait differences (μd) between groups. Members of the reference group were generated from *N* (0, 1). Members of the focal group were generated from *N* (0, 1), *N* (-0.5, 1), or *N* (-1, 1). The mean latent trait differences between groups μd = μ*R* – μ*F*, where μ*R* and μ*F* were the mean latent traits of the reference and focal groups, respectively were determined. Consequently, there were three levels of μ*d*: 0, 0.5, and 1. The study found out that a difference in mean ability of 1 standard deviation between certain reference and focal groups occurred frequently in real testing situations between groups (French & Maller, 2007). This factor was varied only in the Type I error portion of this study. Ability distribution in this study was given in terms of difference in mean ability. However the current study used mean and standard deviation to indicate the Ability distribution condition when generating data.

Roussos and Stout (1996) analyzed DIF detection while setting the ability distributions as normal, for both focal and reference groups with a variance of one, but with varying means. The differences in means used were 0.0, 0.5, and 1.0. Roussos and Stout (1996) chose these amounts of mean differences based on an examination of real data and discussions with test data specialists. This study used real data and only varied the means. The study did not use simulated data and did not give ability distribution in terms of mean and standard deviation. The current study used simulated data generated using computer software and used mean and standard deviation to generate the data. The software allowed one to set the ability

distribution in terms of mean and standard deviation. The study used a normal ability distribution, mean 0 and standard deviation 1, and that of mean 1 and standard deviation 2. Unlike the study by Roussos and Stout (1996), the current study investigated the effect of Ability distribution on the detection of DIF by comparing two DIF detection methods namely Mantel-Haenszel and Logistic Regression statistics. Studies have been done to detect the type of DIF but those that determine the effect of Ability distribution on the type of DIF are still limited. Roussos and Stout (1996) reported that a slight tendency toward increasing Type I error with increasing Sample size and increasing *d*T [difference in ability distribution means], with MH seeming to have very slightly lower Type I error rates for *d*T >0.

Jodoin and Gierl (2002) simulated data for equal Ability Distributions and unequal Ability Distributions under small sample sizes (250/250) to larger sample sizes (1000/1000). When comparing the Type I error and power rates between equal and unequal Ability Distributions rates were slightly (i.e., 0.2 points to 4.0 points) higher for unequal Ability Distributions for smaller samples, but evened out as the sample sizes increased. Jodion and Gierl (2002) set the unequal Ability Distributions with a difference of .50 for the means of the reference and focal group with the same standard deviation. In real testing situations the total test score is used as a criterion for determining the ability levels of the examinees. The smaller difference in this study might attribute to not finding differences in Type I error and power rates for the larger sample sizes. The current study compared a much smaller sample sizes and larger sample size and set the ability levels using means and standard deviations unlike this study which used mean differences between the focal and reference groups. The study used unequal ability distributions but did not indicate how the ability distributions were set. The current study also set ability distributions with a mean of 0 and standard deviation 1, and mean 1 and standard deviation 2.

### 2.3.4 Logistic Regression DIF Detection and Test Length

Determining the number of items on the entire test or measure itself is of major importance in DIF research. A typical recommendation as noted by Zumbo (1999) is to have a minimum of 20 items. The reasoning for a minimum of 20 items directly relates to the formation of matching criteria. As noted earlier, a total test score is typically used as a method for matching individuals on ability. The total test score is divided up into normally 3-5 ability levels which are then used to match individuals on ability prior to DIF analysis procedures. Using a minimum of 20 items allows for greater variance in the score distribution which results in more meaningful ability level groups (Güler & Penfield, 2009). The recommendation of 20 items enabled the present study to use a test with items fewer than 20 considered as short tests and items more than 20 considered as long tests to determine the effect of test length on the DIF detection procedure. The short test lengths are based on the number of items that are often observed on personality inventories (10-15 items). Although the psychometric properties of the instrument should have been assessed prior to being utilized, it is important that the validity and reliability of an instrument be adequate. Test items need to accurately tap into the construct of interest in order to derive meaningful ability level groups. Of course, one does not want to inflate reliability coefficients by simply adding redundant items. The key is to have a valid and reliable measure with sufficient items to develop meaningful matching groups. Gadermann et al. (2012), and John and Soto (2007) offer more information on modern approaches to structural validation and more precise and appropriate methods for assessing reliability.

Lau and Arce (2011) used the Monte Carlo simulation technique to compare different detecting methods for anchor item stability. The simulated test form contained 30 dichotomous items (score points 0, 1) and 30 polytomous items (score points 0, 1, 2...). A set

of 18 common items including 9 dichotomous and 9 polytomous were included in the internal anchor item design. Among the 18 common items, 4 of them contained DIF including two dichotomous and two polytomous items. For the two polytomous common items with DIF, one had net DIF and the other had global DIF. Computer program WINGEN3 (Han 2010) was used for simulating the response data. This study analyzed both polytomous and dichotomous items unlike previous studies which used either polytomous or dichotomous items. The study did not compare different DIF detection procedures to determine whether DIF detection was affected by the DIF procedure used. The study did not state the conditions such as different test lengths, under which the data was generated. The current study also used the WINGEN 3 software to generate dichotomously scored data. The software enabled the user to simulate different test lengths (Han & Hambleton, 2007). The current study used the MH DIF method and determined the effect of test length on the DIF detection procedure using inferential statistics such as ANOVA and statistical graphs.

In a related study Lopez-Rivas (2012) used the three-parameter logistic model (3PLM: Birnbaum, 1968) for response data generation because it has been shown to fit cognitive ability data well in a multitude of studies. To generate item response data using the 3-PLM trait scores and item parameters are needed. Trait scores for examinees in the focal and reference groups of various sizes were obtained by sampling values from independent normal distributions. Tests consisting of 15 and 30 items were created for the Monte Carlo study by randomly selecting item parameters from tables published by Narayanan and Swaminathan (1996) which showed item calibration results for an administration of the graduate Management admissions Test. A study by Truancy (2009) was obtained from data obtained from the reading achievement tests administered by the Turkish and American students who participated in the PIRLS-2001 (the Progress in International Reading Literacy 2001) by the

International Association for the evaluation of Educational Achievement (IEA). Within the scope of the study the data on the MICE reading tests – one of the reading literacy tests if PIRLS were used. The MICE reading tests according to Truancy (2009) consisted of 14 questions. 7 of these items ie 1,2,3,5,8,9and 13, were multiple choice questions scored as 1-10, whereas the $6^{th}$ and $12^{th}$ items required long answers. The first was scored as 0-1-2 and the second was scored as 0-1-2-3 through partial scoring. It was examined whether the application of the MICE reading and its items as part of the PITLS project in Turkey and USA display differential item functioning, across cultures using parameter comparison method. The findings indicated that no DIF items were detected across cultures in Turkey and USA. The effect of test length on the number of DIF items was not determined using inferential statistics by comparing different test lengths. The study used real data with the focal and reference groups based on cultures. This research compared a test consisting of 14 items with both multiple choice and long answer questions. The scoring criteria were also different from the current study which used a dichotomous scale of 0-1. The current study used simulated data under different test lengths and determined the effect of test length on the effect size and the number of DIF items detected.

## 2.4 Mantel-Haenszel verses Logistic Regression and Different conditions

Although a number of methods are available for detecting DIF items, the statistical conditions under which they operate differ from one method to another. Methods that have some similar statistical properties can be compared when conducting DIF studies. For instance Mantel-Haenszel and Logistic Regression can be used with small sample sizes. IRT methods and SIBTEST are robust to large sample sizes. The use of small sample sizes is still a major concern to DIF researchers and has not been done exhaustively. Classroom settings can have students as low as 20 and 60 while a school can have students as high as 1000. A study by

Salubayaba (2013) compared small and large sample sizes using SIBTEST method and found no DIF items detected for small sample sizes. The method used was not robust to small sample sizes and therefore was not suitable for studying small sample sizes. Since MH and LR methods can be used for small sample sizes they were compared in the current study to determine their statistical power under different conditions (Zwick & Ercikan, 1989). A number of DIF detection studies have been done but rarely have these analyses compared the power of DIF methods under different sample sizes especially small and large sample sizes (Cromwell, 2006).

A study by Swaminathan and Rodgers (1993) compared Logistic Regression and Mantel-Haenszel procedures for detecting uniform and non-uniform DIF in a simulation study which also examined their distributional properties. For the distributional properties, two factors were selected: sample size and degree of model data fit. Two levels of model data fit namely good fit and poor fit were crossed with two levels of sample size; 250 per group and 500 per group. For each combination of sample size and model data fit, 100 replications were done. Five items were used to calculate MH and LR statistics and to construct empirical sampling distributions. The Kolmogorov-Smirnov test was performed to determine if the test statistics had the expected distributions. The results showed that model-data fit did not affect the results for LR procedure but had a significant effect for the MH procedure. Sample size had a strong effect on the detection rates of both procedures. The LR and MH procedures were almost equally effective in detecting uniform DIF while the MH procedure was ineffective in detecting non-uniform DIF.

This study compared the performance of MH and LR DIF in detecting uniform and non-uniform DIF under model data fit conditions and sample size. It did not however look at the effect of the different conditions on the effect size and the number of DIF detections under

uniform and non uniform conditions. The current study compared MH and LR DIF methods under different sample size conditions and determined the effect of these conditions on the effect size and number of DIF items. The current study also used graphs to compare the effect of these conditions on the number of DIF detections.

DIF detection studies consider examinees of similar ability who respond differently to a particular item due to group membership. However most DIF detection studies rarely indicate the ability level of the examinees. Studies comparing DIF detection methods at different ability distributions are limited. A study by Awuor (2008) on the effect of unequal sample sizes on the power of DIF detection compared SIBTEST and Mantel-Haenszel procedures. In the simulation study the ability distribution in terms of mean and standard deviation was not stated and remained unknown. Also the study compared two DIF methods namely SIBTEST and MH. SIBTEST could only be used with large sample sizes while MH can be used with both small and large sample sizes. Also SIBTEST requires the use of only items identified as DIF items while MH considers all items as DIF items. Although both methods are non-parametric methods, their statistical properties are quite different. It may not be appropriate to compare them when detecting DIF items especially where small sample sizes are required. MH and LR DIF methods were compared in the current study because they are robust to small sample sizes and can be used with dichotomously scored items. Both methods can detect uniform DIF even though MH method can also detect non-uniform DIF. In the current study only uniform DIF was considered.

Studies that compared DIF statistics for detecting DIF items have not used a strictly statistical decision making framework. A study by Cromwell (2006) compared the use of effect size for LR and MH methods for predicting DIF. In this method a statistical decision making

framework such as analyzing the effect of sample size and test length using inferential statistics such as Analysis of variance (ANOVA) and statistical graphs were not used. Inferential statistics are quite vital when comparing the statistical power of DIF methods such as MH and LR under certain examinee conditions such as sample size, ability distribution and test length. The current study used inferential statistics such as ANOVA to determine the effect of sample size, ability distribution and test length on the effect size and number of DIF detections.

Most studies on DIF detection have used only one DIF detection method to detect DIF items (Khalid, 2011; Fidalgo & Mellenbergh, 2000). Those that have compared two or more methods are still inadequate. SIBTEST method was used in studies by Salubyaba (2013) and Gierl, Gotzman and Boughton (2004). They did not manipulate DIF conditions such as sample size and test lengths. Some studies that have also compared DIF methods with different statistical properties. Such comparison may lead to inconsistent findings. Methods that had similar statistical properties such as MH and LR were selected for use in the current study. These methods can be used with small sample sizes and non-normal data.

The quality of a procedure is assessed in terms of Type I error rate and Type II error rate. In a DIF detection study, Type II error possess a critical challenge to the validity of test scores that psychometricians have to investigate and find techniques that can offer the best solution to the error rate inflation. Simulation studies allow researchers to determine with certainty which techniques are better than others and in what conditions. However, when empirical data are analyzed, it may not be possible to determine whether an item identified with DIF is a correct detection or a false positive and how many items with DIF have failed to be identified. (Fidalgo, Ferreres, & Muňiz, 2004). Fidalgo et al., (2004) calculated the SIBTEST statistics in two stages. In the first stage they analyzed each item for DIF with the rest of the

items forming the matching subtest, as is normally done with M-H procedure. In the second stage they conducted standard single-item DIF analyses using items not identified as DIF in the first stage as the matching subtest. In both procedures they tested the two-tailed hypothesis of DIF against each group at two significant levels (.05 and .01). They concluded that the quality of a statistical test can be assessed by its robustness and power. It is known that a statistical power is robust if its probability of a Type I error, is approximately equal to the normal significance level. This study flagged items as having DIF or not having DIF. The current study considered all items as DIF items and did not flag any item as having DIF or not having DIF.

According to Kristjansson et al (2005), the variance of estimated parameters is reduced through the replication of data. The number of replications is influenced by the purpose of the study, the desire to minimize the sampling variance of the estimated parameters and by the need for statistical test results to have adequate power to detect the effects of interest. Using the same seed value in the generation of the varying sample sizes helps to minimize the effect of random error on parameter estimates. Studies have reported that the number of replications has a direct influence on the precision (accuracy) of the estimated parameters. The studies have also reported that more replications produce parameter estimate with less sampling variance (Hays & Olds, 1992). While estimated other parameters may require large number of replications, Kristjansson et al (2005) proposed that when comparing the number of DIF items correctly detected, a small number of replications such as 10 may be sufficient. This proposal of 10 replications was debatable because it was argued that there were many variables that influenced the statistical power of a DIF detection procedure.

Several studies have stressed how crucial the number of replications is in a study, because the results of such studies may vary depending on the number of replications. However using no replications or a very small number of replications may result in sampling variance that is large enough to seriously bias the parameters being estimated (Hambleton, Jones & Rodgers, 1993). These studies do not suggest the number of replications that are sufficient for DIF detection. Although previous simulation studies (Kristjansson et al., 2005; Rogers & Swaminathan, 1993; Su & Wang, 2005; Wang & Su, 2004) employed 100 replications, in the current study, one thousand replications was completed for each condition to ensure the accuracy of the empirical estimations of the sampling distribution characteristics and to produce parameter estimates with less sampling variance. Since there was no criterion for selecting the number of replications and the more the number of replications the less the sampling variance, one thousand replications were chosen for the current study. This was to ensure that the sampling variance was reduced as much as possible.

In psychological research and psychometric evaluation, statistics play a vital role but should by no means be the sole basis for decisions and conclusions reached. Reasoned judgment is of critical importance when evaluating items for DIF (Walker, 2011). For instance, depending on the statistical procedure used for DIF detection, differing results may be yielded. Some procedures are more precise while others less so. For instance, the Mantel-Haenszel procedure requires the researcher to construct ability levels based on total test scores whereas IRT more effectively places individuals along the latent trait or ability continuum (Walker, 2011). Thus, one procedure may indicate DIF for certain items while others do not. Another issue is that sometimes DIF may be indicated but there is no clear reason why DIF exists. This is where reasoned judgment comes into play. The researcher can use common sense to derive meaning from DIF analyses. It is not enough to report that items function differently

for groups, there needs to be a theoretical reason for why it occurs. Furthermore, evidence of DIF does not directly translate into unfairness in the test (Edelen & Reeve, 2007). It is common in DIF studies to identify some items that suggest DIF. This may be an indication of problematic items that need to be revised or omitted and not necessarily an indication of an unfair test. Therefore, DIF analysis can be considered a useful tool for item analysis but is more effective when combined with theoretical reasoning.

Examination items on which one group of test-takers performs significantly better than another group (Roever, 2005) can be identified. DIF is a collection of statistical methods utilized to determine if examination items are appropriate and fair for testing the knowledge of different groups of examinees (e.g., male vs. female Caucasian vs. African-American (Schumacher, 2005). DIF detection procedures, such as MH and LR, can therefore be used to identify biased items under different conditions of Sample size, Ability distribution and Test length. As such, DIF aids in the identification of test items that are potentially biased. In assessing response patterns, the comparison groups are initially matched on the underlying construct of interest (e.g., verbal ability or mathematics achievement). By matching groups on the measured variable, researchers/test developers are better able to determine whether item responses are equally valid for distinct groups of test-takers (Zumbo, 1999).

DIF methods therefore assess the test-takers' response patterns to specific test items. DIF occurs when a statistically significant difference is evident in the probability that test-takers from the two distinct groups who have the same underlying ability on the measured construct, demonstrate differing probabilities of correctly answering the item (Zumbo, 1999). As stated, examinees' ability 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 analyses of the other test items (Zumbo, 1999).

A test item is considered to be biased when a dimension on the test is deemed to be irrelevant to the construct that is being measured, placing one group of examinees at a disadvantage in taking the test (Hambleton & Rodgers, 1995). Thus, if DIF is *not* evident for an item, then there is no item bias. DIF is required but is not sufficient for item bias. That is, if DIF is apparent, then its presence is not sufficient to declare item bias. An item might show DIF, but not be considered biased if the difference is a result of the actual difference in the groups' ability to respond to the item for instance, if one group of test-takers is at a high level and the other is at a low level, the lower group would perform significantly lower; (Roever, 2005). If test-takers differed in knowledge, a difference in item responses would be expected. Consequently, a difference in the performance of groups of examinees with different abilities on specific items is not indicative of test bias, but rather of item impact (Schumacher, 2005).

Upon seeing evidence for the occurrence of DIF, one would need to apply subsequent item-bias analyses (e.g., empirical evaluation or content analysis) in order to determine if item biasness is present (Zumbo, 1999). 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. In items exhibiting test bias, an additional construct is evident, apart from the construct that the items are supposed to measure (Roever, 2005).

A study by Parshall & Miller (2004) identified biased test items through Differential Item Functioning analysis using four contingency table approaches: Chi-Square, Distracter Response Analysis, Logistic Regression, and Mantel-Haenszel Statistic. The study made use

of test scores of 200 junior high school students. One hundred students came from a public school, and the other 100 were private school examinees. One hundred students were males and 100 were females. Basing from their English II grades, 95 students were classified as low ability and 105 as high ability students. A researcher-constructed and validated Chemistry Achievement Test was used as research instrument. The results from the four methods used were compared, and it was found that school type, gender, and English ability bias existed. There was a high degree of agreement between the Logistic Regression and the Mantel-Haenszel statistics in identifying biased test items. In this study the test length and ability distribution of the examinees used was unknown. It was not stated whether the items were polytomous or dichotomous. While this study used real data with a small Sample size of 100 respondents, the current study used simulated data generated using computer software and whose sample size and ability distribution could be varied. The current study determined DIF of Type A, B and C to represent small medium and large DIF, and compared them using MH and LR statistics.

A study by Adedoyin (2010) using IRT approach to detect gender biased items in public examinations attempted to detect gender bias test items from the Botswana Junior Certificate Examination in Mathematics. To detect gender bias test items, a randomly selected sample of 4000 students response to mathematics paper 1 of the Botswana Junior Certificate examination were selected from 36,000 students who sat for the examination. Out of which 2,000 were males and 2000 were females. The examination paper consisted of 38 test items. To detect the gender biased test items, the study used 3PL (Multilog software) item response theory (IRT) statistical analysis. This generated the item characteristics curves (ICC) for the two groups (male/female). The study compared the results generated from the ICC curves for the male and female groups, and found that, out of 16 test items that fitted the 3PL item

response theory (IRT) statistical analysis, 5 items were gender biased. The current study used two methods namely LR and MH and compared their ability to detect biased items under different conditions. This study also used real data whereas the current study used simulated data generated using computer software. Simulation studies have been developed to better understand the statistical power of many DIF detection methods. Researchers can alter such factors as Sample size, latent Ability distribution, Test length, Effect size, and DIF pattern to better understand how they affect statistical power of different DIF detection methods.

The studies that have been reviewed have shown that different conditions have an effect on the detection of DIF using different methods. The selection of a method for DIF detection is of major concern since different methods have different statistical properties and operate under different conditions. This has informed the current study to use Mantel-Haenszel and Logistic Regression DIF methods since they can provide effect sizes, and can also be used with various sample sizes. The literature has also informed the current study to use simulation in order to vary the conditions and study their effect on the variable of interest.

# CHAPTER THREE

# RESEARCH METHODOLOGY

## 3.1 Introduction

This chapter presents the research methodology of the study. The chapter is divided into eight sections. Section two presents the research design and description of variables. Section three presents the population. Section four presents the sample size and sampling technique. Section five presents the instruments for data collection. Section six presents the validity and reliability of the study. Section seven presents the data collection procedure. Section eight presents the methods of data analysis.

## 3.2 Research Design

A factorial research design was used in this study. According to Kerlinger (1986), in a factorial design the researcher can modify certain factors and observe the effects of these modifications on the variable of interest. This design enables the researcher to study the effects of two or more independent variables at the same time. It allows one to determine whether the effect of one independent variable depends on the level of other independent variables. In a factorial design, all levels of each independent variable are combined with all levels of the other independent variables to produce all possible conditions.

This design enabled the researcher to simulate samples for three Sample size conditions, three types of Test lengths and two types of Ability Distribution conditions. This resulted into a 3x3x2 factorial design giving 18 data sets. The design was therefore used to investigate the relative accuracy of Mantel-Haenszel and Logistic Regression statistics, in DIF detection for a fifty, thirty and ten item dichotomous test, a sample size of twenty, sixty and one thousand examinees, and ability distribution mean 0, standard deviation 1 and mean 1 standard

deviation 2. This design was therefore appropriate for partialing out variances due to various factors. An important advantage of the simulation methodology is that it allows the investigator to manipulate design conditions that would otherwise be impossible to include in a study.

The independent variables in the present study were Sample size, Ability distribution, and Test length. This was due to the effect these variables can have on the power of DIF procedures. These conditions were selected to simulate realistic data samples. The dependent variables were the weighted-least squares $R^2$, Effect size for Logistic Regression DIF, the log odds ratio, Effect size for Mantel-Haenszel DIF statistics and the number of detections of DIF items across 3 DIF Types.

### 3.3 Area of Study

This was a simulation study that involved the generation of data using computer software. The researcher did not go the field in any particular area to collect the required data. This study therefore was not done in a particular geographical area. The map of the area of study was therefore not provided. The study was computer based and therefore did not constitute any particular area of study.

### 3.4 Population

The population of the study consisted of 2000 examinee responses. It consisted of responses from the reference group and the focal group. Though the focal group represents the minority examinees, this study had an equal number of 1000 examinee responses for the reference and 1000 for the focal group. These were used for simulation of DIF conditions and to ensure similarity to real data.

### 3.5 Sample Size and Sampling Technique

A stratified random sampling technique was used to select the sample from a pool of 2000 examinee responses. The stratifying criterion was based on the examinee responses designated as reference and focal. A balanced sample was therefore drawn for the reference and the focal groups. The reference group and focal group had three sample sizes: 20, 60, and 1000 each. These were used to establish three sample size conditions namely small sample sizes [($20_r/20_f$), moderate sample sizes ($60_r/60_f$)], and large sample sizes ($1000_r/1000_f$). These gave a realistic proportion to sample sizes between reference groups and focal groups of $100\%_r$, $100\%_f$ (large sample size), and [$2\%_r$, $2\%_{f;}$ $6\%_r$, $6\%_f$] (small and moderate sample sizes), respectively. The sample sizes gave a reflection the examinees in a classroom (small and moderate) and those in a school (large sample size).

The sample size combinations included combinations of three balanced sample sizes for the reference and focal groups to model a variety of research situations. The sample sizes were selected because some studies (e.g. Ankenmann, Witt, & Dunbar, 1999) had reported Type I error inflation in DIF detection with sample sizes as large as 500/500 with Likelihood Ratio Goodness of Fit Statistics (LR) while other studies (e.g., Roussos & Stout, 1996) did not report any significant Type I error inflation with as small sample as 100 for both the reference and the focal groups when SIBTEST and M-H were used with uniform DIF and identical ability distributions.

### 3.6 Instruments for Data Collection

WinGen3 statistical software was used to generate dichotomous item response data for many conditions that arose in practice (Han, 2007). The computer screenshot of the software is shown in Appendix A. The main window consisted of examinee characteristics which included the number of examinees and the Ability distribution in terms of mean and standard

deviation. The mean and standard deviation were used to model two equal Ability distribution conditions. The first condition was set with means of 0.0 for the focal and reference groups and standard deviations 1.0 for focal and reference groups. The second condition was also set with the focal group having a mean of 1.0 and a standard deviation of 2.0 and the reference group a mean of 1.0 and a standard deviation of 2.0.

The software also consisted of item characteristics which included the number of items, the number of response categories, the model to be used i.e. 1PLM, 2PLM, 3PLM or non-parametric. The data was binary therefore two response categories were selected. The distribution in terms of parameter a, b and c was then selected. When appropriate entries were made, true scores and true item parameters were then generated. Replication data sets and response data sets were also generated. The software allowed examinee graphs and Item graphs to be displayed. The DIF/IPD window consisted of introduction to DIF/Item parameter drift via the direct input mode or the multiple file read in mode. This consisted of data files for the reference group/test 1 and focal group's later tests. The examinee and item parameter output files were saved in the comma separated values (csv) extension.

The software generated model parameter values from various distributions for realistic data. It generated item parameters to create realistic item response data from various kinds of distributions. The user generated binary response data representing examinee responses on a test. The user chose various test lengths for the examinees so that the researcher conducted the study with more realistic datasets. The tests had 10 items, 30 items and 50 items respectively. The test length of 10 items was selected according to the number of items that are often observed on personality inventories (10-15 items) Achievement tests and questionnaires can have over 20 items. It was also be used to vary the ability distribution of

the data. A good random number generator must have specified statistical properties namely, adequate period length, ease of implementation, efficiency, portability and reproducibility. This software was suitable because it could be used to generate data with all the different conditions that were used in the study. This included the two Ability distribution conditions, the number of items and the sample size. The software also allowed the data to be replicated any number of times. For the purposes of this study 1000 replications were performed.

## 3.7 Validity and Reliability

A pilot study was carried out to enhance the instrument's validity and reliability. The objectives for the pilot study were to try out the process of data generation and DIF detection with LR and MH DIF methods. The pilot study also checked the adequacy of the computer software in carrying out data simulation. For the pilot study, ten dichotomously scored items were generated randomly using the computer software. Item response data for 20 examinees were generated. The data was not replicated prior to analysis. For details refer to Appendix C

### 3.7.1 Validity

The software was assessed by experts from the department of Educational Psychology and the department of Pure and Applied Mathematics at Maseno University to enhance its face validity.

### 3.7.2 Reliability

Another purpose of the pilot study was to obtain the reliability estimate of the generated data. Since the data was simulated the test-retest method could not be used in this study. The data obtained was dichotomous; hence the Kuder-Richardson-20 (KR-20) method was used to estimate the reliability coefficient. This is a measure of internal consistency reliability for

measures with dichotomous choices. A reliability coefficient of .75 was obtained. For details refer to Appendix C.

### 3.8 Data Collection Procedure

The binary response data was generated with WinGen3 (Han, 2007) computer software. The obtained data represented responses on tests with varied length. The data was replicated up to 1,000 times for every cell in the study, resulting into 18,000 data sets. Replication was necessary to reduce chances of sampling error that would likely result in variance large enough to have a confounding effect. Although previous simulation studies (Kristjansson et al., 2005; Rogers & Swaminathan, 1993; Su & Wang, 2005; Wang & Su, 2004) employed 100 replications, in this investigation, one thousand replications were used for each condition to ensure the accuracy of the empirical estimations of the sampling distribution characteristics. In each replication, new item scores for the 20, 60, and 1000 examinees in each group were generated. Once the item scores were generated the Effect size was calculated for the studied item. The average value of the Effect sizes across the 1000 replications was calculated. Permission to undertake the study was obtained from Maseno University Research Ethics Committee (MUERC).

### 3.9 Methods of Data Analysis

Statistical analysis was done on the Mantel-Haenszel DIF method with the dependent variable as the Log odds ratio Effect size for MH DIF, then the Logistic Regression DIF method and the dependent variable was the weighted-least-squares $R^2$ Effect size for LR DIF. The independent variables were the Sample sizes, Test length and Ability distribution. Analysis was done on the raw data in order to obtain the Effect sizes for both MH and LR methods. For the MH method analysis was done using a routine was written, according to the

MH formulae, which gave the Effect size for MH analysis. The procedure was also repeated for 1000 replications and the average Effect size values were determined. The number of items displaying various categories of DIF were then determined.

For LR method, the Statistical Package for Social Sciences (SPSS) (IBM SPSS Version 20) was used for analysis. Analysis was done using the General Linear model, multivariate analysis which gave $R^2$ values for model 1 and model 2. The $R^2$ values were then entered into coding sheets on MS Excel worksheet to obtain the Effect size, $R^2\Delta$ which was the difference between $R^2$ values for model 1 and model 2. The procedure was repeated for 1000 replications and the average Effect size value was determined. The number of items displaying various categories of DIF were then determined for each category of Test length. Descriptive statistics such as the mean was then used to obtain the mean Effect sizes (ES) and number of detections of DIF items across 3 DIF Types; A, B and C using Mantel-Haenszel and Logistic Regression statistics.

Prior to this analysis, it was necessary to identify some outlying points which may have had more influence on the regression than others. The Statistical Package for Social Sciences (SPSS) (IBM SPSS Version 20) was used to identify possible influential outliers known as *Cook's Distance* or simply *Cook's D*. This was described as follows;

Given a regression of $Y$ on $(x_1,..,x_k)$ using data set $(y_j, x_{1j},..,x_{kj})$, j=1,..,$n$, If

$S$ = estimated root mean square error,

$\hat{y}_j$ = regression estimate of the conditional mean $E(Y_j|x_{1j},..,x_{kj})$ ,

$\hat{y}_j(i)$ = regression estimate of the conditional mean $E(Y_j|x_{1j},..,x_{kj})$ , with the

$i^{th}$ data point $(y_i, x_{1i}, .., x_{ki})$ removed,

then *Cook's Distance* for point *i* is given by

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - y_i)^2}{(k+1)s^2}, \; i = 1,..,n$$

Intuitively, $D_i$ is a normalized measure of the influence of point *i* on all predicted mean values, $\hat{y}_j$ , j=1,..,*n*. *Cook's D* was obtained using **Fit model** on SPSS as follows:

i) Right click on heading of the **Parameter Estimates** table,

ii) Select the **Save Columns** options and click in the **Cook's D Influence**.

iii) A new data column will appear, containing the Cook's D Influence values.

To identify potential outliers, one *Rule of Thumb* was to treat point *i* as an outlier when:

$$D_i \geq \frac{4}{n-(k+1)}$$

As with all *Rules of Thumb*, this provided only a rough guideline and often tends to identify too many points as potential outliers. The best strategy was to look at the distribution of Cook's D values and see whether there were any conspicuously large values relative to others. If these values were roughly of magnitude $4/(n-k-1)$ or larger, then they were worth investigating further.

This procedure only served to identify points that were suspicious from a statistical perspective. It did not mean that these points were to be automatically eliminated. The removal of data points can be dangerous. While it will always improve the "fit" of a regression, it may end up destroying some of the most important information in the data.

The question that was addressed was whether there existed some substantial information about these points that suggested that they be removed. Did they involve special properties or circumstances not relevant to the situation under investigation? Did they involve possible measurement errors? Such distinguishing features were not found therefore there were no clear grounds for eliminating the outliers.

The logistic regression was performed both with and without the outliers and their specific influence on the results was examined. This influence was minor and therefore it did not matter whether or not the outliers were omitted.

Statistical Package for Social Sciences (SPSS) (IBM SPSS Version 20) was used to perform One Way Analysis of Variance (ANOVA) in order to determine the effect of Sample Size, Ability Distribution and Test Length on the Effect Size (ES) of DIF across three types of DIF; A, B and C for both MH and LR methods and also used to draw line graphs. ANOVA uses the concept of F-distribution. This sampling distribution can be used to test hypotheses about two or more population variance. The most common use of the F-ratio is testing hypotheses regarding equality of two or more means. One way ANOVA was therefore used to test if there were significant differences between the Effect sizes under varied conditions. The level of significance used was 0.05 with 2, 15 degrees of freedom for the main effects of Sample size and Test length and 1, 16 degrees of freedom for the main effects of Ability distribution conditions. The dependent measures were the mean Effect size (ES) values for Mantel-Haenszel and Logistic Regression statistics. The F-ratio in one way ANOVA provided a test of the null hypothesis that two or more population means were equal. Post-hoc Bonferroni statistic was used to compare the difference among the means where the difference was statistically significant.

Line graphs for mean Effect size against Test length across DIF types and for each level of Ability distribution and Sample size were constructed to aid interpretation. A similar display for the mean number of items across various categories of DIF was constructed. This was to address the objectives of determining the effect of Sample Size, Ability Distribution and Test Length on the Effect Size and the number of detections of DIF items across 3 DIF Types; A, B and C using MH and LR statistics.

### 3.10 Ethical Considerations

The current study is a simulation study that used simulated data rather than real data. Real subjects were therefore not used in the study. The researcher did not go to the field to collect data from real subjects, but used WINGEN 3 computer software to generate the required data. The software met all the conditions required in generating data. The data was replicated 1000 times in every cell in order to reduce chances of sampling error that would result in variance large enough to have a confounding effect. The data was stored on Ms Exell operating system computer work sheets prior to analysis.

The amount of data involved in the study was large therefore computer software such as Statistical Package for Social Sciences was used in the analysis of the data. The SPSS was easy to use unlike other methods and it had enough statistical power to carry out data analysis. The findings of the study were presented using tables showing Effect sizes of various categories, the number of items displaying various DIF categories and graphs comparing various Effect size categories for Logistic Regression and Mantel-Heanszel methods. From the findings the conclusions were drawn from the study and the recommendations made.

# CHAPTER FOUR

# RESULTS AND DISCUSSION

## 4.1 Introduction

This chapter presents the data analysis and interpretation. The chapter is divided into three sections. Section one presents the findings for the effect of the Mantel-Haenszel statistic in detecting Differential Item functioning under different conditions. Section two presents the findings for the effect of Logistic Regression statistic in detecting Differential Item Functioning under different conditions. Section three presents the findings for comparison of the effect of Mantel-Haenszel and Logistic Regression statistics in Differential Item Functioning under different conditions. The study was guided by the following objectives;

1) To determine the effect of Sample Size, Ability Distribution and Test Length on the Effect size and the number of detections of DIF items across the DIF types using Mantel Haenszel Statistic;
2) To determine the effect of Sample Size, Ability Distribution and Test Length on the Effect size and the number of detections of DIF items across the DIF types using Logistic Regression statistic;
3) To compare the effect of Sample size, Ability distribution and Test Length on the number of detections of DIF items across the DIF types using Mantel Haenszel and Logistic Regression statistics.

The data was analyzed according to these objectives.

## 4.2 Objective 1: Effect of different conditions on Effect size and the number of detections of DIF using MH statistic

In order to achieve the first objective, the Effect size values obtained after analysis of data using MH statistic for 10 items, 30 items and 50 items; a Sample size of 20, 60 and 1000; and

Ability distribution in terms of mean=0, standard deviation=1 and mean=1, standard deviation=2 were summarized. The short test length was selected according to the number of items that are often observed on personality inventories (10-15) items. Table 4.1 shows the mean Effect size $\Delta\alpha_{MH}$ values for Mantel-Haenszel statistic under different conditions.

**Table 4.1: Effect size for DIF items under different conditions using MH statistic**

| No. of items | Ability distribution (Mean, SD) | Sample size | Effect size | | |
|---|---|---|---|---|---|
| | | | Type A | Type B | Type C |
| 10 | (0, 1) | 20 | .2355 | 1.2605 | 2.9469 |
| 10 | (1, 2) | 20 | .6862 | 1.2053 | 4.8528 |
| 10 | (0, 1) | 60 | .7964 | 1.0250 | 4.4606 |
| 10 | (1, 2) | 60 | .4636 | 1.0596 | 5.5727 |
| 10 | (0, 1) | 1000 | .1644 | 1.2986 | 2.3936 |
| 10 | (1, 2) | 1000 | .5530 | 1.3772 | 3.4000 |
| 30 | (0, 1) | 20 | .4857 | 1.2485 | 3.7856 |
| 30 | (1, 2) | 20 | .8626 | 1.2322 | 4.2349 |
| 30 | (0, 1) | 60 | .7735 | 1.2953 | 2.9986 |
| 30 | (1, 2) | 60 | .6616 | 1.1273 | 4.7330 |
| 30 | (0, 1) | 1000 | .5664 | 1.2431 | 3.3960 |
| 30 | (1, 2) | 1000 | .6434 | 1.3500 | 7.3604 |
| 50 | (0, 1) | 20 | .5655 | 1.2815 | 3.2351 |
| 50 | (1, 2) | 20 | .8907 | 1.2000 | 5.1542 |
| 50 | (0, 1) | 60 | .7935 | 1.2595 | 2.7136 |
| 50 | (1, 2) | 60 | .6003 | 1.2601 | 4.0831 |
| 50 | (0, 1) | 1000 | .5544 | 1.2356 | 3.7119 |
| 50 | (1, 2) | 1000 | .4573 | 1.2934 | 4.7178 |

Key:

Type A=Negligible DIF, Type B=Moderate DIF, Type C=Large DIF

According to Table 4.1, the first column shows the number of items. The second column shows Ability distribution values in terms of mean and standard deviation. The third column shows the Sample size. The forth column shows the mean Effect size (ES) values obtained for Type A DIF items. The fifth column shows the mean Effect size (ES) values obtained for Type B DIF items. The sixth column shows the mean Effect size (ES) values obtained for Type C DIF items. It is clear from table 4.1 that, the ES for Type A DIF items had the

smallest values ( $|\Delta\alpha_{MH}| < 1$), those for Type B had moderate values ($1 \leq |\Delta\alpha_{MH}| \leq 1.5$) and those for Type C items had the largest values ( $|\Delta\alpha_{MH}| > 1.5$). These values were used to determine the effect of varied conditions on the Effect sizes of different DIF types.

In order to determine the effect of Sample Size on Effect size for each type of DIF items, One-way Analysis of Variance (ANOVA) was conducted with Effect Size as the dependent variable and Sample Size as the independent variable. Table 4.2 shows results of One-way Analysis of Variance for the effect of Sample Size on the ES of DIF across 3 DIF Types using MH statistic at a level of significance of 0.05. Statistically significant differences between means were recorded for Type B of DIF only ($F=4.234$, $df_b=2$, $df_w=15$, $p=.035$). From Table 4.2 it was evident that Sample size had a significant effect on the detection of Type B DIF items but not Type A ($F=1.605$, $df_b=2$, $df_w=15$, $p=.234$) and C ($F=.0150$, $df_b=2$, $df_w=15$, $p=.985$) items. Post-hoc analysis using Bonferroni method for pairwise comparisons revealed that for Type B DIF items, differences existed between Sample size 60 and 1000 only as displayed in Table 4.3

**Table 4.2: ANOVA Summary for effect of Sample size on Effect size of DIF across 3 DIF types using Mantel-Haenszel statistic**

| Type of DIF | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| **A** | Between Groups | .115 | 2 | .058 | 1.605 | .234 |
| | Within Groups | .539 | 15 | .036 | | |
| | **Total** | **.654** | **17** | | | |
| **B** | Between Groups | .050 | 2 | .025 | 4.234 | .035 |
| | Within Groups | .088 | 15 | .006 | | |
| | **Total** | **.137** | **17** | | | |
| **C** | Between Groups | .050 | 2 | .025 | .015 | .985 |
| | Within Groups | 24.718 | 15 | 1.648 | | |
| | **Total** | **24.767** | **17** | | | |

A strong association therefore existed between Sample size and the Effect size only for the Type B DIF items and not the Type A and Type C. This association was quite sizable in a predictive sense for any population corresponding to the current study.

**Table 4.3: Pairwise comparisons of Effect sizes across different Sample sizes for Type B DIF**

*Dependent Variable:* Effect Size *Post-hoc test:* Bonferroni

| (I) Sample size | (J) Sample size | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| 20 | 60 | .0668720 | .0441727 | .453 | -.052118 | .185862 |
| | 1000 | -.0616358 | .0441727 | .550 | -.180626 | .057354 |
| 60 | 20 | -.0668720 | .0441727 | .453 | -.185862 | .052118 |
| | 1000 | -.1285079* | .0441727 | .032 | -.247498 | -.009518 |
| 1000 | 20 | .0616358 | .0441727 | .550 | -.057354 | .180626 |
| | 60 | .1285079* | .0441727 | .032 | .009518 | .247498 |

* The mean difference is significant at the 0.05 level.

In order to determine the effect of Test Length on Effect Size for each type of DIF items, One-way Analysis of Variance was conducted with ES as the dependent variable and Test Length as the independent variable. Table 4.4 summarizes the ANOVA results for the effect of Test Length on the ES of DIF across 3 DIF Types using MH statistic. The findings indicate that Test Length had no statistically significant effect on ES of DIF items regardless of the type of DIF ($p$>.05).

For Type A DIF items, there was no significant effect of Test length on Effect size ES (F=1.668, $df_b$=2, $df_w$=15, p=.222). For Type B DIF items, there was no significant effect of Test length on Effect size ES (F=.541, $df_b$=2, $df_w$=15, p=.593), and for Type C DIF items, there was no significant effect of Test length on Effect size ES (F=.291, $df_b$=2, $df_w$=15, p=.751).

**Table 4.4: ANOVA Summary for effect of Test length on Effect size of DIF across 3 DIF types**

| Type of DIF | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| | Between Groups | .119 | 2 | .059 | 1.668 | .222 |
| **A** | Within Groups | .535 | 15 | .036 | | |
| | **Total** | **.654** | **17** | | | |
| | Between Groups | .009 | 2 | .005 | .541 | .593 |
| **B** | Within Groups | .128 | 15 | .009 | | |
| | **Total** | **.137** | **17** | | | |
| | Between Groups | .926 | 2 | .463 | .291 | .751 |
| **C** | Within Groups | 23.841 | 15 | 1.589 | | |
| | **Total** | **24.767** | **17** | | | |

In order to determine the effect of Ability Distribution on ES for each Type of DIF items, One-way Analysis of Variance was conducted with ES as the dependent variable and Ability Distribution as the independent variable. Table 4.5 summarizes the ANOVA results for the effect of Ability Distribution on the ES of DIF across 3 DIF Types using MH statistic.

**Table 4.5: ANOVA Summary for effect of Ability Distribution, on Effect size of DIF across 3 DIF types**

| Type of DIF | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| | Between Groups | .043 | 1 | .043 | 1.136 | .302 |
| **A** | Within Groups | .610 | 16 | .038 | | |
| | **Total** | **.654** | **17** | | | |
| | Between Groups | .000 | 1 | .000 | .012 | .915 |
| **B** | Within Groups | .137 | 16 | .009 | | |
| | **Total** | **.137** | **17** | | | |
| | Between Groups | 11.627 | 1 | 11.627 | 14.158 | .002 |
| **C** | Within Groups | 13.140 | 16 | .821 | | |
| | **Total** | **24.767** | **17** | | | |

Statistically significant differences for the effect of Ability Distribution on ES was noted for Type C DIF only ($F_{obs}$.=14.158, $df_b$=1, $df_w$=16, $p$=.002). Ability distribution had a significant effect on Effect size only for Type C DIF items. There was no significant effect of Ability distribution on Effect size ES for Type A DIF items ($F_{obs}$=1.136, $df_b$=1, $df_w$=16, p=.302); and also for Type B DIF items ($F_{obs}$. =012, $df_b$=1, $df_w$=16, p=.915).

Further to the above analyses, line graphs were constructed for mean ES against Test Length across DIF types and for each level of Ability Distribution and Sample Size. This outcome is presented in Figure 4.1 to aid more detailed interpretation of data. The largest mean ES was recorded for Type C DIF items. This was followed by Type B and A, respectively. This outcome was regardless of Ability Distribution, Sample Size and Test Length. However, differences in ES between Type B and C items were not as large as those between either Type A or B or Type A and C items.

More specifically, for Ability Distribution with (Mean, SD) = (0, 1) and Sample Size = 20, mean ES was largest for Type C items followed by Type B and A. However, the highest ES for Type C items occurred for 30 items. For Type C items, when Ability Distribution had (Mean, SD) = (1, 2) and Sample Size = 20, the smallest ES was recorded at Test Length = 30 items. For Ability Distribution with (Mean, SD) = (1, 2) and Sample Size = 20, the mean ES was largest for Type C items followed by B and A. For Type C DIF items, the largest ES was recorded for 10 items and the smallest for 50 items with the magnitude of ES decreasing with Test Length. For Type A and B, ES tended to slightly increase with Test Length. For Ability Distribution with (Mean, SD) = (0, 1) and Sample Size = 60, the mean ES was largest for Type C items followed by B and A. For Type C DIF items in this category, the largest ES was recorded for 10 items and the smallest for 50 items with the magnitude of ES decreasing

**ABILITY DISTRIBUTION WITH MEAN=0, SD=1**

**ABILITY DISTRIBUTION WITH MEAN=1, SD=2**

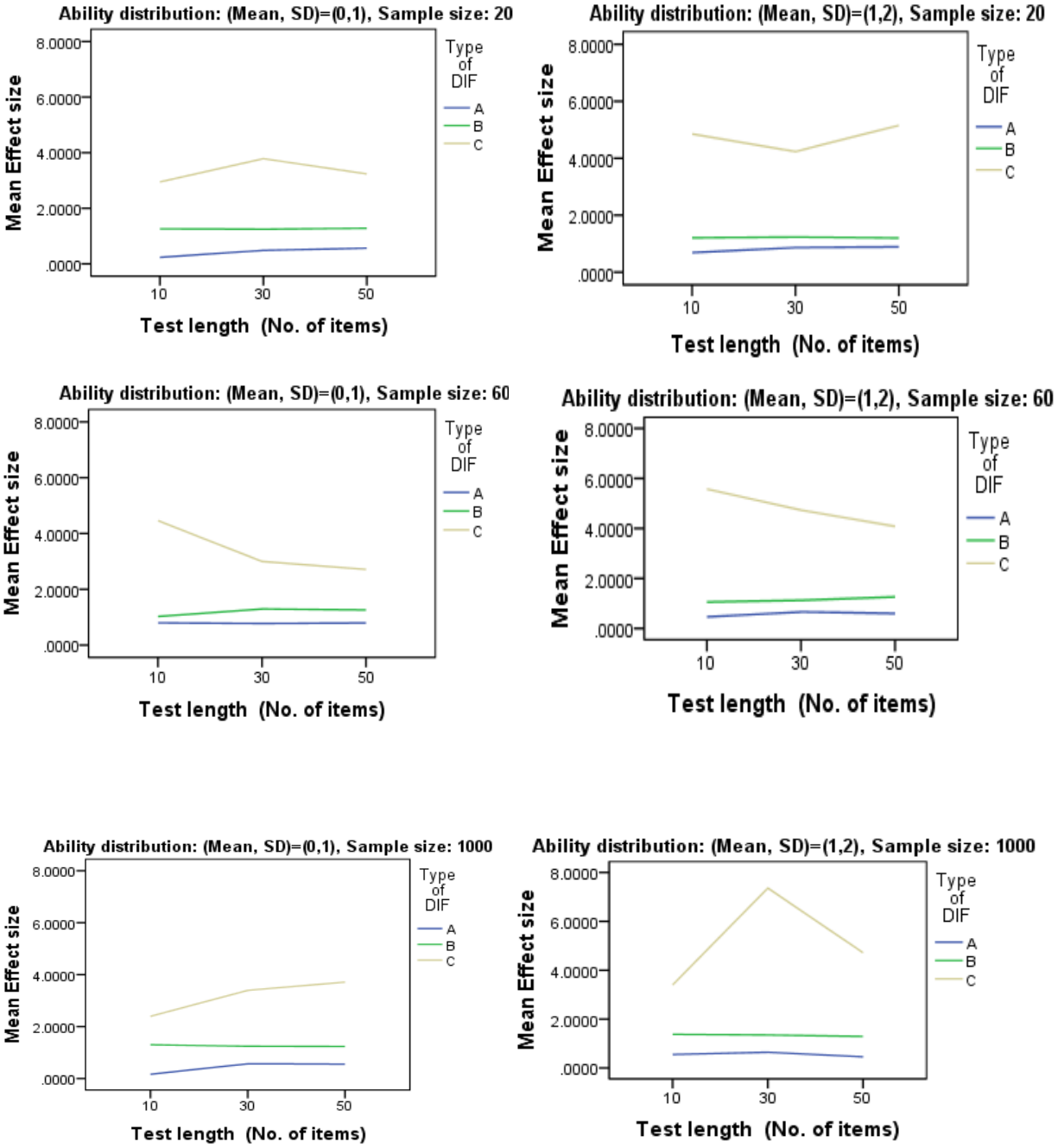



***Figure 4.1: Mean Effect sizes for different types of DIF under different conditions using MH statistic.***

with Test Length. For Type A and B, ES tended to slightly increase with Test Length. This trend was reasonably maintained when the Ability Distribution with (Mean, SD) = (1, 2) and Sample Size = 60. For Ability Distribution with (Mean, SD) = (0, 1) and Sample size = 1000, mean ES was largest for Type C items followed by Type B and A. The largest ES for Type C items in this category was recorded for 50 items and the smallest for 10 items. For Type C items, when Ability Distribution had (Mean, SD) = (1, 2) and Sample Size = 1000, the largest ES was recorded at Test Length of 30 items.

Table 4.6 shows the number of items showing different kinds of DIF for Mantel-Haenszel under different conditions. The first column shows the number of items. The second column shows the Ability distribution condition in terms of mean and standard deviation. The third column shows the Sample size. The fourth, fifth and sixth columns show the number of items showing DIF of various categories for MH statistic.

**Table 4.6: Number of DIF items detected under different conditions for MH statistic**

| No. of items | Ability distribution (Mean, SD) | Sample size | Number of DIF detections | | |
|---|---|---|---|---|---|
| | | | Type A | Type B | Type C |
| 10 | (0, 1) | 20 | 0 | 4 | 6 |
| 10 | (1, 2) | 20 | 1 | 0 | 9 |
| 10 | (0, 1) | 60 | 1 | 1 | 8 |
| 10 | (1, 2) | 60 | 0 | 1 | 9 |
| 10 | (0, 1) | 1000 | 3 | 4 | 3 |
| 10 | (1, 2) | 1000 | 3 | 2 | 6 |
| 30 | (0, 1) | 20 | 0 | 8 | 22 |
| 30 | (1, 2) | 20 | 1 | 3 | 26 |
| 30 | (0, 1) | 60 | 5 | 4 | 21 |
| 30 | (1, 2) | 60 | 5 | 4 | 21 |
| 30 | (0, 1) | 1000 | 10 | 7 | 13 |
| 30 | (1, 2) | 1000 | 2 | 2 | 26 |
| 50 | (0, 1) | 20 | 0 | 23 | 27 |
| 50 | (1, 2) | 20 | 3 | 6 | 42 |
| 50 | (0, 1) | 60 | 16 | 13 | 21 |
| 50 | (1, 2) | 60 | 5 | 5 | 40 |
| 50 | (0, 1) | 1000 | 23 | 11 | 16 |
| 50 | (1, 2) | 1000 | 11 | 6 | 33 |

Key:

Type A=Negligible DIF, Type B=Moderate DIF, Type C=Large DIF

Line graphs were constructed for mean number of DIF detections against Test Length across DIF types and for each level of Ability Distribution and Sample Size. This outcome is presented in Figure 4.2 to aid more detailed interpretation of data. Figure 4.2 shows graphs of the mean number of detections for different types of DIF under different conditions of Sample Size, Ability Distribution and Test length. In addition, the largest difference in DIF detection was recorded when Test Length was 30 items (Moderate Test Length). The same pattern was maintained when Sample Size increased to 60 except that the DIF detection between Type A and Type B DIF items at this level tended to increase as Test Length increased to 30 and then to 50 items.

When Sample Size = 1000 and Ability Distribution is (Mean, SD) = (0, 1), differences in mean DIF detection were minimal across the three types of DIF items i.e. A, B and C. However, differences in mean DIF detection tended to increase with Test Length, with the largest difference occurring when Test Length was 50 items i.e. for the longest test. A point of departure from the previous two trends is that in this case (i.e. Sample size of 1000 and Ability Distribution with (Mean, SD) = (0,1)), Type A items were detected much more than Type C items for the case of the longest test with 50 items.

At Sample Size = 20 and Ability Distribution with (Mean, SD) = (1, 2), Type C items consistently recorded the highest mean number of DIF detections across the three levels of Test length (i.e. 10, 30 and 50 items). The smallest difference in mean number of DIF detections in this case was found to exist between Type A and Type B items for the shortest test of 10 items. A similar outcome was recorded for a Sample size 60, except that the difference in mean DIF detection for Type A and Type B items was minimal. When Sample size got increased to 1000, results were similar to those for Sample size of 60 except that

**ABILITY DISTRIBUTION WITH MEAN=0, SD=1**

**ABILITY DISTRIBUTION WITH MEAN=1, SD=2**

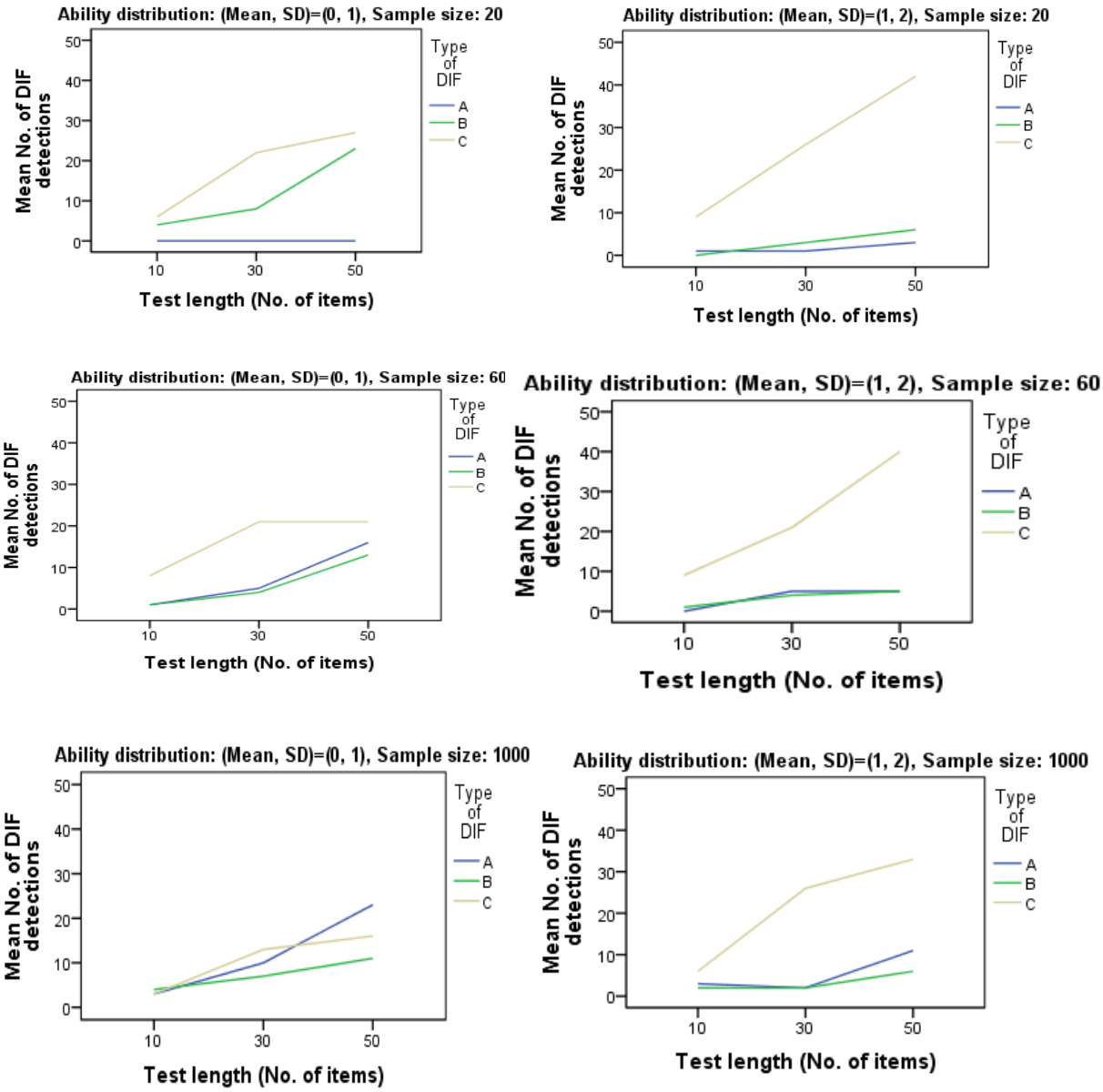



***Figure 4.2: Mean number of DIF detections for different types of DIF under different conditions using MH statistic.***

Type A and Type B items exhibited relatively larger differences in mean DIF detection at a Test length of 50 items. Thus, when the Ability distribution has (Mean, SD) = (0, 1), and number of items is large (50), MH statistic gives optimal results for Type A items than for Type B or C items.

The purpose of this study was to investigate the effect of Sample Size, Ability Distribution and Test Length on Effect Size (ES) of DIF, and the influence of the same variables on detection of DIF using Mantel-Haenszel (MH) statistic. Results indicate that Sample Size had a statistically significant effect on ES for Type B items (Moderate DIF items) and not for Types A or C. Post-hoc test indicated that significant differences in ES for Type B items existed between Sample Size=60 and Sample Size=1000 only. This suggests that it is Type B items that may be problematic when measuring DIF using MH statistic, particularly for moderate to large Sample sizes. Ability Distribution was found to have a statistically significant effect on ES for Type C items (i.e. Large DIF items) only. This suggests that for items with large DIF, the nature of Ability Distribution remains crucial when using the MH statistic.

Whereas Test Length had no statistically significant effect on ES for all the three item Types, there was a general trend for ES to increase with Test Length. The reason could have been that longer tests are more reliable than shorter tests. Line graphs indicated an effect of test length though this was not significant. This was consistent with the findings of Rogers and Swaminathan (1993) as well as Uttaro and Millsap (1994), who found that the greatest impact on ES was for Type C items (i.e. items with large DIF). This notwithstanding, the finding in the present study that MH works best for Type C items compared to either Type B or Type A items concurs with that of Zwick and Ercikan (1989). In a similar token, detection of DIF

using MH statistic tends to improve slightly with Test Length, and this becomes more prominent with Type C items. Indeed, differences in detection of DIF across item Types was more manifest in longer tests than shorter ones, with Type C items generally associated with the highest detection rates.

These findings are not consistent with previous research by Hidalgo and Lopez Pina (2004) which stated that MH analysis generally detected few items with DIF. Their research did not indicate the category of DIF detected. They also stated that Test length (40 to 80 items), resulted in improved performance of the MH procedure. They did not determine the effect of test length on the effect size and the number of detections of DIF using significance tests.

The findings of the study indicated that Ability distribution had a significant effect on the effect size only for Type C DIF items and not Type A and Type C DIF items. The findings are consistent with previous research by Swaminathan and Narayanan (1994) which stated that ability distribution had a significant effect on DIF detection by the MH procedure but had no significant effect by the SIBTEST procedure. The study did not however look at the effect of ability distribution on the DIF type. The study also determined DIF for equal and unequal ability distributions and found that MH was powerful in detecting DIF for equal ability distributions.

The current study indicated that sample size had a statistically significant effect on the effect size only for Type B DIF items and not Type A and Type C dif items. The findings are not consistent with previous research by Gierl, Gotzmann, and Boughton (2004) which stated that when DIF percentage and sample size were small, adverse effects in DIF detection rates were

not experienced. Their study did not indicate whether the effect of sample size depended on the DIF type.

This study made use of dichotomous item response data and not polytomously scored items. It is important that care is taken not to generalize findings to polytomous data as this was outside the scope of the present study. While the results reveal significant findings and draw important implications in the field of DIF, Harrison et al. (2007) argue that simulation is prone to misspecification errors. Further, Davies, Eisenhardt and Bingham (2007) also observed that generalization based on simulation studies must be treated with caution beyond the parameter range specified in the model. This notwithstanding, it is important to mention that Othuon (1998), and Davies, Eisenhardt and Bingham (2007) observed that the key strength of simulation is its ability to support investigation of phenomena that are hard to research by conventional means, particularly in situations where empirical data are limited.

## 4.3 Objective 2: Effect of different conditions on Effect size and the number of detections of DIF using LR statistic

In order to achieve the second objective, the Effect size values obtained after analysis of data using LR statistic for 10 items, 30 items and 50 items; a Sample size of 20, 60 and 1000; and Ability distribution in terms of mean=0,standard deviation=1 and mean=1,standard deviation=2 were summarized. Table 4.7 shows the mean Effect size $R^2\Delta$ values for Logistic Regression statistic under different conditions.

According to Table 4.7, the first column shows the number of items. The second column shows Ability distribution values in terms of mean and standard deviation. The third column shows the Sample size. The forth column shows the mean Effect Size (ES) values obtained for Type A DIF items. The fifth column shows the mean Effect size (ES) values obtained for

**Table 4.7: Effect size for different types of DIF items under different conditions using LR statistic**

| No. of items | Ability distribution (Mean, SD) | Sample size | Effect size | | |
|---|---|---|---|---|---|
| | | | Type A | Type B | Type C |
| 10 | (0, 1) | 20 | 0.02313 | 0.04440 | 0.28740 |
| 10 | (1, 2) | 20 | 0.02020 | 0.04343 | 0.21416 |
| 10 | (0, 1) | 60 | 0.02185 | 0.04270 | 0.17816 |
| 10 | (1, 2) | 60 | 0.01551 | 0.06333 | 0.28640 |
| 10 | (0, 1) | 1000 | 0.00783 | 0.05232 | 0.15490 |
| 10 | (1, 2) | 1000 | 0.00592 | 0.05500 | 0.17635 |
| 30 | (0, 1) | 20 | 0.02842 | 0.04803 | 0.14392 |
| 30 | (1, 2) | 20 | 0.02484 | 0.04406 | 0.19029 |
| 30 | (0, 1) | 60 | 0.01647 | 0.04652 | 0.13831 |
| 30 | (1, 2) | 60 | 0.01890 | 0.05111 | 0.21542 |
| 30 | (0, 1) | 1000 | 0.00999 | 0.04242 | 0.28019 |
| 30 | (1, 2) | 1000 | 0.01242 | 0.05753 | 0.27430 |
| 50 | (0, 1) | 20 | 0.02727 | 0.04878 | 0.22616 |
| 50 | (1, 2) | 20 | 0.02579 | 0.04738 | 0.20307 |
| 50 | (0, 1) | 60 | 0.01977 | 0.05089 | 0.18840 |
| 50 | (1, 2) | 60 | 0.01599 | 0.05474 | 0.35606 |
| 50 | (0, 1) | 1000 | 0.00793 | 0.04673 | 0.20589 |
| 50 | (1, 2) | 1000 | 0.00865 | 0.05390 | 0.25412 |

Key:

Type A=Negligible DIF, Type B=Moderate DIF, Type C=Large DIF

Type B DIF items. The sixth column shows the mean Effect size (ES) values obtained for Type C DIF items. As would be expected, the ES for Type A DIF items had the smallest values ($R^2\Delta < 0.035$), those for Type B had moderate values ($0.035 \leq R^2\ \Delta < 0.070$) and those for Type C items had the largest values ($R^2\Delta \geq 0.070$). These values were used to determine the effect of varied conditions on the Effect sizes of different DIF types.

In order to determine the effect of Sample Size on Effect size (ES) for each type of DIF items, One-way Analysis of Variance (ANOVA) was conducted with Effect Size as the dependent variable and Sample Size as the independent variable. Table 4.8 shows results of One-way Analysis of Variance for the effect of Sample Size on the ES of DIF across 3 DIF Types using LR statistic at a level of significance of 0.05.

**Table 4.8: ANOVA Summary for effect of Sample size on Effect size of DIF for LR across 3 DIF types**

| Type of DIF | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| **A** | Between Groups | .00078854 | 2 | .00039427 | 59.2256 | .000000076 |
| | Within Groups | .00009986 | 15 | .00000666 | | |
| | **Total** | **.00088840** | **17** | | | |
| **B** | Between Groups | .00083037 | 2 | .00041518 | 1.36845 | .28451104 |
| | Within Groups | .00455094 | 15 | .00030340 | | |
| | **Total** | **.00538131** | **17** | | | |
| **C** | Between Groups | .00090915 | 2 | .00045457 | .119424 | .88826636 |
| | Within Groups | .05709560 | 15 | .00380637 | | |
| | **Total** | **.05800475** | **17** | | | |

Statistically significant differences between means were recorded for the Type A DIF ($F=59.2256$, $df_b=2$, $df_w=15$, $p=.000000076$) and not for Type B DIF ($F=1.36845$, $df_b=2$, $df_w=15$, $p=.28451104$) and Type C ($F=.119424$, $df_b=2$, $df_w=15$, $p=.88826636$). From Table 4.8 it was evident that Sample size had a significant effect on the detection of Type A DIF items but not Type B and C items.

Post-hoc analysis using Bonferroni method for pairwise comparisons revealed that for Type A DIF items, differences existed between Sample size 20 and 60; and 20 and 1000; and 60 and 1000 only as displayed in Table 4.9. A strong association therefore existed between Sample size and the Effect size only for the Type A items and not the Type B and Type C DIF items. This association was quite sizable in a predictive sense for any population corresponding to the current study.

**Table 4.9: Pairwise comparisons of Effect sizes across different Sample sizes for Type A DIF**

*Dependent Variable:* Effect Size

*Post-hoc test:* Bonferroni

| (I) Sample size | (J) Sample size | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval Lower Bound | Upper Bound |
|---|---|---|---|---|---|---|
| 20 | 60 | .0068600* | .00148964 | .001 | .0028473 | .0108727 |
| | 1000 | .0161517* | .00148964 | .000 | .0121390 | .0201644 |
| 60 | 20 | -.0068600* | .00148964 | .001 | -.0108727 | -.0028473 |
| | 1000 | .0092917* | .00148964 | .000 | . 0052790 | .0133044 |
| 1000 | 20 | -.0161517* | .00148964 | .000 | -.0201644 | -.0121390 |
| | 60 | .0092917* | .00148964 | .000 | -. 0133044 | -.0052790 |

* The mean difference is significant at the 0.05 level

In order to determine the effect of Test Length on ES for each type of DIF items, One-way Analysis of Variance was conducted with ES as the dependent variable and Test Length as the independent variable. Table 4.10 summarizes the ANOVA results for the effect of Test Length on the ES of DIF across 3 DIF Types using LR statistic.

**Table 4.10: ANOVA Summary for effect of Test Length on Effect size of DIF for LR across 3 DIF Types**

| Type of DIF | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| A | Between Groups | .00002375 | 2 | .00001187 | .206004 | .81609397 |
| | Within Groups | .00086465 | 15 | .00005764 | | |
| | **Total** | **.00088840** | **17** | | | |
| B | Between Groups | .00078593 | 2 | .00039297 | 1.28270 | .30601893 |
| | Within Groups | .00459380 | 15 | .00030636 | | |
| | **Total** | **.00538131** | **17** | | | |
| C | Between Groups | .00323269 | 2 | .00161635 | .442656 | .65045334 |
| | Within Groups | .05477205 | 15 | .00365147 | | |
| | **Total** | **.05800475** | **17** | | | |

The findings indicate that Test Length had no statistically significant effect on ES of DIF items regardless of the type of DIF ($p$>.05). For Type A DIF items, there was no significant effect of Test length on Effect size ES (F=.206004, $df_b$=2, $df_w$=15, p=.81609397). For Type B DIF items, there was no significant effect of Test length on Effect size ES (F=1.28270, $df_b$=2, $df_w$=15, p=.30601893) and for Type C DIF items, there was no significant effect of Test length on Effect size ES (F=.442656, $df_b$=2, $df_w$=15, p=.65045334).

In order to determine the effect of Ability Distribution on ES for each type of DIF items, One-way analysis of variance was conducted with ES as the dependent variable and Ability Distribution as the independent variable.

Table 4.11 summarizes the ANOVA results for the effect of Ability Distribution on the ES of DIF across 3 DIF Types using LR statistic. Ability Distribution had no statistically significant effect on ES for Type C DIF (p>0.05) ($F_{obs}$.= 2.36736, $df_b$=1, $df_w$=16, $p$=.14343733). For Type A DIF items, there was no significant effect of Ability Distribution on Effect size ES (F=.211385, $df_b$=1, $df_w$=16, p=.65186968) and for Type B DIF items, there was no significant effect of Ability Distribution on Effect size ES (F=.045537, $df_b$=1, $df_w$=16, p=.83371535).

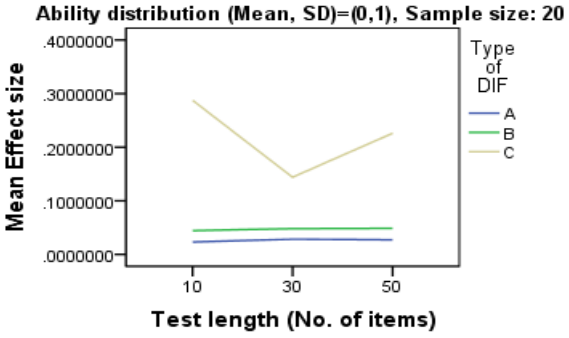
**Table 4.11: ANOVA Summary for effect of Ability Distribution on effect size of DIF for LR across 3 DIF types**

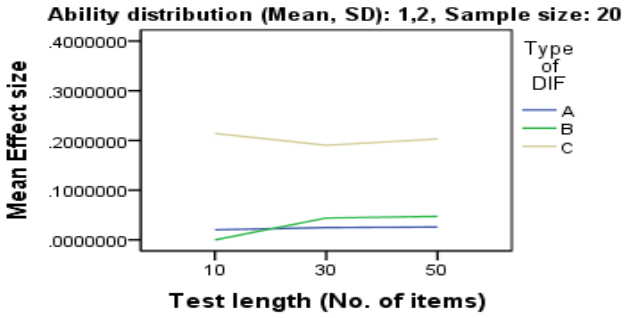
| Type of DIF | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| | Between Groups | .00001158 | 1 | .00001158 | .211385 | .65186968 |
| **A** | Within Groups | .00087681 | 16 | .00005480 | | |
| | **Total** | **.00088840** | **17** | | | |
| | Between Groups | .000015272 | 1 | .000015272 | .045537 | .83371535 |
| **B** | Within Groups | .005366038 | 16 | .000335377 | | |
| | **Total** | **.005381310** | **17** | | | |
| | Between Groups | .007476120 | 1 | .007476120 | 2.36736 | .14343733 |
| **C** | Within Groups | .050528549 | 16 | .003158034 | | |
| | **Total** | **.058004748** | **17** | | | |

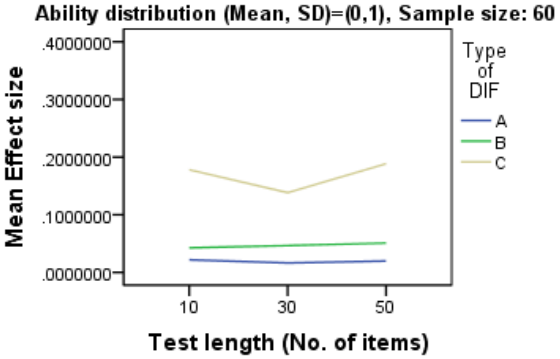
Further to the above analyses, line graphs were constructed for mean ES against Test Length across DIF types and for each level of Ability Distribution and Sample Size. This outcome is presented in Figure 4.3 to aid more detailed interpretation of data. The largest mean ES was recorded for Type C DIF items. This was followed by Type B and A, respectively. This outcome was regardless of Ability Distribution, Sample Size and Test Length. However, differences in ES between Type A and B items were not as large as those between both Type A and C or Type B and C items. More specifically, for Ability Distribution with (Mean, SD) = (0, 1) and Sample size=20, mean ES was largest for Type C items followed by Type B and Type A items. However, the highest ES for Type C items occurred for 10 items. For Type C items, when Ability distribution had (Mean, SD) = (1, 2) and sample size=20, the smallest ES was recorded at Test length =30 items. For ability distribution with (Mean,SD) = (1,2) and Sample size=20, the mean effect size was largest for Type C items followed by Type B and Type A items.

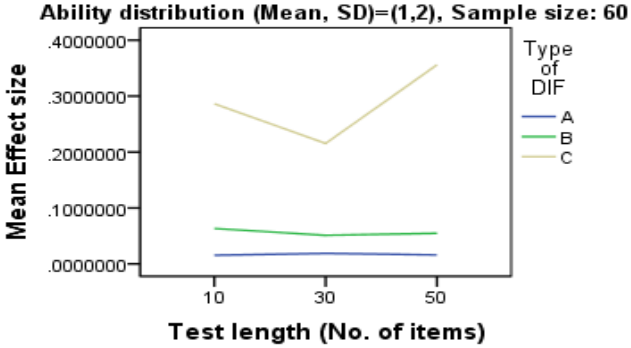
For Type C DIF items, the largest ES was recorded for 10 items and the smallest for 30 items with the magnitude of ES decreasing with Test Length. For Type B items, ES tended to slightly increase with Test Length while for Type A it remained constant with an increase in Test length. For Ability Distribution with (Mean, SD) = (0, 1) and Sample Size=60, the mean ES was largest for Type C items followed by Type B and A items. For Type C DIF items in this category, the largest ES was recorded for 10 items and the smallest for 50 items with the magnitude of ES decreasing with Test Length. For Type A and B, ES tended to marginally increase with Test Length. This trend was reasonably maintained when the Ability Distribution with (Mean, SD) = (1, 2) and Sample Size=60. For Ability Distribution with (Mean, SD) = (0, 1) and Sample size=1000, mean ES was largest for Type C items followed by Type B and A. The largest ES for Type C items in this category was recorded for 50 items

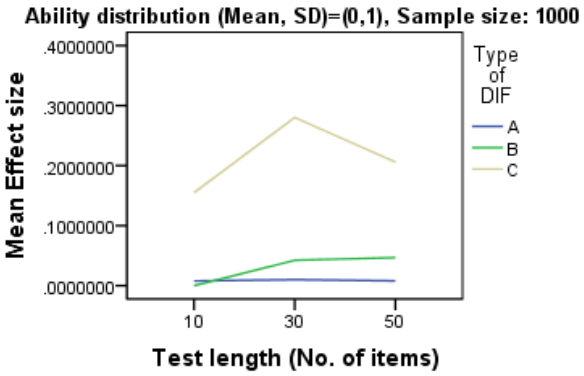
**ABILITY DISTRIBUTION WITH MEAN=0,SD=1** **ABILITY DISTRIBUTION WITH MEAN=1,SD=2**



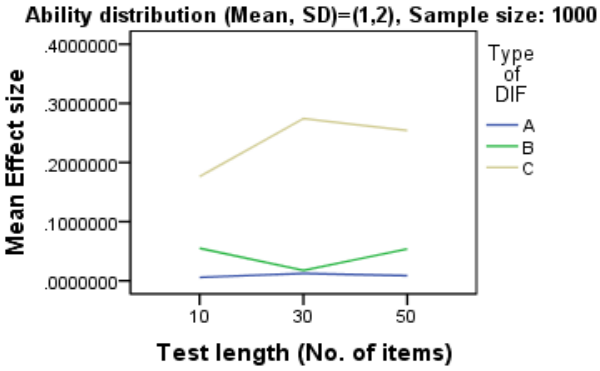




***Figure 4.3: Mean Effect sizes for different types of DIF under different conditions using LR statistc***

and the smallest for 10 items. For Type C items, when Ability Distribution had (Mean, SD) = (1, 2) and Sample Size=1000, the largest ES was recorded at Test Length of 30.

For Ability Distribution with (Mean, SD) = (1, 2) and Sample Size=20, the mean ES was largest for Type C items followed by Type B and A. For Type C DIF items, the largest ES was recorded for 10 items and the smallest for 30 items with the magnitude of ES decreasing with Test Length. For Type B items, the ES tended to marginally increase with Test Length while for Type A it remained constant with an increase in Test length. For Ability Distribution with (Mean, SD) = (0, 1) and Sample Size=60, the mean ES was largest for Type C items followed by Type B and A. For Type C DIF items in this category, the largest ES was recorded for 50 items and the smallest for 30 items. For Type A and B, ES tended to remain constant with Test Length. This trend was reasonably maintained when the Ability Distribution with (Mean, SD) = (1, 2) and Sample Size=60 though the mean Effect size for Type C was larger than that for (Mean, SD) = (0, 1) and Sample Size=60.

For Ability Distribution with (Mean, SD) = (0, 1) and Sample size=1000, mean ES was largest for Type C items followed by Type B and A items. The largest ES for Type C items in this category was recorded for 30 items and the smallest for 10 items. For Type C items, when Ability Distribution had (Mean, SD) = (1, 2) and Sample Size=1000, the largest ES was recorded at Test Length of 30 items. The mean Effect size of Type C items was much almost the same as when the Ability Distribution had (Mean, SD) = (0, 1) and Sample Size=1000. For Type A and B items the ES was very low but also tended to be the same across the various Test lengths. Ability distribution therefore tended to have an effect on the ES regardless of the Sample size and Test length. Table 4.12 shows the number of items showing different kinds of DIF for Logistic Regression under different conditions.

**Table 4.12: Number of DIF items detected under different conditions for LR statistic**

| No. of items | Ability distribution (Mean, SD) | Sample size | Number of DIF detections | | |
|---|---|---|---|---|---|
| | | | Type A | Type B | Type C |
| 10 | (0, 1) | 20 | 3 | 4 | 3 |
| 10 | (1, 2) | 20 | 2 | 0 | 8 |
| 10 | (0, 1) | 60 | 2 | 3 | 5 |
| 10 | (1, 2) | 60 | 7 | 1 | 2 |
| 10 | (0, 1) | 1000 | 9 | 0 | 1 |
| 10 | (1, 2) | 1000 | 5 | 3 | 2 |
| 30 | (0, 1) | 20 | 11 | 8 | 11 |
| 30 | (1, 2) | 20 | 9 | 5 | 16 |
| 30 | (0, 1) | 60 | 13 | 5 | 12 |
| 30 | (1, 2) | 60 | 9 | 9 | 12 |
| 30 | (0, 1) | 1000 | 18 | 4 | 8 |
| 30 | (1, 2) | 1000 | 7 | 1 | 22 |
| 50 | (0, 1) | 20 | 18 | 17 | 15 |
| 50 | (1, 2) | 20 | 12 | 6 | 32 |
| 50 | (0, 1) | 60 | 34 | 8 | 8 |
| 50 | (1, 2) | 60 | 13 | 8 | 29 |
| 50 | (0, 1) | 1000 | 32 | 4 | 14 |
| 50 | (1, 2) | 1000 | 21 | 5 | 24 |

Key:

Type A=Negligible DIF, Type B=Moderate DIF, Type C=Large DIF

The first column shows the number of items. The second column shows the Ability distribution condition in terms of mean and standard deviation. The third column shows the Sample size. The fourth, fifth and sixth columns show the number of items showing DIF of various categories for LR statistic. The number of DIF items detected under different conditions is shown in Table 4.12 for three types of DIF items; A, B and C.

The information in Table 4.12 is summarized using line graphs in Figure 4.4. The graphs show the mean number of detections for different types of DIF under different conditions of Sample Size, Ability Distribution and Test length. In general, the mean number of DIF detections using LR statistic increased with Test Length regardless of the nature of Ability Distribution, Sample size and Type of DIF. When the Ability distribution was such that (Mean, SD)=(0,1), and the Sample size was at its lowest level of 20, only small differences in DIF detection occurred between Type A and Type C items. However there were reasonable

differences in DIF detection between the two item Types and Type B items with the highest mean DIF detection being recorded for Type B items. In addition, the largest difference in DIF detection was recorded when Test Length was 50 items (Large Test Length). The same pattern was maintained when Sample Size increased to 60 except that the DIF detection between Type A and Type B DIF items at this level tended to increase as Test Length increased to 30 and then to 50 items.

When Sample Size=1000 and Ability Distribution was (Mean, SD) = (0, 1), differences in mean DIF detection were large across the three types of DIF items i.e. A, B and C. However, differences in mean number of DIF detection tended to increase with Test Length, with the largest difference occurring when Test Length was 50 items i.e. for the longest test. A point of departure from the previous two trends is that in this case (i.e. Sample size of 1000 and Ability Distribution with (Mean, SD) = (0,1), Type A items were detected much more than Type C items for the case of the longest test with 50 items.

At Sample Size=20 and Ability Distribution with (Mean, SD) = (1, 2), Type C items consistently recorded the highest mean number of DIF detections across the three levels of Test length (i.e. 10, 30 and 50 items). The smallest difference in mean number of DIF detections in this case was found to exist between Type A and Type B items for the shortest test of 10 items. For a sample of size 60, the difference in mean DIF detection for Type B and Type C items was minimal for a Test length of 10 items but it was very large for a Test length of 50 items. The same number of DIF items was recorded for Type A and B for a Test length of 30 items. When Sample size got increased to 1000, results were similar to those for Sample size of 60 except that Type A and Type B items exhibited relatively larger differences in mean DIF detection at a Test length of 50 items. Thus, when the Ability

**ABILITY DISTRIBUTION WITH MEAN=0,SD=1** **ABILITY DISTRIBUTION WITH MEAN=1,SD=2**

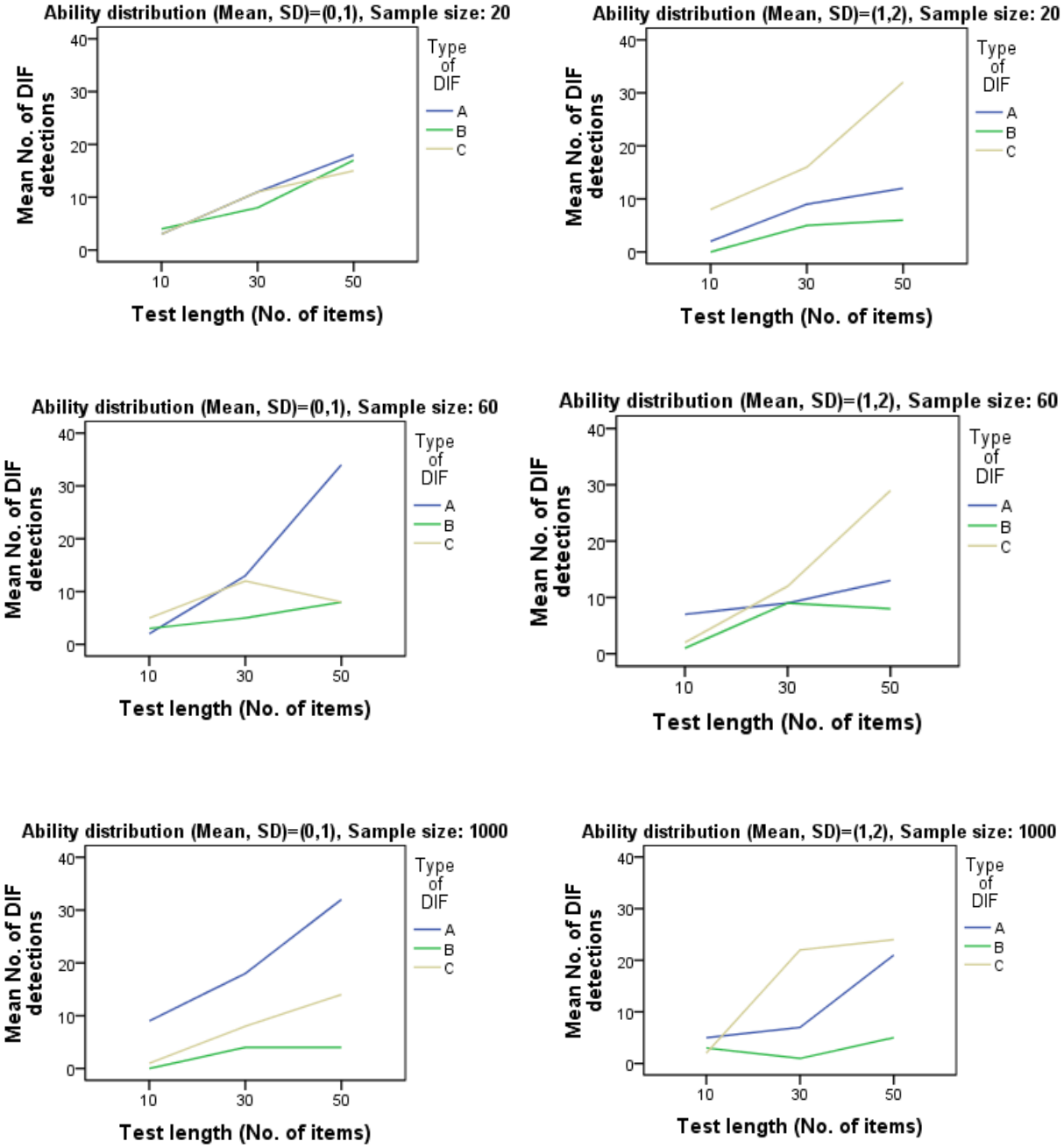



***Figure 4.4: Mean number of DIF detections for different types of DIF under different conditions using LR statistic***

distribution has (Mean, SD) = (0, 1), and number of items is large (30), LR statistic gives optimal results for Type A items than for Type B or C items.

This objective was to investigate the effect of Sample Size, Ability Distribution and Test Length on Effect Size (ES) of DIF, and the influence of the same variables on detection of DIF using Logistic Regression (LR) statistic. Results indicate that Sample Size had a statistically significant effect on ES for Type A items (Negligible DIF items) and not for Type B or C. Post-hoc test indicated that significant differences in ES for Type A items existed between Sample size 20 and 60; and 20 and 1000; and 60 and 1000 only. This suggests that it is Type A items that may be problematic when measuring DIF using LR statistic, particularly for small to large sample sizes. Ability Distribution was found to have no statistically significant effect on ES for all the DIF Types. This suggests that for all the DIF Types, the nature of Ability Distribution was not crucial when using the LR statistic.

Sample size had a statistically significant effect on the effect size only for Type A DIF items. This indicates that the higher the number of responses the higher the chances of varied responses. The findings are consistent with those of a study by Hernandez and Gomez-Bento (2006) who found out that in all the sample sizes studied DIF items were detected. The study did not however determine the effect of sample size on the procedure of DIF detection. The study only used the SIBTEST DIF detection method which was robust to large sample sizes. The current study used the LR method with both large and small sample sizes.

In a similar token, detection of DIF using LR statistic tends to improve slightly with Test Length, and this becomes more prominent with Type C items when Ability distribution is Mean=1 SD=2. Indeed, differences in detection of DIF across item Types was more manifest

in longer tests than shorter ones, with Type C items generally associated with the highest detection rates. These findings are consistent with previous research by Khalid (2011) who examined the power of MH procedure by varying the magnitude of DIF, Test length (40-80 items) and Sample size. It was found that the influence of Test length was rather modest. The findings have shown that the number of items does not greatly affect the detection of DIF of any kind by LR method.

Ability distribution had no significant effect on the effect size of all the DIF types. However statistical graphs showed that the effect varied slightly depending on the ability distribution. This was consistent with a study by Jodoin and Gierl (2002) who found no differences in DIF detection on all the ability levels studied. The study however set the unequal with a difference of .5 for the means of the reference and focal group and with the same standard deviation. The current study set the mean and standard deviation as equal for the reference and focal groups.

Test Length had no statistically significant effect on ES for the entire DIF item Types. There was a general trend for ES to increase with Test Length. This was inconsistent with the findings of Rogers and Swaminathan (1993) as well as Uttaro and Millsap (1994), who found that the greatest impact on ES was for Type C items (i.e. items with large DIF). This notwithstanding, the finding in the present study that LR works best for Type C items compared to either Type B or Type A items does not concur with that of Zwick and Ercikan (1989).

This study also made use of dichotomous item response data and not polytomously scored items. It is important that care is taken not to generalize findings to polytomous data as this

was outside the scope of the present study. While the results reveal significant findings and draw important implications in the field of DIF, Harrison et al. (2007) argue that simulation is prone to misspecification errors. Further, Davies, Eisenhardt and Bingham (2007) also observed that generalization based on simulation studies must be treated with caution beyond the parameter range specified in the model. This notwithstanding, it is important to mention that Othuon (1998), and Davies, Eisenhardt and Bingham (2007) observed that the key strength of simulation is its ability to support investigation of phenomena that are hard to research by conventional means, particularly in situations where empirical data are limited.

## 4.4 Objective 3: Effect of different conditions on the number of detections across the DIF types using MH and LR Statistics

In order to achieve the third objective, the number of items showing different kinds of DIF for MH and LR under different conditions was compared. This information was first summarized in Table 4.13.Table 4.13 shows the Number of Type A DIF items detected under different conditions for MH and LR statistics. The first column shows the number of items. The second column shows the Ability distribution condition in terms of mean and standard deviation. The third column shows the Sample size. The fourth and fifth columns show the number of items showing Type A DIF for MH and LR statistics respectively.

Line graphs showing the mean number of detections for Type A DIF under different conditions of Sample size, Ability distribution and Test length were compared for MH and LR statistics. Figure 4.5 shows the mean number of Type A DIF detections under different conditions using MH and LR statistics. From the graphs it can generally be seen that LR statistic detected more Type A DIF items than MH statistic regardless of the Sample size, Ability distribution and Test Length.

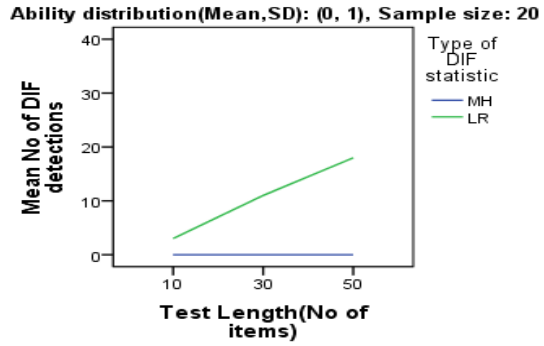
**Table 4.13: Number of Type A DIF items detected under different conditions for MH and LR statistics**

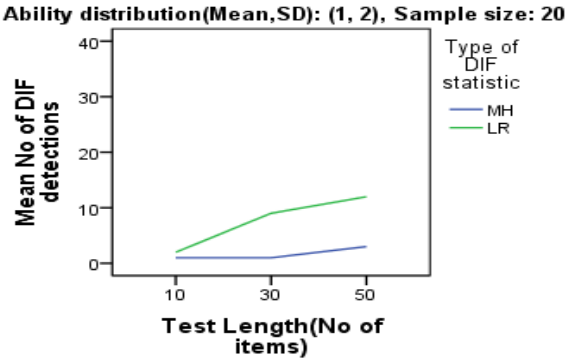
| No. of items | Ability distribution (Mean, SD) | Sample size | Number of DIF detections | |
|---|---|---|---|---|
| | | | MH | LR |
| 10 | (0, 1) | 20 | 0 | 3 |
| 10 | (1, 2) | 20 | 1 | 2 |
| 10 | (0, 1) | 60 | 1 | 2 |
| 10 | (1, 2) | 60 | 0 | 7 |
| 10 | (0, 1) | 1000 | 3 | 9 |
| 10 | (1, 2) | 1000 | 3 | 5 |
| 30 | (0, 1) | 20 | 0 | 11 |
| 30 | (1, 2) | 20 | 1 | 9 |
| 30 | (0, 1) | 60 | 5 | 13 |
| 30 | (1, 2) | 60 | 5 | 9 |
| 30 | (0, 1) | 1000 | 10 | 18 |
| 30 | (1, 2) | 1000 | 2 | 7 |
| 50 | (0, 1) | 20 | 0 | 18 |
| 50 | (1, 2) | 20 | 3 | 12 |
| 50 | (0, 1) | 60 | 16 | 34 |
| 50 | (1, 2) | 60 | 5 | 13 |
| 50 | (0, 1) | 1000 | 23 | 32 |
| 50 | (1, 2) | 1000 | 11 | 21 |

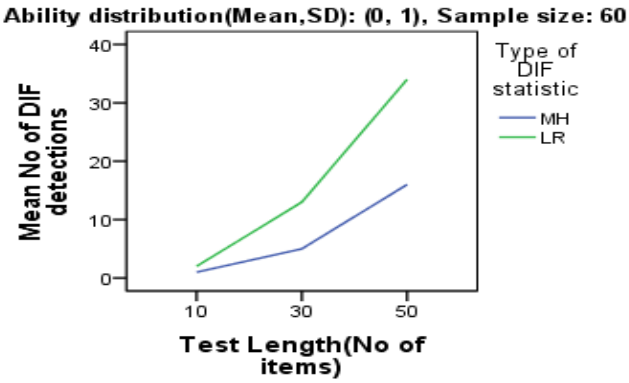
This indicates that LR is a better statistic for detecting Type A DIF than MH. When the Ability Distribution was such that (Mean, SD) = (0, 1), and the Sample Size was 20, only small differences in DIF detection occurred for Type A items between MH and LR statistics for 10 items while large differences occurred for 50 items. The number of items detected remained the same regardless of the Test length for MH statistic while it increased with Test length for the LR statistic. When the Ability Distribution was such that (Mean, SD) = (0, 1), and the Sample Size was 60, the number of DIF detections increased with Test length regardless of the DIF statistic.

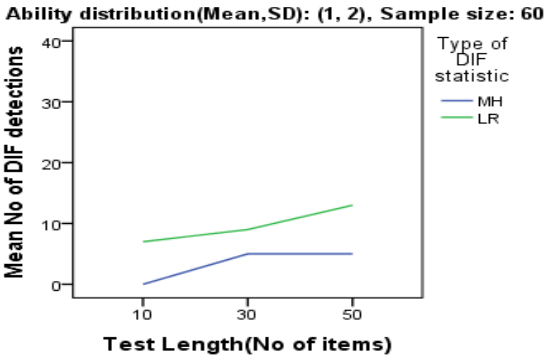
However LR detected more Type A items than MH with the highest number detected for a Test length of 50 items. From the graphs it can be seen that LR statistic detected more Type A DIF items than MH statistic regardless of the Sample size, Ability distribution and Test Length. The number of items detected by the MH statistic by was almost the same regardless

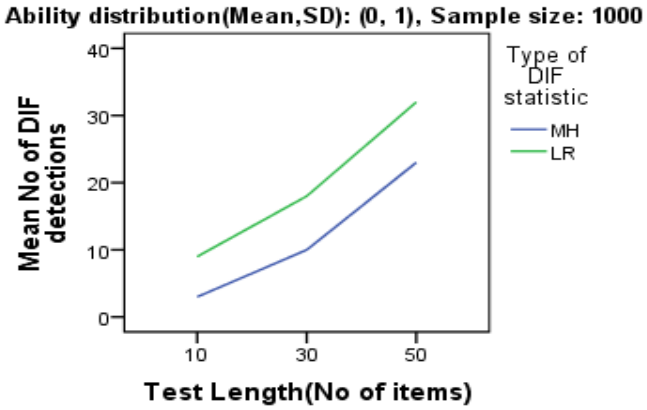
**ABILITY DISTRIBUTION WITH MEAN=0,SD=1** **ABILITY DISTRIBUTION WITH MEAN=1,SD=2**



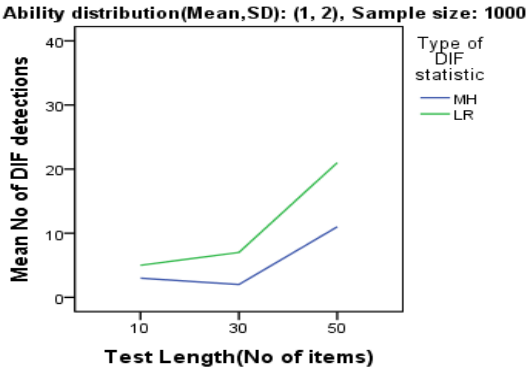




***Figure 4.5: Mean number of DIF detections for Type A DIF under different conditions using MH and LR statistics***

of the Test length while it increased with Test length for the LR statistic. When the Ability Distribution was such that (Mean, SD) = (0, 1), and the Sample Size was 60, the number of DIF detections increased with Test length regardless of the DIF statistic. However LR detected more Type A items than MH with the highest number detected for a Test length of 50 items. When the Ability Distribution was such that (Mean, SD) = (0, 1), and the Sample Size was 1000, the number of DIF detections increased with Test length regardless of the DIF statistic. However LR detected more Type A items than MH with the highest number detected for a test length of 50 items. This number was however lower than when the Sample size was 60.

When the Ability Distribution was such that (Mean, SD) = (1, 2), and the Sample Size was 20, the number of DIF detections increased with Test length regardless of the LR DIF statistic but remained at a low level for MH statistic. The number of DIF detections were however lower than when the Ability Distribution was such that (Mean, SD) = (0, 1). This indicated that Ability distribution had an effect on the number of DIF detections for LR but not for MH statistic. However LR detected more Type A items than MH with the highest number detected for a test length of 50 items.

When the Ability Distribution was such that (Mean, SD) = (1, 2), and the Sample Size was 60, the number of DIF detections was lower than when Ability Distribution was such that (Mean, SD) = (0, 1), for 30 and 50 items for both MH and LR statistics. A large difference was noted in the number of DIF detections between MH and LR statistics for 30 and 50 items. This further indicated that Ability distribution and Test length had an effect on the number of DIF detections by both MH and LR statistics. However LR detected more Type A items than MH with the highest number detected for a Test length of 50 items. When the Ability Distribution was such that (Mean, SD) = (1, 2), and the Sample Size was 1000, the

number of DIF detections decreased with Test length for 30 items using the MH statistic. As earlier noted the number of DIF detections were less than when the Ability distribution was such that (Mean, SD) = (0, 1) for both statistics.

It was therefore noted that LR detected more Type A DIF items than MH statistic regardless of the Ability distribution, Test length and Sample size. However more detections were noted when the Ability distribution was such that (Mean, SD) = (0, 1) than when the Ability distribution was such that (Mean, SD) = (1, 2) for both MH and LR statistics. Also the number of DIF detections increased with Test length regardless of the Sample size and Ability distribution

This objective was to compare the effect of Sample size, Ability Distribution and Test Length on the number of Type A DIF detections using MH and LR Statistics. The findings indicate that the Sample size had a small effect on the detection of Type A DIF items, regardless of the Ability distribution using either MH or LR statistics. However Ability distribution did have an effect in the detection of Type A items, by both MH and LR statistics.

These findings are consistent with a previous research by Hidalgo and Lopez Pina (2004) which stated that Logistic regression analysis generally detected more items with DIF than the standard MH procedure. Their research did not indicate the category of DIF detected. They also stated that Test length (40 to 80 items), resulted in improved performance of the MH procedure which is consistent with the findings of this study. These findings are also consistent with previous research by Khalid (2011) who examined the power of MH procedure by varying the magnitude of DIF, Test length and Sample size. It was found that

the influence of Test length was rather low. The findings showed that the number of items do not greatly affect the detection of DIF of any kind by MH method.

The results were not consistent with those of a study by Gonzalez-Romá, Hernandez and Gomez-Benito (2006) who indicated that power of DIF statistics increased as Sample sizes and DIF magnitude increased and that the control for Type I error was better when sample sizes were large. This study used many large Sample size conditions which differed between the reference and the focal groups. The study was also not consistent with a study by Swaminathan and Rodgers (1993) which stated that the MH procedure is as powerful as the LR procedure in detecting uniform DIF. This indicates that the number of DIF items detected by MH was almost the same as those of the LR procedure. The study however used conditions such as the model-fit and sample size. It also determined the power of MH and LR procedures in detecting uniform and non uniform DIF. The MH procedure can only detect uniform DIF and therefore it was not suitable for detecting non-uniform DIF. The current study only detected uniform DIF and compared LR and MH procedures under different conditions of sample size, ability distribution and test length.

Table 4.14 shows the Number of Type B DIF items detected under different conditions for MH and LR statistics. The first column shows the number of items. The second column shows the Ability distribution condition in terms of mean and standard deviation. The third column shows the Sample size. The fourth and fifth columns show the number of items showing Type B DIF for MH and LR statistics respectively.

**Table 4.14: Number of Type B DIF items detected under different conditions for MH and LR statistics**

| No. of items | Ability distribution (Mean, SD) | Sample size | Number of DIF detections | |
|---|---|---|---|---|
| | | | MH | LR |
| 10 | (0, 1) | 20 | 4 | 4 |
| 10 | (1, 2) | 20 | 0 | 0 |
| 10 | (0, 1) | 60 | 1 | 3 |
| 10 | (1, 2) | 60 | 1 | 0 |
| 10 | (0, 1) | 1000 | 4 | 3 |
| 10 | (1, 2) | 1000 | 2 | 8 |
| 30 | (0, 1) | 20 | 8 | 5 |
| 30 | (1, 2) | 20 | 3 | 5 |
| 30 | (0, 1) | 60 | 4 | 9 |
| 30 | (1, 2) | 60 | 4 | 4 |
| 30 | (0, 1) | 1000 | 7 | 1 |
| 30 | (1, 2) | 1000 | 2 | 17 |
| 50 | (0, 1) | 20 | 23 | 6 |
| 50 | (1, 2) | 20 | 6 | 8 |
| 50 | (0, 1) | 60 | 13 | 8 |
| 50 | (1, 2) | 60 | 5 | 8 |
| 50 | (0, 1) | 1000 | 11 | 4 |
| 50 | (1, 2) | 1000 | 6 | 5 |

Line graphs showing the mean number of detections for Type B DIF under different conditions of Sample size, Ability distribution and Test length were compared for MH and LR statistics. Figure 4.6 shows the mean number of Type B DIF detections under different conditions using and LR statistics. From the graphs it can be seen that LR statistic detected more Type B DIF items than MH statistic regardless of the Sample size, Ability distribution and Test Length. When the Ability Distribution was such that (Mean, SD) = (0, 1), and the Sample Size was 20, only small differences in DIF detection occurred for Type B items between MH and LR statistics for 10 items while large differences occurred for 50 items. The number of items detected was almost the same regardless of the Test length for MH statistic while it increased with Test length for the LR statistic.

This result is similar to that of Type A DIF items for the same Ability distribution. When the Ability Distribution was such that (Mean, SD) = (0, 1), and the Sample Size was 60, the

**ABILITY DISTRIBUTION WITH MEAN=0,SD=1** **ABILITY DISTRIBUTION WITH MEAN=1,SD=2**

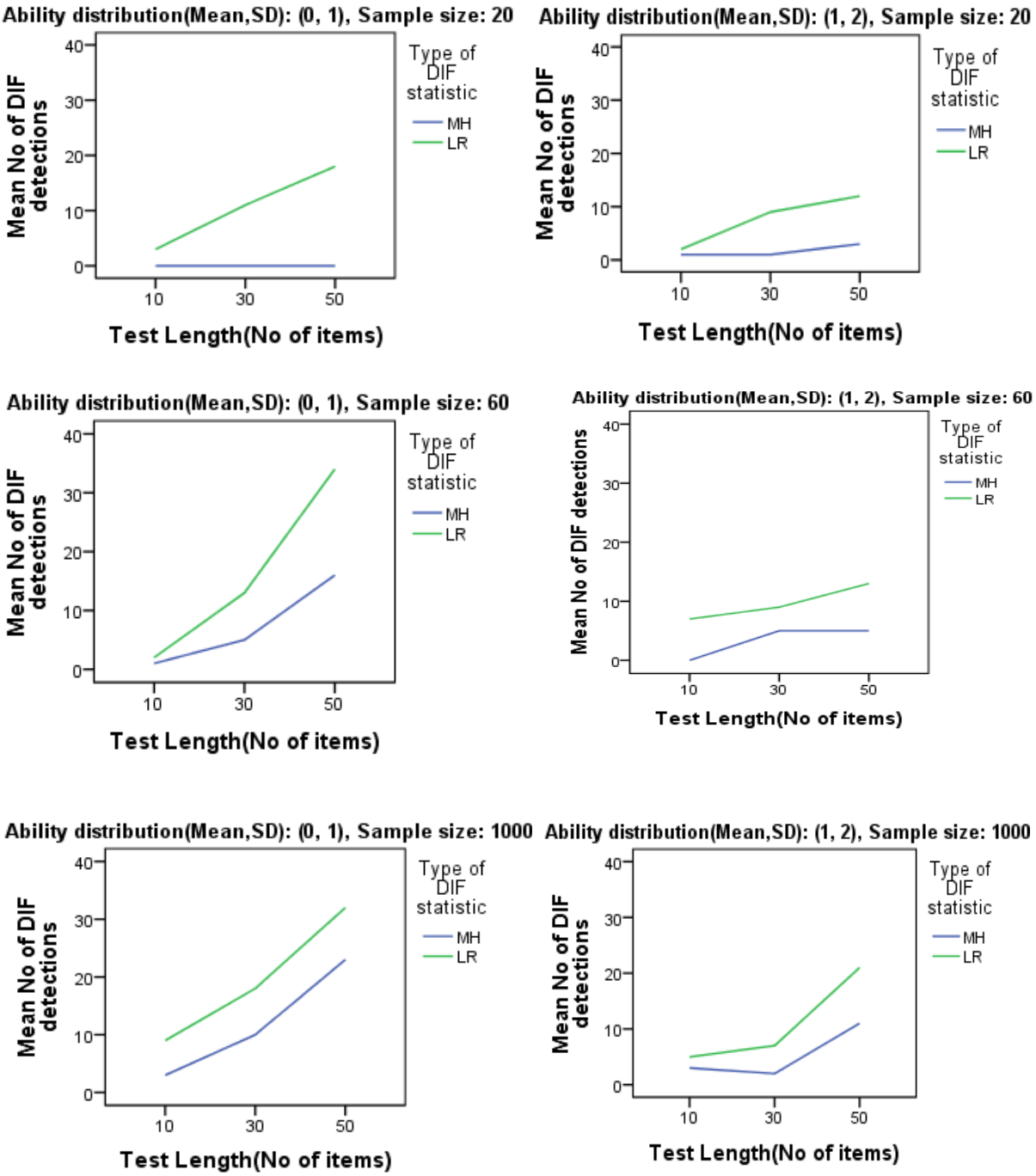



***Figure 4.6: Mean number of DIF detections for Type B DIF under different conditions using MH and LR statistics***

number of DIF detections increased with Test length regardless of the DIF statistic. However LR detected more Type B items than MH with the highest number detected for a Test length of 50 items. This result was also similar to that of Type A DIF items for the same Ability distribution. When the Ability Distribution was such that (Mean, SD) = (0, 1), and the Sample Size was 1000, the number of DIF detections increased with Test length regardless of the DIF statistic. However LR detected more Type B items than MH with the highest number detected for a Test length of 50 items. This number was however lower than that when the Sample size was 60. The result showed more Type B items detected for Test length of 10 items than Type A DIF items.

When the Ability Distribution was such that (Mean, SD) = (1, 2), and the Sample Size was 20, the number of Type B DIF detections increased with Test length for the LR DIF statistic but remained at a low level for MH statistic. The number of DIF detections were however lower than when the Ability Distribution was such that (Mean, SD) = (0, 1). This indicated that Ability distribution had an effect on the number of DIF detections for LR but not for MH statistic. However LR detected more Type B items than MH with the highest number detected for a Test length of 50 items. When the Ability Distribution was such that (Mean, SD) = (1, 2), and the Sample Size was 60, the number of DIF detections for LR increased with Test length while that for MH decreased from 30 to 50 items. This indicated that Test length had a significant effect on the detection of DIF for MH statistics.

However LR still detected more DIF items regardless of the Test length with the greatest number being detected for a Test length of 50. This further indicated that Ability distribution and Test length had an effect on the number of DIF detections by both MH and LR statistics. When the Ability Distribution was such that (Mean, SD) = (1, 2), and the Sample Size was

1000, the number of DIF detections decreased with Test length from 10 to 30 items using the MH statistic and then increased with a Test length of 50 . For LR the mean number of DIF items detected increased with the Test length. As earlier noted the number of DIF detections for Sample size 1000 were less than when the Ability distribution was such that (Mean, SD) = (0, 1) for both statistics. This further indicated that Ability distribution had an effect on the detection of Type B DIF items for both the MH and LR DIF statistics. It was therefore noted that LR detected more Type B DIF items than MH statistic regardless of the Ability distribution, Test length and Sample size.

However more detections were noted when the Ability distribution was such that (Mean, SD) = (0, 1) than when the Ability distribution was such that (Mean, SD) = (1, 2) for both MH and LR. Also the number of DIF detections increased with Test length regardless of the Sample size and Ability distribution. It was also noted that when the Ability distribution was such that (Mean, SD) = (1, 2) the number of DIF detections for MH statistic was less for Test length 30 than that of Test length 10 but increased when the Test length was 50.

This objective was to compare the effect of Sample size, Ability Distribution and Test Length on the number of Type B DIF detections using MH and LR Statistics. The findings indicate that the Sample size had a small effect on the detection of Type B DIF items, regardless of the Ability distribution using either MH or LR statistics. However Ability distribution did have an effect in the detection of Type B items, by both MH and LR statistics. Also that LR statistic detected more Type B DIF items than MH statistic notwithstanding the Ability distribution, Sample size and Test length.

These findings are consistent with previous research by Hidalgo and Lopez Pina (2004) which stated that Logistic regression analysis generally detected more items with DIF than

the standard MH procedure. Their research did not indicate the category of DIF detected. They also stated that Test length (40 to 80 items), resulted in improved performance of the MH procedure which is consistent with the findings of this study. These findings are also consistent with previous research by Khalid (2011) who examined the power of MH procedure by varying the magnitude of DIF, Test length and Sample size. It was found that the influence of Test length was rather modest. The findings showed that the number of items do not greatly affect the detection of DIF of any kind by MH method.

The results were not consistent with those of a study by Gonzalez-Romá, Hernandez and Gomez-Benito (2006) who indicated that power of DIF statistics increased as Sample sizes and DIF magnitude increased and that the control for Type I error was better when Sample sizes were large. This study used many large Sample size conditions which differed between the reference and the focal groups. The current study used small and large sample sizes to determine their effect on the DIF detection procedure. It also used statistical graphs to compare the number of DIF detections between the MH and LR methods of different DIF types.

Table 4.15 shows the Number of Type C DIF items detected under different conditions for MH and LR statistics. The first column shows the number of items. The second column shows the Ability distribution condition in terms of mean and standard deviation. This table was used to draw line graphs showing the mean number of detections for Type C DIF under different conditions of Sample size, Ability distribution and Test length for MH and LR statistics. Figure 4.7 shows the mean number of Type C DIF detections under different conditions using MH and LR statistics. From the graphs it can be seen that the MH statistic detected more Type C DIF items than LR statistic regardless of the Sample size, Ability distribution

and Test Length. When the Ability Distribution was such that (Mean, SD) = (0, 1), and the Sample Size was 20, only small differences in DIF detection occurred for Type C items between MH and LR statistics for 10 items while large differences occurred for 30 and 50 items. The number DIF items detected increased with Test length for both MH and LR statistics.

**Table 4.15: Number of Type C DIF items detected under different conditions for MH and LR statistics**

| No. of items | Ability distribution (Mean, SD) | Sample size | Number of DIF detections | |
|---|---|---|---|---|
| | | | MH | LR |
| 10 | (0, 1) | 20 | 6 | 3 |
| 10 | (1, 2) | 20 | 9 | 8 |
| 10 | (0, 1) | 60 | 8 | 5 |
| 10 | (1, 2) | 60 | 9 | 2 |
| 10 | (0, 1) | 1000 | 3 | 1 |
| 10 | (1, 2) | 1000 | 6 | 2 |
| 30 | (0, 1) | 20 | 22 | 11 |
| 30 | (1, 2) | 20 | 26 | 16 |
| 30 | (0, 1) | 60 | 21 | 12 |
| 30 | (1, 2) | 60 | 21 | 12 |
| 30 | (0, 1) | 1000 | 13 | 8 |
| 30 | (1, 2) | 1000 | 26 | 22 |
| 50 | (0, 1) | 20 | 27 | 15 |
| 50 | (1, 2) | 20 | 42 | 32 |
| 50 | (0, 1) | 60 | 21 | 8 |
| 50 | (1, 2) | 60 | 40 | 29 |
| 50 | (0, 1) | 1000 | 16 | 14 |
| 50 | (1, 2) | 1000 | 33 | 24 |

When the Ability Distribution was such that (Mean, SD) = (0, 1), and the Sample Size was 60, the number of Type C DIF detections increased with Test length between 10 to 30 items and was the same for 30 and 50 items using MH the statistic. For LR statistic, the number of Type C DIF items detected increased with Test length between 10 to 30 items and decreased with Test length between 30 and 50 items. When the Ability Distribution was such that (Mean, SD ) = (0, 1), and the Sample Size was 1000, the number of DIF detections increased with Test length regardless of the DIF statistic ,small differences occurred between MH and LR in the number of Type C DIF items detected.

**ABILITY DISTRIBUTION WITH MEAN=0, SD=1** **ABILITY DISTRIBUTION WITH MEAN=1, SD=2**

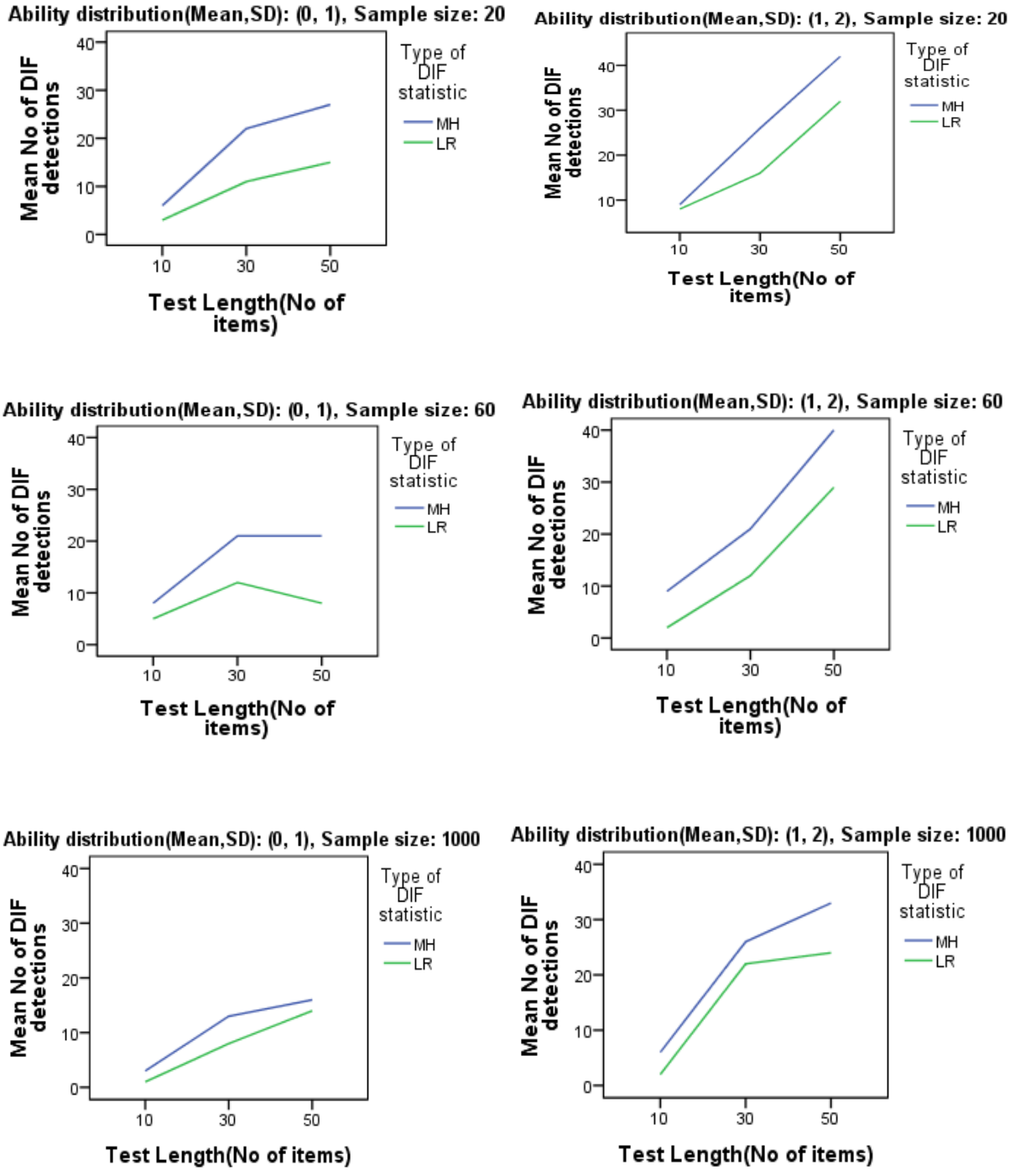



*Figure 4.7: Mean number of DIF detections for Type C DIF under different conditions using MH and LR statistics*

When the Ability Distribution was such that (Mean, SD) = (1, 2), and the Sample Size was 20, the number of DIF detections increased with Test length regardless of the Ability distribution for both the MH and LR statistics. Small differences in the Type C DIF detection between MH and LR statistics occurred for Test length 10 while large differences occurred for 30 and 50 items. When the Ability Distribution was such that (Mean, SD) = (1, 2), and the Sample Size was 60, the number of Type C DIF detections increased with Test length for both MH and LR statistics. Large differences in the DIF detection occurred between the two statistics with the largest difference occurring for 50 items. When the Ability Distribution was such that (Mean, SD) = (1, 2), and the Sample Size was 1000, the number of Type C DIF detections increased with Test length for both statistics. The difference in the detection of Type C DIF items between MH and LR statistics was small for a Test length of 10 and 30 but it was large for a Test length of 50 items. This indicated that Test length had an effect on the number of Type C DIF detections using the MH and LR statistics

It was therefore noted that MH detected more Type C DIF items than LR statistic regardless of the Ability distribution, Test length and Sample size. However more Type C DIF detections were noted when the Ability distribution was such that (Mean, SD) = (1, 2) than when the Ability distribution was such that (Mean, SD) = (0, 1) for both MH and LR statistics. Also the number of DIF detections increased with Test length regardless of the Sample size and Ability distribution in some instances while it decreased with Test length in other instances. Ability distribution, Test length and Sample size therefore had an effect on the number of Type C DIF detections by both the MH and LR statistics.

This objective was to compare the effect of Sample size, Ability Distribution and Test Length on detecting Type C DIF items using MH and LR Statistics. The findings indicate that the Sample size had a small effect on the detection of Type C DIF items of regardless of

the Ability distribution using either MH or LR statistics. However Ability distribution did have an effect in the detection of Type C items, by both MH and LR statistics. Also that MH statistic detected more Type C DIF items than LR statistic notwithstanding the Ability distribution, Sample size and Test length. These findings are also not consistent with previous research by Hidalgo and Lopez Pina (2004) which stated that Logistic regression analysis generally detected more items with DIF than the standard MH procedure. Their research did not indicate the category of DIF detected. They also stated that Test length (40 to 80 items), resulted in improved performance of the MH procedure.

The findings were also not consistent with previous research by Pedrajita and Talisayon (2009) who found a high degree of agreement between LR and MH statistics in identifying biased items (Type C DIF items). The study however used real data from Junior high school students from public and private schools. However the test length and ability distribution of the examinees was unknown. The current study compared MH and LR DIF methods under known conditions of test length, ability distribution and sample size. The findings were also not consistent with previous research by Adedoyin (2010) which used IRT method to detect gender biased items in public examinations. The study found out that out of 16 test items that fitted the 3PL IRT analysis 5 were gender biased. The study used 2000 males and 2000 females which was a large Sample size. However the effect of sample size on the DIF detection method was not investigated. The study also used only one DIF detection method. The current study used two DIF detection methods namely MH and LR to detect biased items under different conditions. The findings were also not consistent with previous research by Fidalgo, Mellenberg, and Muñiz, (2000) who reported that DIF detection by either M-H or an IRT based procedure resulted in inflated Type I error. These studies however did not find the

effect of different conditions such as sample size, ability distribution and test length on the various DIF types such as Type C.

The findings of this study show that different methods have different statistical powers for detecting DIF of different types. For instance MH is more powerful than LR in detecting Type C DIF items while LR is more powerful in detecting type A and B DIF items. This gives one direction on the DIF statistic to choose when detecting DIF of various types. These statistics are also affected by different conditions which should be considered before a DIF detection method is chosen.

# CHAPTER FIVE

# SUMMARY OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

## 5.1 Introduction

This chapter presents a summary of the study, conclusions drawn from the study and recommendations based on the findings. This is presented in four sections. Section one presents the summary of the findings. Section two presents the conclusions that were drawn from the study. Section three gives the recommendations that were made from the findings, while the last section gives recommendations for further research.

## 5.2 Summary of the Findings

The study was mainly concerned with the effectiveness of Mantel-Haenszel (MH) and Logistic Regression (LR) statistics in detecting Differential Item Functioning under different conditions. The study determined the effect of MH and LR DIF methods in detecting Differential Item Functioning under different conditions of Ability distribution, Sample size and Test length, on the Effect size and the number of DIF detections. It also compared the mean number of DIF detections through DIF analysis using LR and MH statistics under different conditions. A factorial research design was used in this study. The independent factors were Sample size, Ability distribution, and Test length. The dependent factors were the Effect sizes and number of DIF detections. The population of the study consisted of 2000 examinee responses with 1000 in the reference and 1000 in the focal group. A stratified random sampling technique was used with the stratifying criterion based on the examinee responses designated as reference and focal. The reference and focal groups had three Sample sizes: 20, 60, and 1000 each.

WinGen3 (Han, 2009) statistical software was used to generate dichotomous item response data. The responses were from 10 items, 30 items and 50 items respectively. The Ability

distribution was mean 0, SD 1 and mean1, SD 2. The data was replicated up to 1,000 times for every cell in the study, resulting into 18,000 data sets. In each replication, new item scores were generated and the Effect size determined for the studied item. The average values of the Effect sizes across the 1000 replications were determined. The Statistical Package for Social Sciences (SPSS) (IBM SPSS Version 20) was used to perform One Way Analysis of Variance (ANOVA) to determine the effect of Sample Size, Ability Distribution and Test Length on the Effect Size (ES) of DIF across three types of DIF; A, B and C for both MH and LR statistics and also to draw line graphs. The line graphs were used to compare the Effect sizes and the number of DIF items showing various Effect sizes of the LR and MH DIF statistics, in order to find out the statistic that detected DIF items of different types under different conditions.

### 5.2.1 Objective 1: Effect of different conditions on the Effect size (ES) and number of DIF detections using MH statistic

The objective of this study was to investigate the effect of Sample size, Ability distribution and Test length on the Effect size (ES) of DIF, and the influence of the same variables on detection of DIF using Mantel-Haenszel statistic. The findings indicated that Sample Size had a statistically significant effect on ES for Type B items (Moderate DIF items) and not for Types A and C. Post-hoc test indicated that significant differences in ES for Type B items existed between Sample Size=60 and Sample Size=1000 only. Ability Distribution was found to have a statistically significant effect on ES for Type C items (i.e. Large DIF items) only. Whereas Test Length had no statistically significant effect on ES for all the three item Types, there was a general trend for ES to increase with Test Length. In a similar token, detection of DIF using MH statistic tends to improve slightly with Test Length, and this becomes more prominent with Type C items. Indeed, differences in detection of DIF across

item Types was more manifest in longer tests than shorter ones, with Type C items generally associated with the highest detection rates.

### 5.2.2 Objective 2: Effect of different conditions on the Effect size (ES) and number of DIF detections using LR statistic

The objective of this study was to investigate the effect of Sample size, Ability distribution and Test length on the Effect size (ES) of DIF, and the influence of the same variables on detection of DIF using Logistic Regression statistic. The findings indicated no statistically significant effect of Ability Distribution on ES for Type A, B and C DIF on the effect size. Sample size had a significant effect on Type A DIF items but not Type B and C items. Post-hoc analysis using Bonferroni method for pairwise comparisons revealed that for Type A DIF items, differences existed between sample size 20 and 60; and 20 and 1000; and 60 and 1000. Sample size had a significant effect on the detection of Type A DIF items but not Type B and C items. The findings also indicated that Test Length had no statistically significant effect on ES of DIF items regardless of the type of DIF. For Type A, B and C DIF items, no statistically significant differences for the effect of Ability Distribution on ES were recorded for all the DIF Types.

Line graphs showed that the largest mean ES was recorded for Type C DIF items followed by Type B and A, respectively regardless of Ability Distribution, Sample Size and Test Length. Differences in ES between A and B items were not as large as those between either A and C or B and C items. The highest ES for Type C items occurred for 10 items, while the smallest ES was recorded at Test Length=30 items. For Type B, ES tended to marginally increase with Test Length while for type A it remained constant with an increase in test length. In general, the mean number of DIF detections using LR statistic increased with Test

Length regardless of the nature of Ability Distribution, Sample Size and Type of DIF. There was a general trend for ES to increase with Test Length.

### 5.2.3 Objective 3: Effect of different conditions on the number of detections across the DIF types using MH and LR Statistics

Line graphs showing the mean number of detections for different Types of DIF under different conditions of Sample size, Ability distribution and Test length were compared for MH and LR statistics. For both MH and LR statistics the mean number of DIF detections increased with Test length regardless of the nature of Ability distribution, Sample size and Type of DIF. Test length seemed to affect DIF detection using the MH and LR statistics. The largest difference in DIF detection between MH and LR statistics was recorded when the Test length was 50 items (large Test length) while the smallest difference was recorded when the Test length was 10 items.

The findings indicated that regardless of the Sample size, Ability distribution, and Test length, LR detected more Type A DIF items than MH statistic. Ability distribution had an effect on the detection of Type A and B DIF items by MH and LR statistics. The findings also indicated that regardless of the Sample size, Ability distribution, and Test length, LR detected more Type A and B DIF items than MH statistic.

The findings indicated that regardless of the Sample size, Ability distribution, and Test length, MH detected more Type C DIF items than LR statistic. Ability distribution had an effect on the detection of Type C DIF items by MH and LR statistics. It was generally noted that the detection of DIF items of any Type depended on DIF statistic. The LR statistic

detected more Type A and Type B DIF items while the MH statistic detected more Type C DIF items.

## 5.3 Conclusions

The conclusions that were drawn from the findings of this study were as follows:

### 5.3.1 Objective 1: Effect of different conditions on Effect size (ES) and number of DIF detections using MH statistic

Sample size had a statistically significant effect on the Effect size for Type B items and not Type A or Type C items and that a difference existed between Sample size 60 and 1000 only when using the MH statistic. Ability distribution had a statistically significant effect on the Effect size for Type C items and not for Type A or B items. This is a clear indication of the importance of making selective use of MH statistic in detecting DIF. Test length had no statistically significant effect on the Effect size of DIF items regardless of the type of DIF.

The detection of DIF using MH statistic generally improved with Test Length regardless of the nature of Ability distribution and Sample size. This confirmed that longer tests are normally more desirable than shorter ones and such detection when MH is used was better achieved for Type C items than either Type A or B items. The highest mean DIF detection was for Type C items and not for Type A and B items. DIF detection between Type A and Type B items at a Sample size of 60 tended to increase as Test Length increased to 30 and then to 50 items.

### 5.3.2 Objective 2: Effect of different conditions on the Effects size (ES) and number of DIF detections using LR statistic

Sample size had a statistically significant effect on the Effect size for Type A items and not Type B or Type C items and that a difference existed between Sample size 20 and 60; 20 and 1000; and 60 and 1000 when using the LR statistic. Ability distribution had no statistically significant effect on the Effect size of all types of DIF items using the LR statistic. This was a clear indication of the importance of making selective use of LR statistic in detecting DIF. Test length had no statistically significant effect on the Effect size of DIF items regardless of the type of DIF.

Detection of DIF using LR statistic also generally improved with Test Length regardless of the nature of Ability distribution and Sample size. This confirms that longer tests are normally more desirable than shorter ones and such detection when LR is used is better achieved for Type C items than either Type A or B items. Differences in Effect size between A and B items were not as large as those between either Type A and C or Type B and C items.

### 5.3.3 Objective 3: Effect of different conditions on the number of detections across the DIF types using MH and LR Statistics

Ability distribution, in terms of mean and standard deviation, contributed significantly to the number of DIF items of all kinds detected by both MH and LR statistics. Sample size did not contribute significantly to the number of DIF items of all kinds detected by both MH and LR methods. Test length did not contribute significantly to the number of DIF items of all kinds detected by both MH and LR methods. For both MH and LR statistics the mean number of

DIF detections increased with Test length regardless of the nature of Ability distribution, Sample size and Type of DIF.

Test length seemed to affect DIF detection using the MH statistic but it had no effect on DIF detection using LR statistic. Regardless of the Sample size and the Ability distribution, LR detected more Type A DIF items than MH. The detection of Type B DIF items by both statistics depended on the Ability distribution. MH detected more Type C DIF items than LR regardless of the Ability distribution and Sample size. The detection of Type C DIF items by the two statistics however depended on the Ability distribution and the Test length. This was also a clear indication of the importance of making selective use of MH or LR statistic in detecting DIF of various types.

## 5.4 Recommendations

Based on the findings of the study, the following recommendations were made;

i. The findings of the study indicated that Sample size and Ability distribution had a statistically significant effect on the Effect size for Type B and C DIF items respectively, using the MH statistic. It was therefore recommended that test developers consider using MH statistic when detecting Type B and C items under varying Sample size and Ability distribution conditions. The findings also indicated that the MH statistic detected more Type C items than Type A and B items. It was therefore recommended that test developers consider using the MH statistic when detecting biased items or items with large DIF.

ii. The findings also indicated that Sample size had a statistically significant effect on Effect size for Type A items only while Ability distribution and Test Length had no

significant effect on Effect size for all the DIF types using the LR statistic. It was therefore recommended that test developers may consider using the LR statistic when detecting Type A DIF items while considering varying Sample sizes. It was therefore recommended that Ability distribution be considered when detecting the number of DIF items using the LR statistic.

iii. The findings also indicated that the LR statistic detected more Type A and B DIF items than MH statistic while the MH statistic detected more Type C DIF items than LR, regardless of the Ability distribution and Sample size. It was therefore recommended that test developers consider the use of the LR statistic for detecting Type A and B DIF items and the MH statistic for detecting Type C DIF items. It was also recommended that Ability distribution be considered when detecting Type B DIF items regardless of the DIF statistic that is to be used.

## 5.5 Suggestions for Further Research

For the purpose of investigating further effects of different conditions on the detection of Differential item functioning, it was suggested that the following be carried out.

i. While the data used was simulated data generated using computer software, there were no real DIF items in an original test. The findings of the study cannot therefore be generalized to real data. It was therefore suggested that similar studies that use items from real data be undertaken.

ii. The study was also limited to only two DIF detection methods namely LR and MH methods and also on only three conditions namely Sample size, Ability distribution and Test length. The findings cannot therefore be generalized to other DIF detection methods such as IRT or SIBTEST and other examinee conditions. It was therefore suggested that a similar study be undertaken using other DIF detection methods and

with other examinee conditions such as Population distribution in terms of skewness and kurtosis.

iii. A limitation of the present study is that of exclusion of analyzing non-uniform DIF items. Recent literature has made it more likely and possible to determine if items contain non-uniform DIF (Hidalgo and Lopez-Pina, 2004; Jodoin & Gierl, 2001). The present study only focused on uniform DIF for a more parsimonious study due to the three conditions and the comparison between two methods (MH DIF and LR DIF). It is recommended that further research be done which could include the analysis of non-uniform DIF items and the interactions with the detection of DIF.

iv. The study was also limited to dichotomously scored items. Polytomous items did not form part of this study. It was therefore suggested that a similar study be conducted using computer software that would analyze DIF in polytomous items.

The current study investigated DIF analysis using simulated data generated using computer software. This has contributed to the little research that has been done in this area in Kenya. This study has determined the statistical power of the procedure for DIF detection under different conditions. However, studies that are focused on the effect of Sample sizes and Ability distribution are still limited. This study has added to the limited existing literature on the effectiveness of various Sample sizes, Test length and type of Ability distribution on the statistical power of DIF detection using LR and M-H procedures. The findings of this study will contribute to research and practice in schools and institutions' testing program, the formulation and implementation of educational policies and decisions related to test development. It will also help test developers and test users to make informed decisions regarding the selection of test item evaluation procedures in the area of Differential item functioning under different conditions. The findings of the study are of great significance not

only to teachers and their classroom practice but also to educational policy makers, test developers and test users.

## REFERENCES


Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91. Retrieved on 4th May 2014 at 1325 hours from http://www.apm.sagepub.com/content/34/3/166.refs.

Adedoyin, O.O. (2010). IRT approach to detect gender biased items in public examinations: A case study from the Botswana junior certificate examination in Mathematics**.** *Educational Research and Reviews* Vol. 5 (7), pp. 385-399. Retrieved on 15th July 2010 at 2105 hours from http://www.academicjournals.org/.../Pdf .

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp 3-23). Dublin: Educational Research Center.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999**).** An investigation of the power of the Likelihood ratio goodness of fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, 36(4), 277-300. Retrieved on 4th May 2010 at 0900 hours from http://www.apm.sagepub.com/content/34/3/166.refs.

Awuor, R. A. (2008). *Effect of Unequal Sample Sizes on the Power of DIF Detection: An IRT-Based Monte Carlo Study with SIBTEST and Mantel-Haenszel Procedures*. (Doctoral Dissertation), Virginia Polytechnic Institute and State University. Retrieved on19th June 2008 at 0605 hours from http://scholar.lib.vt.duu/theses/.../RAA_ETD.pdf.

Birnbaum, A. (1968). Some Latent trait models and their use in inferring an examinee's ability. In F.M. Lord and Novick (Eds). *Statistical Theories of Mental scores* (pp. 397-472).Reading, MA: Addition-Wesley. Retrieved on 3rd June 2018 at 2005 hours from https//www.researchgate.net/publication/326136908.

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement 43,* 313-334.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23*, 67-95.

Breslow, N.E. and Day, N.E. (1980). *Statistical methods in cancer research, vol. I: The analysis of case-control studies*. Scientific Publication No32. International Agency for Research on Cancer, Lyon. Retrieved on 3rd April 2015 at 0700 hours from http://iassr.org/journal (c) EJRE published by International Association of Social Science Research – IASSRISSN: 2147-6284 European Journal of Research on Education, 2015, 3(1), 7-16.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased items***.** Newbury Park, CA: Sage. Retrieved on 5th March 2013 at 2300 hours from http://education.gsu.edu/coshima/EPRS8410/dif.pdf.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*, 129-147.

Chahine, S., & Childs, R. A. (2010). Detecting Differential Item Functioning and differential Step Functioning Due to differences that should matter in Practical Assessment, Research and Evaluation. *A peer reviewed electronic journal*, 15, 10, 1-3.

Clark, P.C. (2010). *An examination of Type I Errors and Power for Two Differential Item Functioning Indices*. (Masters Thesis), Wright State University.

Clauser, B., Mazor, K. M., & Hambleton, R. K. (1998). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31, 67-78.*

Clauser, B., & Mazor, K. (1998). Using statistical procedures to identify differential functioning test items. *Educ. Measure. Issues Practice* 31-44.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133-148.

Crocker, L., & Algina, J. (1986). *Introduction to Classical Modern Test Theory*. Rinehard and Winston Inc. United States.

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.

Cromwell, S.D. (2006). *Improving the Prediction of Differential Item Functioning: A comparison of the use of an Effect size for Logistic Regression DIF and Mantel-Haenszel DIF methods*. (Doctoral Dissertation), Texas A&M University.

Davis, J.P., Eisenhardt, K. M., & Bingham,C. B. (2007). Developing theory through simulation methods. *Academy of Management Review*,32(2), 480-499.

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2*, 243-276.

De Boeck, P. (2008). Random IRT models. *Psychometrika, 73*, 533-559.

De Boeck, P., & Wilson, M. (Eds) (2004). *Explanatory Item Response Models. A generalized linear and non- linear Approach*. New York: Springer. Retrieved on 10th June 2011 at 0523 hours from http://www.jaqm.ro/.../9_Marella_Chiacchio_Bove.pdf.

DeMars, C. E. (2015). Modeling DIF for simulations: Continuous or categorical secondary trait? *Psychological Test and Assessment Modeling*, 57, 2015 (3), 279-300.

DeMars, C. E. (2009). Modification of Mantel-Haenszel and Logistic Regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics*.

DeMars, C. (2010). *Item Response Theory*. Ney York: Oxford University Press Inc.

DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education, 24*, 189-209.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*, 465-488.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2*, 217- 233.

Edelen, M. O., Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*, 5-18.

Elosua, P., & Wells, C. (2013). Detecting DIF in Polytomous Items using MACS, IRT and Ordinal Logistic Regression .*Psychometrica, 34*,324-342.

Engelhard, G. (2016). Using Item Response Theory model-Data fit to conceptualize Differential item and person functioning for students with Disabilities. *Journal of Educational Measurement, 40*, 485-548.

Erdem, K. (2014). Comparison of Mantel-Haenszel and Logistic Regression Techniques in Detecting Differential Item Functioning. *Journal of Measurement and Evaluation*. Educational. Psychology. 5(2):12-25.

Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. (2003). *SAS for Monte Carlo studies: A guide for quantitative researchers***.** Cary, NC: SAS Institute, Inc. Retrieved on 1stSeptember 2009 at 1720 hours from http://www.docstoc.com/.

Fidalgo, A. M., Ferreres, D. & Muñiz, J. (2004**).** Liberal and conservative Differential Item Functioning detection using Mantel-Hanszel and SIBTEST: Implications for Type I and Type II error rates .*Journal of Experimental Education, 73*(1), 23-39. Retrieved on 17th January, 2008 at 2100 hours from http://www.mendeley.com/.../angel-m-fidalgo/.

Fidalgo, A. M., & Laura, M. (2018). Effects of the Ability Distribution shape on the Generalised Mantel-Haenszel statistics used for DIF detection. *Methods of Psychological Research Online, 5, 439-453.* Retrieved on 17th June, 2018 at 0830 hours from http://www.researchgate.net/publication/326/36908.

Fidalgo, A. M., Mellenbergh, G. J., & Muñniz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online,* 5, 43-53. Retrieved on 17th January, 2010 at 1545 hours from http://www.mendeley.com/.../angel-m-fidalgo/.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement,* 61, 181-210. Retrieved on 12th December 2011 at 0532 hours from http*://*psycnet.apa.org/journals/ccp/73/1/136/.

Finch, W. H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST and IRT likelihood ratio. *Applied Psychological Measurement,* 29, 278-295. Retrieved on April 3rd, 2013 at 1230 hours from http://intl-epm.sagepub.com/.../0013164412472341.ful.

Finch, W. H., & French. B. F. (2007**).** Detection of crossing deferential item functioning. *Educational Psychological Measurement*, 67(4), 565-582. doi: 10.1177/0013164406296975.

Finch, W. H. (2016**).** Detection of Deferential Item Functioning for more than two groups: A Monte Carlo comparison of Methods. *Applied measurement in Education*, 29(1), 30-45 http://doi.org/10.1080/08957347.2015.1102916/.

Fleishman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika.* 43: 521-532. Retrieved on 1st May 2011 at 2025 hours from http://link.springer.com/article/10.../s13428-012-0196-.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with Logistic Regression for differential item functioning. *Educational Psychological Measurement*, 67(3), 373-393. Retrieved on 2nd March, 2012 at 1600 hours. from http://epm.sagepub.com/content/67/3/373.refs.html?patientinform-links...;67/3/

Gadermann, A., M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, & Evaluation, 17(3),* 1-13.

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests**.** *Educational Measurement: Issues and Practice*, 20, 26-36. Retrieved on 21st, November 2012 at 1100 hours from http://www.ualberta.ca/~mgierl/publications.ht.

Gierl, M. J., Gotzmann, A., & Boyghton, K. A. (2004). Performance of SIBTEST when the percent of DIF items is large. *Applied Measurement in Education*, 17(3), 241-264. Retrieved on 21st, November 2012 at 1500 hours from http://www.ualberta.ca/~mgierl/publications.ht...

Gold, M. S., Bentler, P. M., & Kim, K. H. (2003). A comparison of maximum likelihood and asymptotically distribution-free methods of treating incomplete non-normal data. *Structural Equation Modeling*, 10 (1), 47-79. Retrieved on 12th January 2012 at 2400 hours from http://www.education.pitt.edu/.../profiledetails.aspx?...K...

González-Romá, V., Hernández, A., & Gómez-Benito, J. (2006). Power and Type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items**.** *Multivariate Behavioral Research*, 41(1), 29- 53.

Güler, N., & Penfield, R. (2009). A Comparison of Logistic Regression and Contingency Table Methods for Simultaneous Detection of Uniform and Non-uniform DIF. *Journal of. Educational. Measurement.* 46(3):314-329.

Hambledon, R.K., & Rodgers, H. J. (1995). Item Bias Review. Practical Assessment, Research and Evaluation. Retrieved in October 2015 at 1600 hours from http://pareonline.net/getvn.asp?v=4&n=6.

Hambledon, R.K., & Swaminathan, H. (1985). *Item Response Theory Principles and Application*. Nijhoff Publishing. Retrieved on 10th January 2015 at 1700 hours from http//ericae.net/ft/tamu/biaspub2.htm.

Hambleton, R.K., Swaminathan, H., & Rodgers, H.J.(1991). *Fundamentals of Item Response Theory*. Newbury Park, C A: Sage Press. Retrieved on January 16th 2012 at 2255 hours from http://www.books.google.co.ke/books?isbn=0805861769.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30*, 143-155.

Han, K. T., & Hambleton, R. K. (2009). User's manual for WinGen: Windows software that generates IRT model parameters and item responses. Center for Educational Assessment Research Report No. 642. University of Massachusetts.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responces. *Applied Psychological Measurement*, 31(5), 457- 459. doi1177/0146621607299271

Hernández, A., & González-Romá, V. (2003). Evaluating the multiple-group mean and covariance structure model for the detection of differential item functioning in polytomous ordered items. *Psichtema*, 15, 322-327. Retrieved on 19th June 2013 at 1100 hours from http://epm.sagepub.com/content/64/6/903.

Hidalgo, M. D., & Lopez-Pina, J.A. (2004). Differential item functioning detection and effect size: a comparison between Logistic Regression and Mantel-Haenszel 137 procedures. *Educational and Psychological Measurement*, 64(6), 903-915. DOI: 10.1177/0013164403261769.

Holland, P.W., & Thayer, H. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.129-145). Hillsdale, NJ: Lawrence Erlbaum Associates. Retrieved on 12th April 2009 at 1200 hours from http://www.books.google.co.ke/books?isbn=1109103204.

Hox, J. J. (2002). *Multilevel analysis***.** Mahwah, New Jersey: Lawrence Erlbaum Associates. Retrieved in 2010 at 0958 hours from http://www.questia.com.

Huberty, C. (2002).A history of effect size indices. *Educational and Psychological Measurement*, 62, 227-240. Retrieved on 1st June 2009 at 1123 hours from http://www.thefreelibrary.com.

Jiang, H., & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics, 23*, 291-322.

Jodoin, M. G., & Gierl, M.J. (2002). Evaluating type I error and power rates using an effect size measure with the Logistic Regression procedure for DIF detection. *Applied*

*Measurement in Education*, 14, 329-349. Retrieved on 4th on November 2011 at 2100 hours from http://www.tandfonline.com/doi/full/10.1080/15305058.2011.60281.

John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 461-494). New York, NY: Cambridge University Press.

Kathleen, M. M.,Clauser, B. E. & Hambleton, R.K. (1992). The Effect of Sample Size on the Functioning of the Mantel-Haenszel Statistic.*Educational and Psychological Measurement*, *52*(2), 443-451. Retrieved on 30th March 2017 at 1000 hours from http://journals.sagepub.com/doi/abs/10.1177/0013164492052002020.

Kerlinger, F, N. (1986**).** *Foundations of Behavioral Research*. CBS publishing New York. Retrieved on 21st January 2013 at 0845 hours from http/www.ets.org/media/AboutETS/pdf/overview.pdf..

Khalid, N, M. (2011). The performance of Mantel-Haenszel procedures in the identification of DIF items. *International Journal of Educational Sciences* 3(2), 435-447.

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B. D. (2005**).** A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, 65, 935. Retrieved on 9th May 2010 at 2033 hours from http://epm.sagepub.com/content/65/6/935.refs.

Kubiak, A. T., & Colwell, W. R. (1990). *Using multiple statistics with the same items appearing in different test forms.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston.

Lau, A.C & Arce, A.J. (2011). Comparing Methods for Detecting Unstable Anchor Items with Net DIF and Global DIF Conceptions. *American Educational Research Association*. New Orleans, Louisiana.

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*, 647-677.

Lopez Rivas, G.E. (2012). Detection and Classification of DIF types using Parametric and Non Parametric methods: A comparison of the IRT-Likelihood Ratio Test, Crossing-SIBTEST and Logistic Regression procedures. A PhD Dissertation submitted to the Department of Psychology, University of South Florida.

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. Retrieved on 17th April, 2013 0520 hours from www. *prezi.com/m1u58qcnpxbc/untitled-prezi/*at.

Mazor, K.M., Hambleton, R.K., & Clauser, B.E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22* (4): 357-367.

McCarty, F. A., Oshima, T. C., & Raju, N.S. (2007). Identifying Possible Sources of Differential Functioning Using Differential Bundle Functioning with Polytomously Scored Data**.** *Applied Measurement in Education*, 20(2), 205–225. Retrieved on 20th May 2011 at 2250 hours from http**://**education.gsu.edu/coshima/.../McCarty.

Miller, T.A., & Spray, J. A. (1993). Logistic Discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, 30,(2), 107-122. Retrieved on 12th July 2013 at 1440 hours from http//wwww.ets.org/research/contract.html.

Millsap, R. E. & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

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

Osterlind, S. J. & Everson, H. T. (2009). *Differential item functioning.* Thousand Oaks, CA: Sage Publishing.

Othuon, L. O. A. (1998).*The accuracy of parameter estimates and coverage probability of population values in regression models upon different treatments of systematically missing data.* Unpublished PhD thesis. University of British Columbia.

Özlem, Y & Özbek, B. (2016). A comparison of four differential Item functioning procedures in the presence of multidimensionality. *Educational Research and Reviews* 11(13), 1251- 1261.

Pae, T. I. (2004**).** DIF for examinees with different academic backgrounds. *Language Testing* 21, 53-73. Retrieved on 3rdJuly, 2011 at 0545 hours from http://ltj.sagepub.com/content/21/1/53.refs.

Parshall, C. G., & Miller, T. R. (2004). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small sample conditions. *Journal of Educational Measurement*, 32(3), 302-316. Retrieved on 19th June 2014 at 1650 hours from http://scholar.lib.vt.duu/theses/.../RAA_ETD.pdf.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias**.** In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics Psychometrics* (Vol. 26, pp. 125–167). Amsterdam: Elsevier. Retrieved 27th June 2011 at 1940 hours from http://samianstats.files.wordpress.com/.../handbook-of-statistics-vol-261.pd...

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*(3), 187-210.Pedrajita, Q J., & Talisayon, V.M. (2009). Identifying Biased Test Items by Differential Item Functioning Analysis Using Contingency Table Approaches: A Comparative Study. Education Quarterly, *University of the Philippines College of Education* Vol. 67 (1), 21-43.

Raju, N. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological. Measurement*. 14(2):197-207.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed.**)**. Thousand Oaks, CA: Sage Publications, Inc. Retrieved on 8th January 2013 at 1320 hours from http://www.psych.umass.edu/.../Cincinatti_syllabus.pdf.

Roever, C. (2005). *"That's not fair!" Fairness, bias, and differential item functioning in language testing*. Retrieved on March 18th 2009 at 0514 hours from: http://www2.hawaii.edu/~roever/brownbag.pdf.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning**.** *Applied Psychological Measurement*, 17, 105-116. Retrieved on 21st April 2013 at 1700 hours from http://apm.sagepub.com/content/17/2/105.refs.

Roussos, L.A., & Stout, W.F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Hanszel Type I error performance. *Journal of Educational Measurement*, 33, 215-230. Retrieved on 22nd December 2010 at 0200 hours.
from http://www.epm.sagepub.com/content/70/6/961 refs

Salubayba, T. M. (2013). *Differential item functioning detection in reading comprehension test using Mantel-Haenszel, Item response Theory, and logical data analysis.* The international Journal of social sciences, 14(1), 76-82.

Sarkar, S.K., Midi, H. and Imon, R. (2011). Diagnostics of fitted binary logistic Regression model based on individual subjects and covariate patterns. *International Journal of Applied Mathematics*,23: 63-81.

Schumacher, R. (2005). *Test bias and differential item functioning***.** Retrieved on 2nd March 2011 at 2300 hours from http://www.appliedmeasurementassociates.com/White Papers/TEST Bias and Differential Item Functioning.pdf.

Shealy R, Stout WF (1993). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (197-239). Hillsdale NJ: Erbaum.

Su, Y., & Wang, W. (2005). Efficiency of the Mantel, Generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items**.** *Applied Measurement in Education*, 18 (4), 313-350. Retrieved on 21st April 2013 at 1215 hours from http://epm.sagepub.com/content/68/6/940.refs.

Swaminathan, H., & Rogers, H. J. (1990**).** Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370. Retrieved on 2nd March 2011 at 0256 hours from http://www.jstatsoft.org/v39/i08/paper.

Tan, X., & Gierl, M. J. (2005). *Using local DIF analyses to assess group differences on multilingual examinations*. Poster presented at the annual meeting of the National Council on Measurement in Education. Montreal, QC, Canada. Measurement. Retrieved on 21st November 2012 at 2000 hours from http://www.ualberta.ca/~mgierl/files/Cv2012.pdf.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider**?** *Journal of Counseling and Development*, 80, 64-71. Retrieved on 1st September 2012 at 1340 hours from http://www.nccbi.nlm.nih.gov/pmc/.../PMC3173710/

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.

Walker, C.M. (2011). A review of DIFACK: Dimensionality-based IRT analysis package. *International Journal of Testing*, 4, 305-317.

Wang, W. C., & Yeh, L. Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, 27, 479-486. Retrieved on 25th June 2015 at 1830 hours from http//www.etds.lib.nchu.edu.tw/.../detail.

Wang, W., & Su, Y. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of Differential Item Functioning in polytomous items. *Applied Psychological Measurement*, 28(6), 450-480. Retrieved on 4th May 2012 at 1345 hours from http://www.apm.sagepub.com/content/34/3/166.refs.

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

Werner, J., Raj E., & Wynne W. (2002).Generating Non-normal Data for Simulation of Structural Equation Models Using Mattson's Method**.**, 37 (2), 227-244, Lawrence Erlbaum Associates, Inc. *Multivariate Behavioral Research*. Retrieved on 18th November 2009 at 1000 hours from http://www.reinartz.unkoeln.de/researchpapers/.

Wiberg, M. (2009). Differential Item Functioning in Mastery Tests; A comparison of Three Methods Using Real Data. *International Journal of Testing*, 9:41-59, DO1|: 10.1080/15305050902733455.

Zeiky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.Holland & H. Wainer (Eds**.),** *Differential item functioning* (pp. 337-347). Hillsdale,NJ: Lawrence Erlbaum Associates, Inc. Retrieved on 19th June 2008 at 2000 hours from http://scholar.lib.vt.edu/theses/.../etd.../RAA_ETD.pdf.

Zhang, D. (2005). *A Monte Carlo Investigation of Robustness to Non normal incomplete Data of Multilevel Modeling*. (Doctoral Dissertation), University of International Business and Economics, China. Retrieved on 1st January 2014 at 0700 hours from http://www.digitalcommons.uconn.edu/cgi/viewcontent.cgi?article=1001&context....

Zheng, Y., Gierl, M. J., & Cui, Y. (2007). *Using real data to compare DIF detection and effect size measures among Mantel-Haenszel, SIBTEST, and Logistic Regression procedures*. Thesis in EducationalPsychology. Retrieved on 12th November 2012 at 2100 hours from http://etda.libraries.psu.edu/theses/approved/WorldWideFiles/ETD-312/hglmdif.pdf.

Zumbo, B.D., & Thomas, D.R. (1996). *A measure of DIF effect size using logistic regression procedures***.** Paper presented at the National Board of Medical Examiners, Philadelphia. Retrieved on 19th September 2012 at 0230 hours from http://www.educ.ubc.ca/faculty/**zumbo**/cv.htm.

Zumbo, B. D. (1999). Logistic Regression Modeling as a unitary framework for Binary and Likert-type (ordinal) Item scores**.** *A Handbook on the Theory and Methods of Differential Item Functioning* (DIF) Ottawa, Canada, K1A 0K2. Retrieved 19th September, 2012 at 2100 hours from http://www.educ.ubc.ca/faculty/**zumbo**/cv.htm.

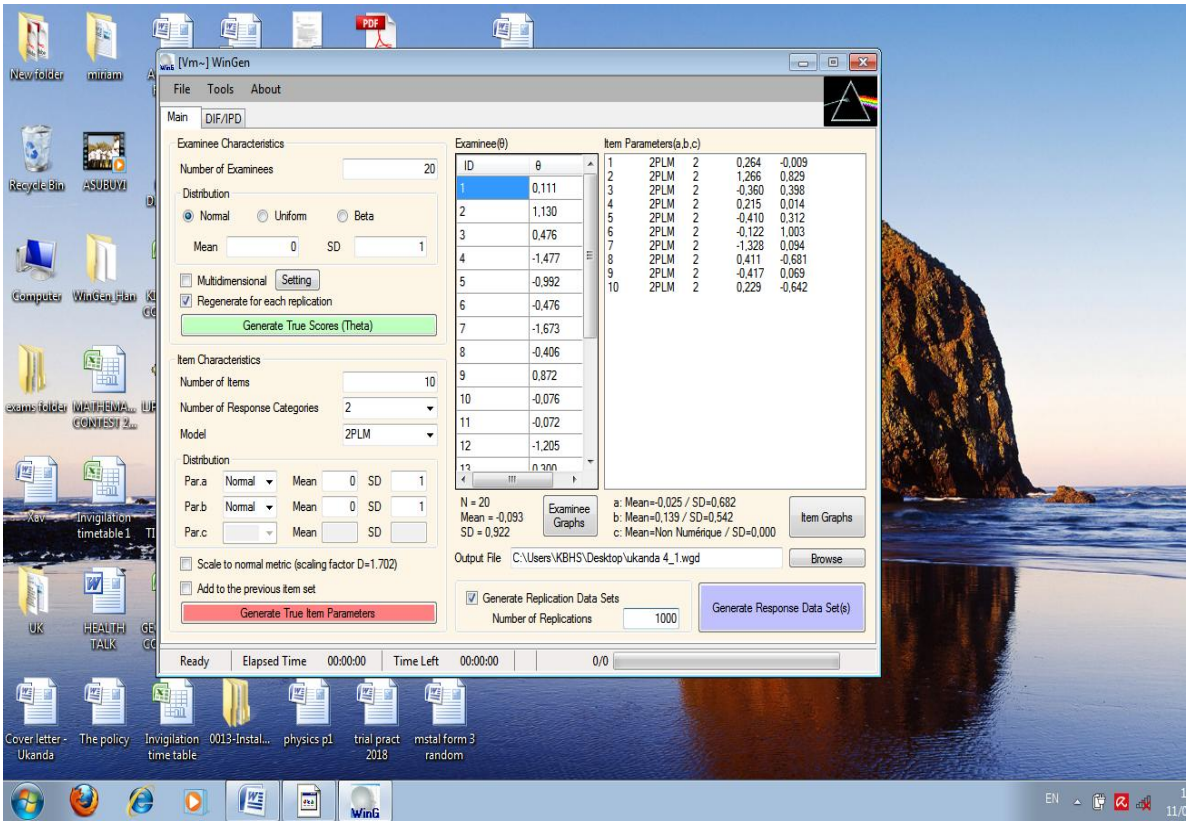
Zwick,R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP History Assessment. *Journal of Educational Measurement*, 26(1), 55-66. Retrieved on 23rd July 2015 at 1532 hours from https://kuis.repo.nii,ac.jp.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: rules, minimum sample size requirements, and criterion refinement* (ETS RR-12-08) NJ: ETS. Online http://www.ets.org/research/contact.html.

# APPENDICES

## APPENDIX A

## DATA COLLECTION INSTRUMENT




Screen shot of the WINGEN 3 computer software used in data generation.

The screenshot of WINGEN 3 software used in data generation is shown in the figure above. The main window consists of examinee characteristics which included the number of examinees and the Ability distribution in terms of mean and standard deviation. It also consists of item characteristics which include the number of items, the number of response categories, the model to be used i.e. 1PLM, 2PLM, 3PLM or non-parametric. The distribution in terms of parameter a, b and can be selected. When appropriate entries are made, true scores and true item parameters can be then generated. Replication data sets and

response data sets can also be generated. The software allows examinee graphs and Item graphs to be displayed. The DIF/IPD window consists of introduction to DIF/Item parameter drift via the direct input mode or the multiple file read in mode. This consists of data files for the reference group/test 1 and focal group's later tests. The examinee and item parameter output files can be saved in the comma separated values (csv) extension.

## APPENDIX B

## SAMPLE ITEM RESPONSE DATA GENERATED USING WINGEN 3 SOFTWARE

a) Sample size 60, 10 items, Mean 0,SD 1

1 ,1,0,1,0,0,0,1,1,1,1
2 ,1,0,0,0,0,0,1,1,1,0
3 ,0,0,0,0,0,1,0,1,1,0
4 ,1,0,1,0,1,0,0,1,1,1
5 ,1,0,0,1,1,0,0,1,0,0
6 ,1,1,1,0,1,0,1,0,1,1
7 ,1,1,1,0,0,1,1,1,1,1
8 ,0,1,0,0,0,1,0,1,1,0
9 ,1,0,0,0,1,1,0,1,0,1
10 ,1,0,0,0,0,0,0,0,1,1
11 ,0,1,1,0,0,0,1,1,1,1
12 ,1,1,1,0,0,1,1,1,1,1
13 ,0,1,0,0,1,1,1,0,0,1
14 ,0,0,0,0,1,1,1,1,1,1
15 ,1,1,0,0,1,0,0,1,1,0
16 ,0,0,0,1,0,0,0,1,1,0
17 ,0,1,1,0,0,0,0,1,1,1
18 ,1,0,1,0,0,0,1,1,0,1
19 ,0,0,0,0,0,0,0,1,1,1
20 ,1,1,1,0,1,0,1,1,1,1
21 ,0,0,1,0,0,0,0,1,1,1
22 ,0,0,1,0,0,0,0,1,1,0
23 ,0,0,0,0,1,0,1,1,1,0
24 ,1,0,1,0,1,1,1,1,1,1
25 ,1,1,1,0,0,1,1,0,1,1
26 ,0,1,1,0,0,0,0,1,1,1
27 ,0,0,0,1,1,0,0,1,1,1
28 ,0,0,1,0,0,1,0,0,1,1
29 ,0,0,0,0,1,1,1,1,1,1
30 ,0,1,0,0,1,0,1,0,0,1
31 ,1,0,1,0,0,1,0,1,1,1
32 ,0,1,1,0,0,1,1,1,1,1
33 ,1,1,1,0,0,1,1,0,0,1
34 ,0,0,0,1,0,1,0,1,1,0
35 ,0,0,1,0,0,1,1,1,0,1
36 ,0,1,1,0,0,0,0,1,1,0
37 ,1,0,0,0,0,1,1,1,1,1
38 ,1,1,1,0,1,0,0,1,1,1
39 ,0,0,1,0,0,1,0,1,1,1
40 ,1,1,0,0,1,0,1,1,1,1
41 ,0,1,0,0,0,0,0,1,1,1
42 ,1,0,0,0,1,1,1,1,1,1
43 ,0,1,1,0,1,1,0,1,1,1
44 ,1,1,0,0,1,0,1,1,0,0
45 ,1,1,0,0,1,1,0,1,1,1
46 ,0,1,1,0,1,0,0,1,1,0
47 ,1,0,0,0,0,0,0,1,1,1
48 ,1,0,0,0,0,0,0,0,1,1
49 ,1,1,1,0,1,1,1,1,0,1
50 ,0,1,0,0,1,1,1,0,0,1
51 ,0,1,0,0,0,0,1,1,1,1
52 ,0,1,1,0,0,0,1,1,1,1
53 ,1,1,1,0,1,1,1,0,0,1
54 ,1,0,1,0,0,0,0,0,1,0
55 ,0,1,0,0,0,1,1,0,1,1
56 ,0,1,0,0,0,1,1,0,1,1
57 ,1,1,0,0,0,0,1,0,1,1
58 ,0,1,1,0,1,1,1,1,1,1
59 ,0,1,0,0,0,0,0,1,1,1
60 ,0,1,0,0,0,1,0,1,1,0

b) Sample size 20,10 items, Mean 0, SD 1

1 1,1,1,0,0,1,0,0,0,1,
2 0,1,1,0,1,1,0,0,1,1,
3 1,1,0,0,1,0,0,1,0,1,
4 1,0,1,1,1,0,1,1,1,1,
5 1,1,0,1,1,0,1,1,0,1,
6 0,1,1,0,1,1,0,0,0,1,
7 1,1,1,0,1,0,0,1,1,1,
8 1,1,1,1,0,1,0,0,0,0,
9 0,1,0,0,1,1,1,0,1,0,
10 0,1,1,0,1,1,0,0,1,0,
11 0,1,1,0,0,0,0,0,1,1,
12 0,1,1,0,0,0,0,0,1,1,
13 0,1,1,1,0,1,0,0,0,0,
14 1,1,1,0,0,1,0,0,0,0,
15 0,1,0,0,1,0,0,0,1,1,
16 1,1,0,0,0,0,0,0,0,0,
17 0,0,1,0,0,0,0,0,0,1,
18 0,1,0,0,1,1,1,0,0,0,
19 1,0,0,1,1,0,1,1,0,0,
20 1,1,0,1,1,0,1,0,1,1,

c) Sample size 1000, 10 items, Mean 0, SD 1

| | | | | | |
|---|---|---|---|---|---|
| 1 | ,1,1,0,0,0,1,1,0,1,1 | 42 | ,1,0,0,1,1,1,0,1,0,0 | 83 | ,1,1,0,0,0,0,0,1,0,0 |
| 2 | ,1,1,0,0,0,1,1,1,0,1 | 43 | ,1,0,1,0,0,0,1,0,1,0 | 84 | ,0,0,0,1,1,1,0,0,1,1 |
| 3 | ,1,1,1,1,0,0,1,0,1,0 | 44 | ,0,1,0,1,1,1,1,1,1,1 | 85 | ,1,0,1,1,1,1,1,1,1,1 |
| 4 | ,0,1,0,0,0,1,0,0,1,0 | 45 | ,0,0,1,0,0,0,1,1,1,1 | 86 | ,1,0,0,1,1,1,1,1,0,1 |
| 5 | ,1,0,1,1,0,1,1,1,0,1 | 46 | ,0,0,0,0,0,1,0,1,0,1 | 87 | ,1,0,0,0,0,0,0,1,1,1 |
| 6 | ,1,1,0,0,1,0,1,0,0,0 | 47 | ,0,0,1,1,0,1,0,1,0,1 | 88 | ,1,0,0,0,0,1,1,0,0,1 |
| 7 | ,0,1,1,1,0,1,1,0,0,1 | 48 | ,1,1,1,1,0,0,0,0,0,1 | 89 | ,1,1,0,1,1,1,1,1,1,1 |
| 8 | ,1,0,1,0,0,0,1,0,1,1 | 49 | ,1,0,0,1,1,1,1,0,0,1 | 90 | ,1,1,1,1,1,0,1,0,0,1 |
| 9 | ,1,1,1,0,0,0,0,0,0,0 | 50 | ,0,0,0,0,0,1,1,1,0,1 | 91 | ,1,1,1,0,1,0,0,1,1,1 |
| 10 | ,0,0,0,1,0,1,1,1,1,1 | 51 | ,1,1,1,1,1,0,1,1,1,0 | 92 | ,1,0,1,0,0,0,0,1,0,0 |
| 11 | ,0,1,1,0,0,0,1,0,1,0 | 52 | ,0,0,1,1,1,1,1,1,1,1 | 93 | ,0,1,0,1,1,0,0,0,1,1 |
| 12 | ,0,0,0,0,0,1,0,1,0,1 | 53 | ,1,0,0,0,0,1,0,1,1,1 | 94 | ,0,0,0,1,1,0,1,0,0,0 |
| 13 | ,1,0,1,1,1,1,0,1,0,1 | 54 | ,1,1,0,1,1,0,0,1,0,1 | 95 | ,1,1,0,0,0,1,1,0,0,0 |
| 14 | ,1,0,0,1,0,1,0,1,1,0 | 55 | ,1,1,0,1,0,0,0,0,0,0 | 96 | ,1,1,1,0,0,0,0,0,0,0 |
| 15 | ,0,0,1,1,0,1,1,0,0,1 | 56 | ,1,0,1,1,0,1,1,0,1,1 | 97 | ,1,0,0,0,0,1,1,1,0,1 |
| 16 | ,1,0,0,0,0,1,0,1,1,1 | 57 | ,0,0,1,0,0,0,1,1,1,1 | 98 | ,1,0,1,1,0,1,1,1,1,1 |
| 17 | ,1,0,0,1,0,1,1,1,1,1 | 58 | ,0,0,1,1,0,1,0,1,1,1 | 99 | ,1,0,1,1,0,1,1,0,0,1 |
| 18 | ,0,1,0,1,0,1,0,0,1,1 | 59 | ,1,1,0,0,0,0,0,0,1,1 | 100 | ,0,0,0,0,0,0,0,1,0,1 |
| 19 | ,1,1,1,1,0,1,1,0,1,1 | 60 | ,1,1,1,0,1,0,1,1,0,1 | 101 | ,1,0,0,1,1,1,1,0,1,1 |
| 20 | ,1,0,0,0,1,0,1,0,1,1 | 61 | ,1,1,1,1,0,0,1,0,1,1 | 102 | ,1,0,1,1,1,0,1,0,0,1 |
| 21 | ,1,1,1,0,0,0,1,0,0,0 | 62 | ,1,0,0,1,1,1,0,1,1,1 | 103 | ,0,0,1,0,1,1,1,1,1,1 |
| 22 | ,1,0,0,0,0,0,1,0,1,1 | 63 | ,0,0,1,1,1,0,1,1,1,1 | 104 | ,1,1,0,0,0,1,0,1,1,0 |
| 23 | ,1,1,1,0,1,0,0,1,1,0 | 64 | ,0,1,1,0,1,0,0,1,0,1 | 105 | ,1,0,0,0,1,1,1,1,1,1 |
| 24 | ,1,0,0,0,1,1,1,1,0,1 | 65 | ,1,1,1,1,1,1,0,1,0,0 | 106 | ,1,1,0,0,0,0,1,0,1,1 |
| 25 | ,1,1,1,0,1,0,0,1,0,0 | 66 | ,0,0,1,0,1,1,1,1,1,1 | 107 | ,0,0,0,1,1,1,0,1,0,1 |
| 26 | ,1,1,1,1,0,1,1,0,1,1 | 67 | ,0,0,0,1,0,1,1,1,1,1 | 108 | ,1,1,1,0,0,1,1,0,0,1 |
| 27 | ,0,1,0,1,1,1,0,0,0,1 | 68 | ,1,0,1,1,1,1,1,0,1,0 | 109 | ,1,1,1,0,0,0,0,0,0,0 |
| 28 | ,1,0,1,1,1,1,0,1,0,1 | 69 | ,1,1,1,0,1,0,0,1,1,1 | 110 | ,1,0,0,0,0,0,0,0,1,0 |
| 29 | ,0,0,1,0,0,1,1,0,0,1 | 70 | ,0,0,0,0,0,0,0,0,0,1 | 111 | ,1,1,1,1,1,1,0,1,1,1 |
| 30 | ,0,0,0,1,1,1,1,1,0,1 | 71 | ,0,0,1,0,0,0,1,0,0,1 | 112 | ,1,0,1,0,0,1,1,1,0,1 |
| 31 | ,1,0,1,0,1,1,1,1,1,1 | 72 | ,0,0,0,1,1,1,1,1,0,1 | 113 | ,1,1,0,0,0,0,0,0,1,1 |
| 32 | ,1,0,1,0,0,1,1,0,0,1 | 73 | ,0,0,1,1,0,0,1,1,1,1 | 114 | ,1,0,1,0,0,0,1,0,0,1 |
| 33 | ,1,1,0,0,1,0,0,1,1,0 | 74 | ,0,1,1,0,1,1,1,0,1,0 | 115 | ,1,0,0,1,1,0,0,1,0,1 |
| 34 | ,0,0,1,0,0,0,0,1,0,1 | 75 | ,1,0,0,1,0,1,1,1,0,1 | 116 | ,1,0,0,0,0,0,1,1,0,0 |
| 35 | ,1,1,1,1,0,1,1,0,0,1 | 76 | ,1,0,0,0,1,1,0,1,1,0 | 117 | ,1,0,0,1,1,1,1,1,0,1 |
| 36 | ,1,1,1,0,0,0,0,1,0,1 | 77 | ,0,0,0,1,0,1,0,0,1,1 | 118 | ,1,1,1,0,0,0,0,1,1,1 |
| 37 | ,0,1,1,1,0,1,0,1,0,1 | 78 | ,1,0,0,0,0,0,1,1,1,0 | 119 | ,0,0,1,1,0,1,0,1,1,1 |
| 38 | ,1,0,0,0,0,0,1,0,1,1 | 79 | ,0,0,1,1,0,0,1,1,1,1 | 120 | ,1,0,0,1,1,0,1,0,1,0 |
| 39 | ,1,0,1,0,0,0,1,1,0,1 | 80 | ,1,1,1,0,0,1,1,0,0,1 | 121 | ,0,0,0,0,1,1,0,0,0,1 |
| 40 | ,0,0,1,1,1,1,1,0,0,1 | 81 | ,1,1,1,0,0,0,1,0,0,0 | 122 | ,1,0,1,0,1,0,1,1,1,0 |
| 41 | ,1,1,1,0,0,0,1,0,1,0 | 82 | ,0,1,0,0,1,0,0,0,1,0 | 123 | ,0,1,1,1,0,1,1,1,1,1 |
| 124 | ,1,1,0,0,1,0,1,0,1,0 | 167 | ,0,0,0,0,1,0,1,0,1,1 | 210 | ,1,1,1,1,1,0,0,0,1,0 |
| 125 | ,0,0,1,1,1,1,1,1,0,1 | 168 | ,1,0,1,1,1,1,1,0,1,1 | 211 | ,1,1,1,0,0,0,1,0,1,0 |

126 ,1,0,0,0,0,0,0,1,0,0
127 ,1,0,1,1,0,1,0,1,0,1
128 ,1,0,1,0,1,0,0,1,1,1
129 ,0,0,1,1,0,0,1,1,1,1
130 ,1,0,0,0,1,1,1,1,0,1
131 ,1,1,1,0,1,0,0,0,0,1
132 ,1,1,0,0,1,0,0,0,1,1
133 ,1,1,0,1,1,0,0,1,0,0
134 ,1,0,1,1,1,0,1,0,0,1
135 ,1,0,1,1,0,0,1,1,1,0
136 ,0,0,0,1,0,1,1,0,1,1
137 ,1,1,0,1,0,0,0,1,1,0
138 ,1,1,1,0,1,0,1,0,0,1
139 ,0,0,1,1,1,1,1,1,1,1
140 ,0,0,0,1,0,1,1,1,0,1
141 ,0,0,1,1,1,1,0,1,1,1
142 ,0,1,0,1,1,1,0,0,0,1
143 ,1,1,1,1,1,0,1,0,1,1
144 ,0,0,0,1,0,1,0,1,0,0
145 ,1,1,1,1,0,0,0,1,1,1
146 ,0,0,1,1,0,1,1,1,1,1
147 ,1,0,0,1,0,1,0,1,1,1
148 ,0,0,1,0,1,0,1,1,1,1
149 ,0,1,0,0,1,1,0,1,0,0
150 ,0,0,0,1,1,1,1,1,0,1
151 ,0,0,0,0,0,0,1,1,1,1
152 ,1,1,0,0,0,0,1,0,1,1
153 ,0,0,0,1,0,1,1,1,0,1
154 ,0,0,0,0,0,0,1,1,0,1
155 ,1,1,1,1,0,0,0,1,1,0
156 ,1,0,0,1,1,1,1,1,0,1
157 ,1,1,1,0,1,0,0,1,1,1
158 ,0,1,0,0,0,0,0,0,1,0
159 ,1,1,0,0,0,1,0,0,0,1
160 ,1,1,1,0,1,0,0,0,0,0
161 ,1,0,1,1,0,1,1,0,0,0
162 ,1,1,1,0,0,1,1,0,1,0
163 ,1,0,1,1,0,1,1,1,1,1
164 ,0,0,1,0,0,1,1,1,1,1
165 ,1,0,0,1,0,1,0,1,1,1
166 ,1,0,0,1,1,0,1,1,0,1
167 ,1,0,0,1,0,1,1,1,0,1
168 ,0,0,0,0,1,1,0,0,1,1
169 ,1,0,1,0,1,0,0,1,1,1
170 ,1,1,0,1,0,0,0,1,1,0
171 ,0,0,0,1,1,1,0,0,0,1
172 ,1,1,1,0,1,0,0,0,1,1
173 ,1,1,1,0,0,1,0,0,1,1
174 ,1,1,1,1,0,0,1,1,0,0
175 ,1,1,0,0,1,1,0,0,0,1
176 ,1,1,0,0,0,1,1,1,0,1
177 ,1,0,0,1,1,1,1,1,1,1
178 ,0,0,0,0,1,1,1,0,0,1
179 ,1,0,0,1,0,1,1,1,0,1
180 ,1,0,1,1,0,0,0,1,1,1
181 ,1,0,0,1,1,1,0,0,1,1
182 ,1,1,0,1,1,1,1,1,1,1
183 ,1,1,1,1,1,1,1,1,0,1
184 ,1,0,0,1,1,0,0,0,1,0
185 ,1,0,1,1,0,1,0,1,1,1
186 ,0,0,0,1,1,0,1,1,0,1
187 ,1,0,0,0,1,1,0,1,1,1
188 ,0,1,0,0,0,1,1,0,1,1
189 ,0,0,1,1,1,1,1,1,0,1
190 ,0,1,1,1,1,1,1,0,1,1
191 ,0,1,1,0,1,0,1,1,0,1
192 ,1,0,1,0,1,0,1,0,1,1
193 ,1,1,0,0,1,0,1,1,0,1
194 ,0,0,1,1,0,1,1,1,0,1
195 ,1,1,1,1,1,0,0,0,1,0
196 ,1,1,0,0,1,0,1,0,0,1
197 ,1,0,0,1,1,0,1,0,1,0
198 ,1,1,0,0,0,0,0,0,0,0
199 ,0,1,0,1,1,1,0,1,1,1
200 ,1,0,1,0,0,1,1,0,1,1
201 ,0,0,1,1,1,0,1,1,0,1
202 ,1,1,0,0,0,0,0,0,0,1
203 ,1,1,0,1,1,1,0,1,0,1
204 ,0,0,0,0,0,1,0,1,1,1
205 ,0,1,0,0,0,0,1,1,0,1
206 ,0,0,1,0,0,1,1,0,0,1
207 ,1,0,1,0,1,1,1,1,0,0
208 ,0,1,1,1,0,1,0,1,0,1
209 ,1,0,1,0,1,1,1,1,0,1
210 ,1,0,0,0,1,1,1,1,1,1
211 ,1,0,0,0,1,1,1,1,0,1
212 ,1,0,0,0,1,1,1,1,0,1
213 ,1,0,1,1,1,0,1,1,0,1
214 ,0,0,1,0,0,1,0,1,0,1
215 ,1,1,1,0,0,1,1,0,1,0
216 ,0,0,0,1,1,1,0,1,1,1
217 ,1,1,0,0,1,0,1,1,1,0
218 ,1,0,0,0,1,0,0,1,1,0
219 ,1,1,0,0,0,1,0,0,1,0
220 ,1,1,0,0,0,0,0,1,0,0
221 ,1,1,0,1,1,0,0,0,1,0
222 ,1,1,1,0,1,0,1,0,0,1
223 ,1,0,1,1,1,1,1,1,1,1
224 ,1,0,1,0,0,0,1,0,0,1
225 ,0,0,1,0,1,1,1,1,0,1
226 ,1,0,1,1,0,1,1,0,0,1
227 ,1,0,1,1,0,0,1,0,1,1
228 ,1,0,0,1,0,1,1,1,1,1
229 ,1,0,1,0,0,1,1,1,1,1
230 ,0,0,0,0,1,0,0,1,0,0
231 ,0,0,0,1,1,1,1,0,0,1
232 ,1,0,0,0,0,1,1,1,0,1
233 ,1,0,1,0,0,1,1,1,1,0
234 ,1,0,0,0,1,1,0,1,1,1
235 ,0,0,1,1,1,1,1,0,0,0
236 ,1,1,0,0,0,1,0,0,0,0
237 ,0,0,1,0,1,0,1,0,1,1
238 ,0,0,1,1,1,1,1,0,1,1
239 ,1,0,1,1,1,1,1,0,0,1
240 ,1,1,0,1,1,0,0,1,1,1
241 ,0,0,1,1,1,1,1,0,1,1
242 ,1,1,0,0,0,0,0,0,1,1
243 ,1,1,0,0,1,1,1,1,1,1
244 ,1,1,1,0,0,0,0,1,0,1
245 ,1,1,0,0,1,0,0,0,1,1
246 ,0,0,0,1,0,1,1,1,0,0
247 ,0,0,0,0,0,1,1,1,0,1
248 ,1,0,0,0,0,1,0,1,1,1
249 ,1,0,1,1,0,0,1,0,1,0
250 ,1,1,0,0,0,0,1,0,1,0
251 ,1,1,1,0,1,1,0,0,1,0
252 ,0,0,1,0,1,1,0,1,0,1
253 ,1,0,0,1,0,1,1,1,0,1
254 ,0,0,0,0,1,1,0,0,1,1
255 ,1,1,0,0,1,1,1,0,0,1
256 ,1,0,1,1,0,1,1,1,0,1
296 ,1,1,1,0,1,0,1,1,0,0
297 ,1,0,0,0,1,1,1,1,0,1
298 ,1,0,1,1,0,0,0,0,1,1
299 ,1,1,1,0,1,0,1,0,0,1
339 ,1,1,1,1,0,1,0,1,0,1
340 ,1,0,1,1,0,0,0,1,1,0
341 ,0,0,0,1,1,1,0,1,0,1
342 ,0,0,0,1,1,1,0,1,1,1

257 ,0,1,0,0,0,0,0,0,0,0
258 ,0,0,0,0,0,1,1,1,1,1
259 ,0,0,0,1,0,1,1,1,0,1
260 ,0,1,0,1,0,1,1,0,1,1
261 ,1,0,0,0,1,1,1,0,1,0
262 ,1,1,1,0,0,0,0,1,1,1
263 ,1,0,0,0,1,1,1,1,1,0
264 ,1,1,1,0,1,0,1,0,0,1
265 ,0,1,1,0,1,1,0,1,1,1
266 ,1,1,0,0,1,1,0,1,0,1
267 ,1,0,0,0,1,0,0,0,1,0
268 ,1,0,1,0,0,1,0,0,0,1
269 ,1,1,0,0,0,0,1,1,1,1
270 ,1,1,1,0,1,0,1,0,1,1
271 ,1,0,0,0,1,0,0,0,1,1
272 ,1,1,0,1,1,1,1,0,1,1
273 ,1,1,0,0,1,1,0,0,1,1
274 ,1,1,1,0,1,0,0,1,0,1
275 ,1,1,0,0,1,0,1,1,0,1
276 ,1,0,1,0,1,0,0,0,1,1
277 ,0,0,0,1,1,1,1,0,0,1
278 ,1,1,1,1,0,0,0,0,0,0
279 ,1,1,1,0,1,0,1,0,0,0
280 ,0,1,1,1,0,1,1,1,1,1
281 ,0,1,0,0,1,1,1,0,0,1
282 ,0,0,1,0,1,1,0,1,1,1
283 ,1,1,1,1,0,1,0,1,1,0
284 ,1,0,0,0,1,1,1,1,0,1
285 ,1,1,1,0,1,0,0,0,0,1
286 ,1,1,1,1,0,1,0,1,0,0
287 ,1,1,1,1,0,1,0,1,0,0
288 ,1,0,0,1,1,0,0,0,0,1
289 ,1,0,1,0,0,0,1,1,1,1
290 ,1,1,1,0,0,0,0,1,1,0
291 ,1,1,1,0,0,0,1,0,0,0
292 ,1,0,1,1,1,1,1,0,1,1
293 ,1,1,0,0,1,0,1,0,1,1
294 ,1,1,0,0,1,0,0,0,1,0
295 ,0,0,1,1,1,1,0,0,1,1
296 ,1,1,1,0,0,0,1,0,1,1
297 ,1,1,0,0,0,0,0,1,1,0
298 ,0,0,0,0,0,0,0,1,1,1
299 ,0,1,0,1,1,1,0,1,1,1
300 ,1,1,0,0,0,0,1,1,1,0
301 ,0,0,0,0,1,0,0,1,0,1
302 ,0,0,1,1,0,0,0,0,0,1
303 ,1,0,1,1,1,0,1,0,0,1
304 ,1,1,1,0,1,1,0,1,1,0
305 ,0,1,1,1,0,1,1,1,0,1
306 ,0,1,1,1,1,1,1,1,1,1
307 ,0,1,0,0,0,0,0,0,1,1
308 ,1,1,1,0,0,1,0,1,1,1
309 ,1,0,1,1,0,0,1,1,0,1
310 ,1,0,0,0,1,1,1,1,1,1
311 ,1,0,0,0,0,1,1,1,0,1
312 ,1,1,1,1,0,0,0,0,0,1
313 ,0,1,1,0,1,0,1,0,1,0
314 ,1,0,0,1,0,1,1,1,0,1
315 ,1,0,0,0,0,1,1,1,1,1
316 ,1,0,1,1,0,1,0,1,1,1
317 ,0,0,1,0,0,1,1,0,1,1
318 ,0,0,1,0,1,1,0,1,1,1
319 ,1,0,1,0,0,1,1,0,1,1
320 ,1,1,0,1,1,0,1,1,0,0
321 ,0,0,0,1,0,1,1,1,0,1
322 ,0,0,1,1,0,1,1,1,1,1
323 ,1,1,1,0,1,0,0,0,0,1
324 ,1,1,0,0,0,0,0,1,1,1
325 ,0,0,0,1,1,1,1,0,0,1
326 ,0,0,1,1,0,1,1,1,0,1
327 ,0,0,1,0,1,1,1,0,1,1
328 ,1,0,0,0,1,1,1,1,0,1
329 ,0,0,0,1,0,1,1,1,1,1
330 ,1,1,1,0,0,1,0,1,0,0
331 ,1,0,1,1,0,1,1,1,1,1
332 ,1,0,1,1,0,1,1,0,1,1
333 ,1,1,1,0,1,1,0,1,0,1
334 ,1,0,1,1,1,1,0,0,0,1
335 ,1,1,0,0,0,0,1,0,1,0
336 ,1,1,1,0,0,0,1,0,1,1
337 ,1,1,0,0,0,0,0,0,1,0
338 ,0,1,1,0,1,0,0,0,1,1
339 ,0,1,1,0,1,0,0,0,1,1
340 ,1,1,1,0,0,0,1,0,1,1
341 ,1,1,0,0,0,0,0,1,1,0
342 ,1,1,0,1,1,0,1,1,0,1
343 ,0,1,1,0,1,1,1,1,1,1
344 ,1,1,1,0,1,0,1,1,1,0
345 ,1,1,1,1,1,0,0,1,1,1
346 ,1,0,0,1,0,1,1,1,1,1
347 ,1,1,0,1,1,0,0,0,1,1
348 ,0,1,0,1,0,0,1,1,0,0
349 ,0,1,1,1,1,1,0,1,0,1
350 ,1,1,0,1,0,0,1,1,1,1
351 ,1,0,0,1,0,1,1,1,1,0
352 ,1,1,0,0,1,1,0,0,1,1
353 ,0,0,1,1,0,1,1,1,1,1
354 ,1,1,0,0,0,0,0,0,1,0
355 ,1,1,1,0,0,0,1,0,1,0
356 ,0,1,0,0,1,0,1,0,0,1
357 ,0,0,1,0,1,1,1,1,1,0
358 ,1,1,0,1,0,0,1,0,1,0
359 ,1,1,1,0,0,0,0,0,0,0
360 ,0,0,1,1,1,1,0,1,0,1
361 ,1,0,0,0,0,0,0,0,0,0
362 ,1,1,0,0,1,0,1,1,1,0
363 ,0,1,1,0,1,1,0,1,0,1
364 ,1,0,0,0,0,0,0,0,0,1
365 ,0,0,1,1,1,1,1,0,1,1
366 ,0,0,1,1,0,1,0,0,0,1
367 ,1,1,1,0,1,0,1,0,0,1
368 ,1,1,1,0,1,0,0,0,0,1
369 ,1,1,0,1,1,0,1,0,1,1
370 ,1,1,1,0,0,0,0,0,1,0
371 ,0,0,1,0,0,1,1,0,1,1
372 ,1,0,1,1,1,1,0,1,1,1
373 ,1,0,1,0,0,1,1,1,0,1
374 ,1,0,1,0,0,1,1,0,0,1
375 ,1,0,0,0,1,0,1,0,0,0
376 ,1,1,0,0,0,0,1,0,1,1
377 ,0,0,0,1,0,1,0,1,0,1
378 ,1,1,1,0,1,0,0,0,0,0
379 ,1,1,0,1,0,0,0,1,0,0
380 ,1,0,1,0,0,0,1,1,1,1
381 ,1,0,1,0,0,1,1,0,1,0
382 ,1,1,1,0,0,0,1,0,1,1
383 ,1,1,0,0,0,0,1,1,0,1
384 ,0,0,0,0,0,0,1,1,1,0
385 ,0,1,0,1,1,1,0,1,1,1
386 ,1,1,0,0,1,1,1,0,0,1
387 ,1,0,1,0,1,0,0,0,0,1
388 ,1,0,1,0,1,0,0,0,0,1
389 ,1,0,1,0,0,0,1,1,1,1
390 ,1,1,1,0,0,0,0,1,1,0
391 ,1,1,1,0,0,0,1,0,0,0
392 ,1,0,1,1,1,1,1,0,1,1
393 ,1,1,0,0,1,0,1,0,1,1
394 ,1,1,0,0,1,0,0,0,1,0
395 ,0,0,1,1,1,1,0,0,1,1
396 ,1,1,1,0,0,0,1,0,1,1
397 ,1,1,0,0,0,0,0,1,1,0
398 ,0,0,0,0,0,0,0,1,1,1
399 ,0,1,0,1,1,1,0,1,1,1
400 ,1,1,0,0,0,0,1,1,1,0
401 ,0,0,0,0,1,0,0,1,0,1
402 ,0,0,1,1,0,0,0,0,0,1
403 ,1,0,1,1,1,0,1,0,0,1
404 ,1,1,1,0,1,1,0,1,1,0
405 ,0,1,1,1,0,1,1,1,0,1
406 ,0,1,1,1,1,1,1,1,1,1
407 ,0,1,0,0,0,0,0,0,1,1
408 ,1,1,1,0,0,1,0,1,1,1
409 ,1,0,1,1,0,0,1,1,0,1
410 ,1,0,0,0,1,1,1,1,1,1
411 ,1,0,0,0,0,1,1,1,0,1
412 ,1,1,1,0,0,0,0,0,1,0
413 ,0,1,0,1,0,1,0,1,0,1
414 ,1,0,1,0,1,0,0,1,1,1
415 ,1,1,0,0,0,0,1,0,1,0
416 ,1,0,1,0,1,0,1,0,1,1
417 ,1,1,0,0,0,0,1,0,1,0
418 ,1,0,0,0,1,1,1,1,1,1
419 ,1,1,0,1,1,0,1,0,0,0
420 ,1,1,1,0,0,1,1,0,1,1
421 ,1,0,0,0,1,0,1,0,0,0
422 ,1,0,0,0,1,1,1,1,0,1
423 ,1,1,0,1,1,0,1,0,0,1
424 ,1,0,1,1,0,1,1,1,1,1
425 ,1,1,0,1,0,1,0,0,0,1
426 ,1,0,1,0,1,0,1,0,1,1
427 ,1,1,0,0,0,0,1,0,1,0
428 ,1,0,1,1,0,0,0,1,1,1
429 ,1,1,0,1,1,0,1,0,0,0
430 ,1,1,1,0,0,1,1,0,1,1
431 ,1,0,1,0,1,0,1,0,1,1
432 ,1,0,1,0,1,0,1,0,1,1
433 ,1,1,1,0,0,1,1,0,1,1
434 ,1,0,1,1,1,1,1,0,1,1
435 ,1,1,0,0,0,0,1,0,1,0
436 ,1,1,1,0,0,0,1,0,1,1
437 ,1,1,0,0,0,0,0,0,1,1
438 ,0,1,1,0,1,0,0,0,1,1
439 ,0,1,1,0,1,0,0,0,1,1
440 ,1,1,1,0,0,0,1,0,1,1
441 ,1,1,0,0,0,0,0,1,1,0
442 ,0,0,0,0,0,0,0,1,1,1
443 ,0,1,0,1,1,1,0,1,1,1
444 ,1,1,0,0,0,0,1,1,1,0
445 ,0,0,0,0,1,0,0,1,0,1
446 ,0,0,1,1,0,0,0,0,0,1
447 ,1,0,1,1,1,0,1,0,0,1
448 ,1,1,1,0,1,1,0,1,1,0
449 ,1,0,1,1,0,1,1,1,0,1
450 ,1,1,0,1,0,0,1,1,1,1
451 ,1,0,0,1,0,1,1,1,1,0
452 ,1,1,0,0,1,1,0,0,1,1
453 ,0,0,1,1,0,1,1,1,1,1
454 ,1,1,0,0,0,0,0,0,1,0
455 ,1,1,1,0,0,0,1,0,1,0
456 ,0,1,0,0,1,0,1,0,0,1
457 ,0,0,1,0,1,1,1,1,1,0
458 ,1,1,0,1,0,0,1,0,1,0
459 ,1,1,1,0,0,0,0,0,0,0
460 ,0,0,1,1,1,1,0,1,0,1
461 ,1,0,0,0,0,0,0,0,0,0
462 ,1,1,0,0,1,0,1,1,1,0
463 ,0,1,1,0,1,1,0,1,0,1
464 ,1,0,0,0,0,0,0,0,0,1
465 ,0,0,1,1,1,1,1,0,1,1
466 ,0,0,1,1,0,1,0,0,0,1
467 ,1,1,1,0,1,0,1,0,0,1
468 ,1,1,1,0,0,1,1,1,0,1
469 ,1,0,1,0,0,1,0,1,0,1
470 ,0,0,0,0,0,1,0,1,1,1
471 ,1,0,0,0,1,0,1,0,0,0
472 ,1,0,0,0,1,1,1,1,0,1
473 ,1,1,0,1,1,0,1,0,0,1

388 ,0,0,1,1,0,1,0,1,1,1
389 ,1,1,0,1,1,0,0,0,0,1
390 ,1,0,0,1,0,1,1,1,0,1
391 ,1,0,1,1,1,1,1,1,0,1
392 ,1,1,1,1,1,0,1,0,0,1
393 ,1,1,1,0,1,1,0,0,0,0
394 ,1,0,1,0,1,1,1,1,1,1
395 ,1,1,1,0,1,1,1,1,1,1
396 ,0,0,0,1,1,1,1,1,0,1
397 ,1,0,1,1,0,1,1,1,0,1
398 ,1,1,0,1,1,1,1,1,1,1
399 ,1,0,0,1,1,1,0,0,1,1
400 ,1,0,0,1,1,1,0,1,0,1
401 ,1,1,0,0,0,0,1,0,1,1
402 ,0,0,1,0,1,0,1,0,0,0
403 ,1,1,1,0,1,1,1,1,0,1
404 ,1,1,1,0,1,1,1,0,0,1
405 ,1,0,1,1,0,1,0,1,0,0
406 ,1,0,1,1,0,0,0,0,1,1
407 ,1,0,1,0,0,0,0,1,0,0
408 ,1,1,1,1,1,0,1,0,1,0
409 ,1,1,0,1,1,0,1,0,0,0
410 ,1,1,0,1,0,0,1,1,1,1
411 ,1,1,1,0,0,0,1,0,0,1
412 ,0,0,0,0,1,1,0,1,1,1
413 ,1,1,0,0,0,0,0,0,1,1
414 ,1,1,1,1,1,0,0,1,1,0
415 ,0,0,0,0,1,1,1,1,1,1
416 ,1,1,1,0,1,1,1,0,0,0
417 ,0,0,0,0,1,1,1,1,0,0
418 ,1,1,0,0,1,0,1,0,1,0
419 ,1,0,1,1,1,1,0,1,1,0
420 ,1,1,0,0,0,1,0,1,1,1
421 ,1,1,0,0,1,1,0,1,1,0
422 ,1,0,0,0,1,0,1,0,1,1
423 ,1,1,0,1,1,0,0,1,1,1
424 ,1,1,0,0,0,0,0,0,1,0
431 ,0,1,1,1,0,1,0,0,1,0
432 ,1,0,0,0,0,0,0,0,0,0
433 ,0,1,0,1,0,0,1,0,1,1
434 ,0,0,1,1,0,0,1,1,1,1
435 ,0,0,1,0,0,1,1,1,0,1
436 ,0,0,0,0,0,1,0,0,0,1
437 ,1,1,0,1,1,1,1,0,1,1
438 ,0,1,0,0,1,1,0,0,1,1
439 ,0,0,1,0,0,1,1,0,1,1
440 ,1,1,0,0,0,0,1,0,0,0
441 ,1,1,0,0,1,0,0,0,1,1
442 ,0,0,1,0,1,0,1,1,0,0
443 ,1,1,1,0,1,0,0,1,0,1
444 ,1,0,0,0,0,1,1,1,0,1
445 ,1,0,0,0,1,0,1,0,1,1
446 ,0,0,0,0,1,0,1,1,1,0
447 ,1,0,1,1,0,1,1,1,0,1
448 ,1,0,1,0,0,1,1,0,1,1
449 ,0,0,0,1,0,1,1,0,0,1
450 ,1,0,0,1,1,1,1,0,1,1
451 ,0,1,0,1,0,1,1,1,0,1
452 ,1,0,0,0,0,1,1,1,0,0
453 ,1,1,1,1,0,0,0,0,0,1
454 ,0,0,1,1,1,0,0,0,0,1
455 ,1,1,0,0,0,0,1,1,0,1
456 ,1,1,1,0,1,1,1,1,1,0
457 ,0,0,1,1,0,0,1,0,0,1
458 ,1,1,0,1,0,1,1,1,0,0
459 ,1,0,1,1,1,1,1,1,1,1
460 ,1,1,1,0,0,0,1,1,1,1
461 ,1,1,0,0,1,0,0,0,0,1
462 ,1,1,0,1,0,1,0,0,0,1
463 ,1,1,1,1,0,0,1,0,0,1
464 ,1,0,0,0,1,0,1,1,1,1
465 ,1,1,1,1,0,0,0,0,1,0
466 ,1,1,0,0,1,1,0,1,0,0
467 ,1,1,0,1,0,1,0,0,1,1
468 ,1,1,0,1,0,1,0,1,0,1
469 ,1,1,0,1,0,1,0,1,1,1
470 ,1,1,0,1,0,1,0,1,1,1
471 ,1,1,0,1,0,1,0,1,1,1
472 ,1,1,0,1,0,1,0,1,1,1
473 ,1,1,0,1,0,1,0,1,1,1
474 ,0,0,0,1,0,0,1,1,1,0
475 ,1,0,1,1,1,1,0,1,1,1
476 ,1,0,0,1,1,1,1,1,1,1
477 ,1,0,0,1,1,0,0,0,0,1
478 ,1,0,0,1,0,0,1,0,0,1
479 ,1,0,0,0,1,1,1,1,0,1
480 ,1,1,0,1,1,0,0,0,1,0
481 ,0,0,1,0,0,0,1,1,0,1
482 ,0,0,1,1,1,1,1,1,0,1
483 ,1,1,0,0,1,0,0,0,1,0
484 ,1,1,0,0,0,0,0,0,0,1
485 ,1,0,0,1,1,1,1,1,0,1
486 ,1,1,1,0,0,0,0,0,0,1
487 ,1,1,1,1,0,1,0,1,1,0
488 ,1,0,0,0,1,1,1,1,1,1
489 ,1,0,0,0,0,1,1,1,1,1
490 ,1,0,1,1,1,1,0,1,1,1
491 ,0,0,1,0,0,0,0,0,1,1
492 ,1,1,1,0,1,0,0,0,0,0
493 ,0,0,0,1,1,1,1,1,0,1
494 ,0,1,1,0,0,0,1,0,1,1
495 ,0,1,0,0,0,1,1,0,0,1
496 ,1,0,1,1,0,1,0,0,0,1
497 ,1,1,1,0,0,1,0,0,0,1
498 ,1,1,1,1,0,0,1,0,1,1
499 ,1,1,1,0,0,0,0,0,0,1
500 ,0,1,1,0,1,0,0,1,0,1
501 ,0,0,1,1,1,1,1,1,0,1
502 ,1,0,1,1,0,1,0,1,0,1
503 ,1,1,1,0,0,0,1,1,1,1
504 ,1,1,1,0,0,1,0,0,1,0
505 ,1,1,0,0,1,0,1,1,0,1
506 ,1,0,1,0,0,1,1,0,0,1
507 ,1,1,1,0,0,0,1,0,1,0
508 ,1,0,0,0,0,1,0,1,0,1
509 ,1,1,1,0,1,1,1,0,1,1
510 ,0,0,1,0,1,0,1,1,1,1
511 ,1,0,1,0,0,1,0,0,0,1
512 ,0,1,1,1,0,1,1,1,0,1
513 ,1,0,0,0,0,1,1,1,0,1
514 ,1,0,0,1,0,1,1,1,1,1
515 ,1,0,1,1,1,1,0,1,1,1
516 ,1,1,1,0,1,0,0,1,0,1
517 ,1,1,1,0,0,1,1,0,0,1
518 ,0,0,1,0,0,1,1,1,0,1
554 ,1,0,0,0,1,0,1,0,1,1
555 ,1,1,1,0,0,0,1,1,0,0
556 ,1,1,0,1,0,1,1,1,1,0
557 ,1,1,0,0,1,0,0,0,0,0
558 ,1,1,1,1,0,1,1,0,1,1
559 ,0,0,0,0,1,1,1,1,0,1
560 ,1,1,0,0,0,0,1,0,1,1
561 ,0,0,0,0,1,1,1,1,1,1
597 ,1,1,1,1,0,1,0,0,0,1
598 ,1,1,0,0,1,0,0,0,1,1
599 ,0,0,0,1,1,1,1,1,0,1
600 ,0,1,0,1,0,1,0,0,0,1
601 ,1,1,0,1,1,1,0,1,0,1
602 ,1,1,0,1,0,1,1,0,1,0
603 ,1,1,0,0,1,1,1,0,1,1
604 ,0,0,0,1,0,1,1,1,1,1

519 ,1,0,0,0,1,1,1,0,0,1
520 ,1,0,0,1,1,1,1,1,1,1
521 ,1,1,0,0,0,0,0,1,1,1
522 ,1,1,1,0,1,0,0,0,1,1
523 ,1,1,1,0,0,1,1,1,1,0
524 ,0,0,0,0,1,1,1,1,0,1
525 ,1,0,0,0,1,1,0,0,0,1
526 ,1,0,1,1,0,1,1,0,1,0
527 ,1,1,1,1,0,0,0,0,0,1
528 ,0,0,1,0,1,1,1,0,1,1
529 ,1,1,1,0,0,1,1,1,1,1
530 ,0,0,1,0,0,1,1,0,0,0
531 ,1,1,1,1,1,0,1,0,0,1
532 ,0,0,0,1,1,1,1,0,0,1
533 ,0,0,0,1,0,1,1,1,0,1
534 ,1,1,1,0,1,1,0,1,0,1
535 ,1,1,1,0,0,1,1,0,1,1
536 ,1,1,0,0,0,0,1,1,1,0
537 ,0,0,0,1,1,1,0,1,1,1
538 ,1,1,0,1,1,0,1,0,0,0
539 ,0,0,1,1,1,0,0,1,0,1
540 ,1,1,0,1,0,1,0,1,1,1
541 ,1,0,0,0,1,1,0,1,0,0
542 ,0,0,0,1,0,1,1,1,1,1
543 ,0,0,0,1,1,1,0,0,1,1
544 ,1,1,1,1,0,0,0,0,0,1
545 ,0,1,1,0,1,0,0,0,0,0
546 ,1,0,1,1,1,1,0,1,1,1
547 ,1,1,0,0,0,0,1,0,1,0
548 ,0,0,1,1,1,0,1,1,1,0
549 ,1,0,1,1,1,1,1,1,0,1
550 ,0,0,1,0,0,1,0,1,0,1
551 ,1,0,0,0,0,1,0,0,0,1
552 ,1,1,1,0,1,0,0,1,0,0
553 ,0,0,0,1,1,1,1,1,1,1
562 ,0,0,0,1,0,1,0,1,0,1
563 ,0,0,1,0,0,0,1,1,0,1
564 ,0,0,1,1,0,1,1,1,0,1
565 ,0,1,1,1,0,1,1,1,1,1
566 ,1,0,1,1,1,1,1,1,1,1
567 ,1,1,1,1,0,1,0,1,1,1
568 ,1,1,0,1,1,1,1,1,0,0
569 ,1,0,1,0,0,1,1,1,0,0
570 ,1,0,0,1,1,0,0,0,0,0
571 ,0,0,1,1,0,0,0,1,0,1
572 ,0,0,1,0,0,0,1,1,1,1
573 ,1,0,1,1,1,1,1,0,0,1
574 ,1,0,0,1,0,1,1,1,0,1
575 ,1,1,1,0,1,0,0,1,1,0
576 ,1,1,0,1,1,1,0,0,1,1
577 ,1,0,0,0,1,0,0,0,0,1
578 ,1,1,0,1,1,0,0,0,1,0
579 ,0,0,1,0,0,1,1,1,1,0
580 ,1,0,1,1,1,1,0,1,0,1
581 ,1,0,0,0,1,1,1,0,1,1
582 ,0,0,1,1,1,0,0,0,1,1
583 ,0,1,0,0,0,0,1,1,1,0
584 ,0,0,0,0,1,1,0,1,1,1
585 ,1,0,1,1,1,0,1,1,1,1
586 ,0,0,1,1,0,0,1,1,1,1
587 ,1,0,1,0,1,1,1,0,0,1
588 ,1,1,1,0,1,1,0,0,1,1
589 ,1,1,0,0,1,0,0,0,1,0
590 ,0,1,0,1,0,0,1,0,0,1
591 ,1,0,0,1,0,1,1,0,1,1
592 ,0,0,0,0,0,1,1,1,0,1
593 ,1,1,1,0,1,1,1,1,0,1
594 ,1,1,0,0,0,1,0,0,1,1
595 ,0,1,0,1,0,1,1,0,0,1
596 ,0,1,1,0,0,1,0,0,0,0
683 ,1,1,0,0,1,1,1,0,0,1
684 ,1,1,0,0,0,1,0,1,1,0
685 ,0,0,1,1,1,1,1,1,1,1
686 ,1,0,0,1,1,1,1,1,1,1
687 ,0,0,0,0,1,1,1,1,1,1
688 ,1,0,0,0,0,1,1,0,1,1
689 ,1,0,0,1,1,1,1,0,0,1
690 ,0,0,1,1,1,1,0,0,1,1
691 ,0,0,1,0,0,1,1,1,1,1
692 ,1,1,0,1,1,0,1,0,0,1
605 ,1,1,1,0,0,0,1,1,1,1
606 ,0,0,0,1,0,1,0,1,0,1
607 ,0,1,0,0,1,1,1,0,1,1
608 ,0,0,0,1,1,0,1,1,0,1
609 ,1,0,0,0,0,0,1,1,1,1
610 ,1,1,1,0,1,0,1,1,1,0
611 ,0,1,0,0,0,1,0,0,1,1
612 ,1,1,1,0,0,0,1,0,1,0
613 ,0,0,1,0,0,0,0,0,1,1
614 ,1,0,1,0,1,0,1,1,0,1
615 ,1,0,0,0,0,1,1,1,1,1
616 ,0,1,0,1,1,1,1,1,1,1
617 ,1,1,0,0,0,0,0,0,1,1
618 ,1,0,1,1,1,1,1,1,0,1
619 ,0,0,1,1,1,1,0,0,1,1
620 ,1,1,1,0,0,0,0,0,0,1
621 ,1,0,0,1,1,1,0,1,0,1
622 ,1,1,1,0,1,1,1,1,1,1
623 ,1,1,1,0,0,0,0,0,0,0
624 ,1,0,1,0,1,0,1,1,0,1
625 ,1,1,1,0,0,1,1,0,1,1
626 ,1,0,0,1,0,1,0,0,0,1
627 ,1,1,0,1,1,1,1,0,1,1
628 ,1,0,0,0,0,1,1,1,0,1
629 ,0,0,0,1,1,0,1,1,0,1
630 ,1,0,1,1,0,0,1,0,1,1
631 ,1,0,0,0,0,0,0,0,1,0
632 ,1,0,0,0,1,0,1,0,0,1
633 ,0,0,1,1,1,1,1,0,1,1
634 ,1,1,0,1,1,0,0,1,1,1
635 ,1,0,1,1,1,1,1,0,1,1
636 ,1,1,1,1,1,1,1,0,0,0
637 ,0,1,0,1,0,0,0,0,1,0
638 ,1,0,1,1,1,0,0,0,1,1
639 ,0,1,1,0,1,0,1,1,1,1
640 ,1,1,0,1,0,1,0,1,0,1
641 ,1,1,1,0,0,0,1,0,0,0
642 ,1,1,0,1,1,1,1,0,0,0
643 ,1,0,0,0,1,1,0,1,1,1
644 ,1,1,0,0,0,1,0,1,0,1
645 ,0,0,0,1,0,1,0,1,1,1
646 ,1,0,0,1,1,1,0,1,1,1
647 ,1,1,0,0,0,0,1,1,0,0
648 ,0,0,0,1,1,1,1,0,1,1
649 ,1,0,0,1,0,1,0,1,1,0
726 ,0,1,1,1,0,1,1,1,0,1
727 ,1,0,1,0,0,0,1,0,1,1
728 ,1,0,1,0,1,1,1,1,1,0
729 ,1,0,0,0,1,1,1,1,1,1
730 ,0,0,0,0,0,1,1,1,0,1
731 ,1,0,1,0,0,1,0,1,1,0
732 ,1,0,1,1,0,1,1,1,1,0
733 ,1,1,0,0,1,0,1,0,0,0
734 ,1,1,0,1,0,0,1,0,1,1
735 ,1,1,0,1,1,1,1,1,0,1

650 ,0,0,1,0,0,1,1,1,0,1
651 ,0,1,0,1,0,1,1,1,0,1
652 ,1,1,0,0,0,0,1,0,0,0
653 ,1,0,0,1,1,1,0,0,0,1
654 ,0,0,0,1,0,1,0,1,0,1
655 ,1,0,0,0,0,0,0,0,0,0
656 ,0,0,0,1,0,1,1,0,1,1
657 ,1,1,0,0,0,1,0,1,0,1
658 ,1,1,1,1,0,0,1,0,0,1
659 ,1,0,0,1,1,1,0,0,1,1
660 ,1,1,0,1,1,0,0,0,0,0
661 ,0,1,1,1,0,0,0,0,0,1
662 ,1,1,1,0,0,0,0,0,1,1
663 ,1,0,1,0,0,0,1,1,1,0
664 ,1,0,0,1,1,0,0,1,1,1
665 ,1,0,1,1,0,1,1,1,1,1
666 ,1,1,0,0,1,1,1,1,0,1
667 ,0,0,1,1,1,1,1,1,0,1
668 ,1,0,0,0,0,1,0,1,1,1
669 ,1,0,0,0,0,0,1,1,0,0
670 ,1,1,1,0,0,0,0,0,1,0
671 ,0,0,1,0,1,1,1,1,1,1
672 ,1,0,1,0,1,1,1,1,0,1
673 ,0,0,1,0,1,1,1,1,0,1
674 ,0,1,0,0,1,0,1,0,0,1
675 ,0,0,0,1,1,1,1,1,0,1
676 ,1,1,1,0,1,1,1,0,0,1
677 ,1,0,0,1,1,1,1,1,1,0
678 ,0,0,1,1,0,1,0,0,1,1
679 ,1,0,0,1,0,1,1,1,1,1
680 ,0,0,0,1,1,1,0,1,0,1
681 ,0,1,0,1,1,1,0,1,1,0
682 ,1,1,1,0,0,0,0,0,0,1
683 ,1,1,0,0,1,0,0,0,0,1
684 ,1,1,1,0,0,0,0,1,0,0
685 ,1,1,1,0,0,1,0,1,0,1
686 ,0,0,1,0,1,1,0,1,0,1
687 ,0,0,0,1,1,1,0,1,1,1
688 ,1,0,0,1,1,1,1,1,0,1
689 ,1,0,0,1,1,1,1,1,1,1
690 ,1,1,0,0,1,0,1,0,0,1
691 ,1,1,0,0,0,1,1,1,0,1
692 ,0,0,0,0,0,1,1,1,0,1
693 ,1,1,0,1,1,1,1,1,0,0
694 ,1,1,0,0,1,0,1,0,0,1
695 ,1,1,0,0,0,1,1,1,0,1
696 ,0,0,0,0,0,1,1,1,0,1
697 ,1,1,0,1,1,1,1,1,1,1
698 ,0,1,1,1,0,1,1,1,1,1
699 ,1,0,0,1,0,0,1,1,0,1
700 ,1,1,0,1,1,0,1,0,0,0
701 ,0,1,1,1,0,1,1,1,0,1
702 ,1,1,1,0,0,1,1,0,1,1
703 ,1,1,1,1,0,1,0,0,1,1
704 ,0,1,1,0,1,1,0,0,0,0
705 ,1,1,1,0,0,1,1,1,1,0
706 ,1,1,1,1,0,0,1,0,0,0
707 ,0,0,1,1,1,1,0,1,0,1
708 ,1,0,1,1,0,1,1,1,1,1
709 ,1,0,1,1,0,1,1,1,0,1
710 ,1,1,0,1,0,0,0,0,1,0
711 ,0,0,0,1,0,1,1,1,0,1
712 ,1,0,0,0,0,1,1,1,1,1
713 ,0,0,1,1,1,1,1,0,1,1
714 ,1,0,1,1,1,1,1,1,1,1
715 ,1,1,0,0,1,1,0,0,1,1
716 ,1,1,0,1,0,0,1,0,1,1
717 ,1,0,0,1,0,1,0,0,1,1
718 ,1,0,1,0,1,1,1,0,1,1
719 ,1,1,0,0,0,1,0,1,1,1
720 ,1,1,1,0,0,0,0,0,0,0
721 ,1,0,0,0,1,1,1,1,0,1
722 ,1,1,0,0,1,0,1,1,1,1
723 ,1,1,0,0,0,1,0,0,1,0
724 ,1,0,0,1,1,1,0,1,1,1
725 ,1,1,1,1,1,0,0,0,1,0
726 ,1,1,1,0,1,1,1,0,0,1
727 ,1,1,0,1,0,1,1,1,0,0
728 ,1,1,0,0,0,0,0,0,0,1
729 ,0,1,0,1,1,0,0,1,1,1
730 ,1,0,1,0,1,0,1,0,0,1
731 ,1,0,1,0,1,0,1,0,0,1
732 ,0,0,0,0,0,1,1,0,0,0
733 ,0,0,0,1,1,0,0,0,0,0
734 ,1,0,1,0,1,0,0,0,1,1
735 ,1,0,1,0,1,1,1,1,0,1
736 ,1,1,1,0,1,1,1,0,0,1
737 ,1,1,0,1,0,1,1,1,0,0
738 ,1,1,0,0,0,0,0,0,0,1
739 ,0,1,0,1,1,0,0,1,1,1
740 ,0,0,1,0,0,0,1,1,0,0
741 ,1,0,1,0,0,1,0,1,0,0
742 ,1,0,1,0,1,0,1,0,0,1
743 ,0,0,0,0,1,1,0,0,0,0
744 ,1,0,1,0,1,0,0,0,1,1
745 ,1,0,1,0,1,1,1,1,0,1
746 ,1,1,0,0,1,1,1,1,1,1
747 ,1,1,1,0,1,1,1,1,0,1
748 ,1,0,1,1,0,1,1,0,0,0
749 ,1,1,0,0,1,0,0,0,0,1
750 ,1,0,0,0,0,1,0,1,1,1
751 ,1,1,0,0,0,0,1,1,0,1
752 ,1,1,1,0,0,0,0,1,0,0
753 ,1,0,1,1,1,1,1,0,1,0
754 ,1,1,0,1,0,1,0,1,1,1
755 ,1,1,0,1,1,0,1,0,0,0
756 ,0,1,0,1,1,0,1,0,1,1
757 ,0,1,1,1,0,1,1,1,1,1
758 ,0,1,0,0,0,1,1,1,0,0
759 ,0,0,1,1,0,1,0,1,0,1
760 ,0,0,1,1,0,1,0,1,0,1
761 ,0,0,1,0,1,1,1,1,0,0
762 ,1,1,0,0,0,0,0,0,0,1
763 ,1,1,0,0,1,0,0,0,0,0
764 ,1,0,1,1,0,1,1,1,1,1
765 ,1,0,1,1,0,1,0,0,0,1
766 ,1,0,1,0,0,1,1,0,0,1
767 ,0,0,0,1,1,1,0,1,0,1
768 ,0,0,1,0,0,1,1,1,1,1
769 ,1,1,0,0,1,0,0,0,0,1
770 ,1,1,1,0,0,0,0,1,0,0
771 ,1,1,1,0,0,1,0,1,0,1
772 ,0,0,1,1,1,1,1,1,0,1
773 ,1,0,0,0,1,0,1,0,1,1
774 ,1,1,0,0,1,0,0,1,1,1
775 ,1,1,1,1,0,1,1,0,1,1
776 ,0,0,1,0,1,1,0,1,0,1
777 ,0,0,0,1,1,1,0,1,1,1
778 ,1,1,1,0,0,1,0,0,1,1
779 ,0,0,0,1,1,0,1,1,0,1
780 ,1,0,0,1,1,1,1,1,0,1
781 ,0,1,0,1,1,1,0,1,1,0
782 ,1,1,1,0,0,0,0,0,0,1
783 ,1,1,0,0,1,0,0,0,0,1
784 ,1,1,1,0,0,0,0,1,0,0
785 ,1,1,1,0,0,1,0,1,0,1
786 ,0,0,1,0,1,1,0,1,0,1
787 ,0,0,0,1,1,1,0,1,1,1
788 ,1,1,1,0,0,1,0,0,1,1
789 ,0,0,0,1,1,0,1,1,0,1
790 ,1,0,0,1,1,1,1,1,0,1
791 ,0,1,0,1,1,1,0,1,1,0
792 ,1,1,1,0,0,0,0,0,0,1
793 ,1,1,0,1,1,1,1,1,0,0
794 ,1,1,0,0,1,0,1,0,0,1
795 ,1,1,0,1,1,1,0,1,0,1
796 ,0,0,0,0,0,1,1,1,0,1
797 ,1,1,1,1,1,1,1,1,1,1
798 ,0,1,1,1,0,1,1,1,1,1
799 ,1,0,0,1,1,1,1,1,0,1
800 ,1,0,0,1,1,1,1,1,1,1
801 ,1,1,0,0,1,0,1,0,0,1
802 ,1,1,1,0,0,1,0,1,0,1
803 ,0,0,1,0,1,1,0,1,0,1
804 ,0,0,0,1,1,1,0,1,1,1
805 ,1,1,0,0,1,0,1,0,1,1
806 ,1,1,1,0,0,1,0,0,1,1
807 ,0,0,0,1,1,0,1,1,0,1
808 ,1,0,0,1,1,1,1,1,0,1
809 ,0,1,0,1,1,1,0,1,1,0
810 ,1,1,1,0,0,0,0,0,0,1
811 ,1,1,0,1,1,1,1,1,0,0
812 ,0,0,0,1,1,1,1,1,1,1
813 ,1,1,0,1,1,0,1,1,1,0
814 ,1,0,0,1,1,0,1,0,1,0
815 ,1,1,1,1,0,0,0,1,0,0
816 ,1,1,0,0,1,0,0,0,0,1
817 ,1,1,0,1,1,0,1,1,0,1
818 ,1,0,1,1,1,1,0,0,0,1
819 ,1,1,0,0,1,1,1,0,1,1
820 ,1,1,1,0,0,0,0,1,1,1
821 ,0,0,0,0,0,1,1,1,0,1
822 ,0,0,1,0,1,0,0,1,0,1
823 ,1,0,1,1,0,1,1,1,1,1
824 ,1,0,1,1,0,1,1,1,1,1
825 ,1,1,1,1,1,1,1,1,1,1
826 ,1,1,1,0,1,1,1,1,1,1
827 ,1,1,1,0,1,1,1,1,1,1
828 ,1,1,1,0,1,1,1,1,1,1
829 ,1,1,1,0,1,1,1,1,1,1
830 ,1,1,1,0,1,1,1,1,1,1
831 ,1,1,1,0,1,1,1,1,1,1
832 ,1,1,1,0,1,1,1,1,1,1
833 ,1,1,1,0,1,1,1,1,1,1
834 ,1,1,1,0,1,1,1,1,1,1
835 ,1,1,1,0,1,1,1,1,1,1
836 ,1,1,1,0,1,1,1,1,1,1
837 ,1,1,1,0,1,1,1,1,1,1
838 ,1,1,1,0,1,1,1,1,1,1
839 ,1,1,1,0,1,1,1,1,1,1
840 ,1,1,1,0,1,1,1,1,1,1
841 ,1,1,1,0,1,1,1,1,1,1
842 ,1,1,1,0,1,1,1,1,1,1
843 ,1,1,1,0,1,1,1,1,1,1
844 ,1,1,1,0,1,1,1,1,1,1
845 ,1,1,1,0,1,1,1,1,1,1
846 ,1,1,1,0,1,1,1,1,1,1
847 ,1,1,1,0,1,1,1,1,1,1
848 ,1,1,1,0,1,1,1,1,1,1
849 ,1,1,1,0,1,1,1,1,1,1
850 ,1,1,1,0,1,1,1,1,1,1
851 ,1,1,1,0,1,1,1,1,1,1
852 ,1,1,1,0,1,1,1,1,1,1
853 ,1,1,1,0,1,1,1,1,1,1
854 ,1,1,1,0,1,1,1,1,1,1
855 ,1,1,1,0,1,1,1,1,1,1
856 ,1,1,1,0,1,1,1,1,1,1
857 ,1,1,1,0,1,1,1,1,1,1
858 ,1,1,1,0,1,1,1,1,1,1
859 ,1,1,1,0,1,1,1,1,1,1
860 ,1,1,1,0,1,1,1,1,1,1
861 ,1,1,1,0,1,1,1,1,1,1
862 ,1,1,1,0,1,1,1,1,1,1
863 ,1,1,1,0,1,1,1,1,1,1
864 ,1,1,1,0,1,1,1,1,1,1
865 ,1,1,1,0,1,1,1,1,1,1
866 ,1,1,1,0,1,1,1,1,1,1

781 ,1,0,1,1,1,0,1,1,1,1
782 ,1,1,0,1,0,1,1,0,1,1
783 ,1,0,1,1,1,0,1,1,1,1
784 ,1,0,1,0,1,0,0,0,1,1
785 ,1,1,1,0,1,0,0,0,0,0
786 ,1,0,0,1,0,1,1,0,1,1
787 ,1,1,0,1,0,0,1,1,1,1
788 ,0,0,0,0,1,0,0,0,0,1
789 ,1,0,1,1,1,1,1,1,0,1
790 ,1,1,0,0,1,0,1,1,1,0
791 ,0,1,1,0,0,1,0,1,1,0
792 ,1,0,1,0,0,1,0,1,1,1
793 ,1,1,0,0,1,0,1,0,0,1
794 ,1,0,0,0,1,1,0,0,1,1
795 ,0,0,1,0,0,1,1,1,0,0
796 ,1,1,0,1,1,0,0,1,1,1
797 ,1,1,1,1,1,0,0,0,1,1
798 ,0,0,1,0,1,1,1,1,0,1
799 ,1,0,1,0,0,0,1,1,1,1
800 ,0,1,0,1,1,1,1,1,0,1
801 ,1,1,1,0,1,1,1,1,1,1
802 ,0,0,1,0,1,1,0,0,0,0
803 ,0,0,1,1,0,1,1,0,1,1
804 ,1,0,0,1,0,1,0,1,0,1
805 ,0,0,0,1,1,1,0,1,1,1
806 ,0,0,0,1,1,1,1,1,0,1
807 ,1,1,0,0,0,0,0,1,1,0
808 ,0,0,1,0,1,1,1,1,0,0
809 ,0,1,1,1,0,1,1,1,0,1
810 ,0,0,1,1,0,1,1,0,0,1
811 ,0,0,1,1,1,1,0,0,1,1
812 ,1,1,1,0,1,0,0,0,0,0
813 ,1,0,1,0,0,0,0,1,0,1
814 ,0,0,1,1,0,1,1,1,0,0
815 ,1,0,1,0,1,1,1,1,1,1
816 ,1,1,1,0,1,1,1,1,1,0
817 ,0,1,1,0,1,1,1,1,1,1
818 ,0,0,1,1,1,1,1,1,1,1
819 ,0,0,0,1,1,1,0,0,0,1
820 ,1,0,0,1,1,1,1,1,1,1
821 ,0,1,1,0,1,0,1,1,1,1
822 ,1,1,0,0,0,1,0,1,0,1
823 ,1,1,0,0,1,1,1,1,1,0
824 ,1,1,1,1,0,0,0,1,1,1
825 ,1,0,0,0,0,1,0,0,0,1
826 ,0,0,0,1,0,1,0,1,0,1
827 ,1,1,0,1,0,0,0,0,0,1
828 ,1,0,1,1,1,1,0,0,0,1
829 ,1,1,1,0,0,0,0,1,0,0
830 ,0,0,1,0,0,1,0,1,1,0
831 ,1,1,0,1,0,0,1,0,1,1
832 ,1,0,0,1,0,1,1,1,1,1
833 ,1,0,0,1,1,1,1,0,0,1
834 ,1,0,0,0,0,1,1,1,1,1
835 ,1,1,0,0,1,0,1,1,0,1
836 ,1,0,0,1,0,0,1,1,0,1
837 ,1,0,1,0,1,0,1,1,0,1
838 ,1,0,0,1,0,0,1,1,1,1
839 ,0,0,0,1,1,1,0,0,0,1
840 ,1,0,0,1,1,1,1,1,1,1
841 ,0,1,1,0,1,0,1,1,1,1
842 ,1,1,0,0,0,1,0,1,0,1
843 ,0,0,0,0,1,0,1,1,1,0
844 ,0,0,0,1,1,1,1,1,1,1
845 ,1,1,0,1,0,0,0,0,1,1
846 ,0,1,1,1,0,1,0,1,1,1
847 ,1,0,1,1,1,1,0,0,1,1
848 ,1,0,1,1,1,1,1,1,1,1
849 ,1,0,0,0,1,0,0,0,0,0
850 ,1,1,0,0,0,1,1,1,1,1
851 ,1,0,1,1,1,1,1,1,0,1
852 ,0,0,0,0,1,1,0,0,0,0
853 ,1,0,1,1,0,1,1,0,1,1
854 ,1,1,0,1,0,0,1,0,1,0
941 ,1,0,0,1,0,0,0,1,1,0
942 ,1,1,1,0,1,0,0,0,0,1
943 ,0,0,1,1,1,0,1,1,0,1
944 ,1,1,1,0,1,0,1,0,0,1
945 ,1,1,1,1,1,1,1,1,1,0
946 ,1,1,0,0,0,1,1,0,0,1
947 ,1,1,0,0,0,0,0,0,1,1
948 ,1,1,1,0,0,0,1,1,1,1
949 ,1,0,0,0,0,1,1,1,0,1
950 ,1,1,0,0,0,0,0,0,0,1
951 ,1,1,0,1,0,0,1,0,1,0
952 ,0,1,1,0,0,1,0,0,0,0
953 ,1,1,1,0,1,1,1,1,1,1
954 ,1,1,0,0,1,1,1,0,0,1
967 ,1,0,1,1,1,1,1,1,0,1
968 ,0,0,1,1,1,1,1,1,1,1
969 ,1,0,1,1,0,0,0,0,1,1
970 ,0,0,1,1,1,1,1,1,1,1
971 ,0,0,1,0,1,1,1,1,0,1
972 ,0,0,1,0,0,0,0,1,1,1
973 ,0,0,0,1,0,0,0,1,1,1
974 ,1,1,0,0,0,0,0,1,1,1
975 ,1,0,1,1,0,1,1,0,1,1
976 ,1,1,0,0,0,0,0,1,0,1
977 ,0,0,0,0,1,0,1,1,1,1
978 ,1,0,0,1,1,0,0,0,0,1
979 ,1,0,1,0,1,0,1,0,0,1
980 ,0,0,1,1,0,1,0,1,0,1
981 ,0,0,0,1,1,1,1,1,0,1
982 ,1,0,0,0,1,0,1,1,1,1
983 ,0,0,0,1,0,1,1,1,1,1
984 ,1,0,1,0,1,0,1,1,1,1
985 ,1,0,0,0,1,1,0,1,0,1
986 ,1,0,0,0,1,0,0,0,0,1
987 ,1,0,1,0,1,1,1,0,1,1
988 ,1,1,1,1,1,0,0,1,1,1
989 ,1,1,1,1,1,0,0,1,1,0
990 ,1,1,1,0,1,0,0,0,1,0
991 ,0,0,0,1,0,1,0,0,1,1
992 ,1,1,0,0,0,1,0,1,1,1
993 ,1,0,0,1,0,1,0,1,1,1
994 ,0,0,0,0,0,1,1,1,1,1
995 ,1,1,1,1,1,1,0,1,1,0
996 ,0,0,0,1,0,1,0,1,0,1
997 ,1,0,0,1,1,0,0,1,1,1

912 ,0,0,1,1,1,1,0,0,1,1
913 ,1,0,1,1,0,0,1,1,0,1
914 ,1,1,1,0,0,0,0,0,0,0
915 ,0,0,0,1,1,0,1,1,0,0
916 ,0,1,1,0,0,1,0,0,0,1
917 ,1,0,1,1,1,1,0,1,0,1
918 ,0,0,1,0,0,1,0,1,1,0
919 ,0,1,0,0,1,1,1,1,0,1
920 ,0,1,0,0,0,1,1,1,1,1
921 ,0,1,0,1,1,1,0,1,0,1
922 ,1,1,0,1,1,0,0,0,1,1
923 ,1,1,1,1,0,0,0,0,1,0
924 ,0,0,1,0,0,1,1,1,1,1
925 ,1,1,0,1,0,0,1,0,0,1
926 ,1,0,1,0,0,0,0,1,1,1
927 ,0,0,0,1,1,1,1,1,0,1
928 ,1,1,0,1,1,1,0,0,0,0
929 ,1,1,0,0,1,1,1,0,1,1
930 ,1,0,1,1,0,1,1,1,0,0
931 ,1,0,1,1,0,1,1,0,0,1
932 ,0,1,0,0,1,0,0,1,0,1
933 ,1,0,1,1,1,0,0,0,1,1
934 ,1,1,0,1,1,1,0,0,0,1
935 ,0,0,0,0,1,0,0,1,1,1
936 ,1,1,1,1,0,0,0,1,0,1
937 ,1,1,1,1,0,1,0,1,1,1
938 ,1,1,0,1,1,1,1,0,0,0
939 ,1,1,0,1,0,0,0,1,1,1
940 ,1,1,0,0,1,1,0,1,1,1
955 ,1,0,1,0,1,1,1,1,0,1
956 ,1,1,0,0,1,1,1,0,1,1
957 ,0,1,0,1,1,0,0,0,1,0
958 ,0,0,1,0,0,1,0,0,1,1
959 ,0,0,1,0,1,0,0,1,1,1
960 ,1,1,0,0,1,0,0,1,0,0
961 ,1,1,0,1,0,0,0,1,1,1
962 ,1,0,0,1,1,0,0,1,0,1
963 ,0,1,0,0,0,0,1,0,1,1
964 ,0,0,1,0,0,1,1,1,0,1
965 ,0,0,0,0,1,1,1,1,0,1
966 ,1,1,1,0,0,0,0,0,0,0
967 ,1,1,1,1,1,1,1,0,1,1
968 ,1,1,0,0,1,0,1,0,1,1
969 ,1,1,0,0,0,0,1,0,1,1
970 ,1,1,1,1,1,1,0,0,1,1
971 ,1,1,1,1,1,0,1,0,0,1
972 ,1,0,0,0,1,0,1,0,1,1
973 ,0,0,0,1,0,1,1,1,0,1
974 ,0,1,0,0,1,0,1,1,1,1
975 ,1,1,0,0,1,0,1,0,1,1
976 ,1,1,0,0,0,0,1,1,0,1
977 ,1,0,0,0,1,0,0,0,0,0
978 ,0,0,0,0,0,1,1,1,1,0
979 ,0,1,1,0,0,0,1,0,0,1
980 ,1,0,0,1,1,1,0,1,1,1
981 ,1,0,1,1,1,0,1,0,0,1
982 ,1,0,0,1,1,1,0,0,1,1
983 ,0,0,0,1,0,1,1,1,0,1
998 ,0,0,1,1,0,1,0,0,0,1
999 ,1,1,1,0,1,1,1,0,1,1
1000 ,0,0,0,1,1,1,0,1,0,1

## APPENDIX C

## PILOT STUDY

The main purpose was to estimate the reliability of the WINGEN 3 software used for generating data. For this study, ten dichotomously scored items were generated randomly using the computer software. Item response data for 20 examinees in both the reference and focal groups were generated. The ability distribution used for the pilot study was mean 0 standard deviation 1. The data was not replicated prior to analysis. The Kuder-Richardson 20 method was used to estimate reliability.

The formula for KR-20 for a test with K test items numbered i=1 to K is;

$$r = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} p_i q_i}{{\delta_X}^2}\right]$$

Where $p_i$ is the proportion of correct responses to test item i, $q_i$ is the proportion of incorrect responses to test item i so that $p_i + q_i = 1$, and the variance for the denominator is

$${\delta_X}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_i)^2}{n}$$

Where n is the total sample size.

If it is important to use unbiased operators then the sum of squares should be divided by degrees of freedom (n-1) and the probabilities are multiplied by

$$\frac{n}{n-1}$$

The number of items was taken as 10 and the sample size of 20 for both the focal and reference groups. The data was taken for normal Ability distribution with mean 0 and standard deviation 1

From the item response K= 10 , $\sum_{i=1}^{10} p_i q_i = 4.3$, $\frac{\sum_{i=1}^{n}(X_i - \bar{X}_i)^2}{20} = 13.23$

$$r = \frac{10}{10-1}\left[1 - \frac{4.3}{13.25}\right]$$

$$r = 0.75$$

The reliability estimate for the item response data was therefore found to be 0.75. It was therefore assumed that the tests were internally consistent and stable. The software was therefore suitable for use in the main study.

The item response data generated for the pilot study for Test length 10, Mean, SD = 0, 1 and Sample size 20 was as follows

| **Size No** | **Scores on each item** | **Total score** |
|---|---|---|
| 1 | 1,1,1,0,0,1,0,0,0,1, | 5 |
| 2 | 0,1,1,0,1,1,0,0,1,1, | 6 |
| 3 | 1,1,0,0,1,0,0,1,0,1, | 5 |
| 4 | 1,0,1,1,1,0,1,1,1,1, | 8 |
| 5 | 1,1,0,1,1,0,1,1,0,1, | 7 |
| 6 | 0,1,1,0,1,1,0,0,0,1, | 5 |
| 7 | 1,1,1,0,1,0,0,1,1,1, | 7 |
| 8 | 1,1,1,1,0,1,0,0,0,0, | 5 |
| 9 | 0,1,0,0,1,1,1,0,1,0, | 5 |
| 10 | 0,1,1,0,1,1,0,0,1,0, | 5 |
| 11 | 0,1,1,0,0,0,0,0,1,1, | 4 |
| 12 | 0,1,1,0,0,0,0,1,1,1, | 5 |
| 13 | 0,1,1,1,0,1,0,0,0,0, | 4 |
| 14 | 1,1,1,0,0,1,0,0,0,0, | 4 |
| 15 | 0,1,0,0,1,0,0,0,1,1, | 4 |
| 16 | 1,1,0,0,0,0,0,0,0,0, | 2 |
| 17 | 0,0,1,0,0,0,0,0,0,1, | 2 |
| 18 | 0,1,0,0,1,1,1,0,0,0, | 4 |
| 19 | 1,0,0,1,1,0,1,1,0,0, | 5 |
| 20 | 1,1,0,1,1,0,1,0,1,1, | 7 |

Source WINGEN 3 Software.

**APPENDIX D:**

ETHICS APPROVAL


**MASENO UNIVERSITY ETHICS REVIEW COMMITTEE**

Tel: +254 057 351 622 Ext: 3050
Fax: +254 057 351 221

Private Bag – 40105, Maseno, Kenya
Email: muerc-secretariate@maseno.ac.ke


**FROM**: Secretary - MUERC

**DATE**: 25th September, 2017

**TO**: Ferdinand Ingubu Ukanda
PG/PHD/093/2010
Department of Educational Psychology
School of Education
P. O. Box, Private Bag, Maseno, Kenya

**REF**:MSU/DRPI/MUERC/00421/17

**RE: Effectiveness of Mantel-Haenszel and Logistic Regression Statistics in Detecting Differential Item Functioning Under Different Conditions. Proposal Reference Number MSU/DRPI/MUERC/00421/17**

This is to inform you that the Maseno University Ethics Review Committee (MUERC) determined that the ethics issues raised at the initial review were adequately addressed in the revised proposal. Consequently, the study is granted approval for implementation effective this 25th day of September, 2017 for a period of one (1) year.

Please note that authorization to conduct this study will automatically expire on 24th September, 2018. If you plan to continue with the study beyond this date, please submit an application for continuation approval to the MUERC Secretariat by 15th August, 2018.

Approval for continuation of the study will be subject to successful submission of an annual progress report that is to reach the MUERC Secretariat by 15th August, 2018.

Please note that any unanticipated problems resulting from the conduct of this study must be reported to MUERC. You are required to submit any proposed changes to this study to MUERC for review and approval prior to initiation. Please advice MUERC when the study is completed or discontinued.

Thank you.

DIRECTORATE OF RESEARCH PUBLICATION & CONSULTANCIES 25 SEP 2017 MASENO UNIVERSITY

Dr. Bonuke Anyona,
Secretary,
Maseno University Ethics Review Committee

Cc: Chairman,
Maseno University Ethics Review Committee.