Effect of Unequal Sample Sizes on the Power of DIF Detection: An IRT-Based Monte Carlo Study with SIBTEST and Mantel-Haenszel Procedures

Risper Akelo Awuor

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Educational Research and Evaluation

Dissertation Committee:
Gary E. Skaggs (Chair)
Edward W. Wolfe
Kirby Deater-Deckard
Yasuo Miyazaki

June 19, 2008
Blacksburg, Virginia

Keywords: Monte Carlo simulation, statistical power, differential item functioning, DIF magnitude, DIF detection, nominal p-value, sample size, combination ratios.

Effect of Unequal Sample Sizes on the Power of DIF Detection: An IRT-Based Monte Carlo Study with SIBTEST and Mantel-Haenszel Procedures

Risper Akelo Awuor

(ABSTRACT)

This simulation study focused on determining the effect of unequal sample sizes on statistical power of SIBTEST and Mantel-Haenszel procedures for detection of DIF of moderate and large magnitudes. Item parameters were estimated by, and generated with the 2PLM using WinGen2 (Han, 2006). MULTISIM was used to simulate ability estimates and to generate response data that were analyzed by SIBTEST. The SIBTEST procedure with regression correction was used to calculate the DIF statistics, namely the DIF effect size and the statistical significance of the bias. The older SIBTEST was used to calculate the DIF statistics for the M-H procedure. SAS provided the environment in which the ability parameters were simulated; response data generated and DIF analyses conducted. Test items were observed to determine if a priori manipulated items demonstrated DIF. The study results indicated that with unequal samples in any ratio, M-H had better Type I error rate control than SIBTEST. The results also indicated that not only the ratios, but also the sample size and the magnitude of DIF influenced the behavior of SIBTEST and M-H with regard to their error rate behavior. With small samples and moderate DIF magnitude, Type II errors were committed by both M-H and SIBTEST when the reference to focal group sample size ratio was 1:.10 due to low observed statistical power and inflated Type I error rates.

DEDICATION

I would like to dedicate this dissertation to my late father, Boaz Onyino Andhoga; my mother Rodah Andeso Onyino; my brother Richard Alang'o Onyino and my brother in-law Albert Ayieko Agwanda for the sacrifices they made to lay foundation for my graduate education.

ACKNOWLEDGEMENTS

First I am grateful to the Almighty God for His favors and for the good health I have enjoyed throughout the doctoral program. Second, I am indebted to the mentoring and continued assistance from my dissertation committee chair, Professor Gary Skaggs. As my academic advisor, Professor Skaggs did not only introduce me to educational measurement research, but he also literally guided my steps in the area of differential item functioning analysis and academic standards setting research. His invaluable support enabled me to complete this dissertation. Much gratitude to all my committee members: Professor Edward W. Wolfe; Professor Yasuo Miyazaki and Professor Kirby Deater-Deckard, for constantly being there for me. Each one of them remained committed to see me through this academic task.

I owe special thanks to Professor Kusum Singh for admitting me to the Educational Research and Evaluation program and for all the support along the way.

Next, I am grateful to my family, first to my husband, Mordoch for his perseverance and his 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: Cherian and Millie; Steve and Kim; Mercy, Victor, Tonya and Madison. Thank you all for your love and understanding. You all have inspired me throughout my academic undertakings.

I extend much appreciation to my church family, the members of Melrose Avenue S.D.A church. Your prayers gave me a push each day to “strive towards the mark”. Special thanks to Elder Pat Hatch; Pastor Harry Britt; Jean Mills; Chaplain/Captain Jonathan Runnels; Mary Mark; and JoAnn Johnson, just to mention a few. In the same spirit I thank the Walsh family that welcomed me to be part of their family in Blacksburg. The love and caring spirit of the late Sally Roraback will live to be remembered. I owe much thanks to Jeff Walsh; Sunshine and Geoff Hula, Brian and Regina Walsh; April and Douglas; and the rest of the Walsh family members.

This acknowledgement would be incomplete if no special mention was made of the contributions of Professor Scott F. Midkiff; Professor David Parks; Professor Elizabeth Creamer; Professor Elizabeth Fine; Professor Michael Herndon and Professor Kathleen C. Arceneaux for providing employment opportunities through which I gained professional experience and academic advancement besides the finances that supported me during the doctoral program.

Once again I thank my dissertation committee chair, Professor Gary Skaggs for the research experience I gained through the Aspires Research Grant that he was awarded. It was a privilege to be under your mentorship.

I am grateful to the leadership of Educational Research and Evaluation Program; the Department of Educational Leadership and Policy Studies; the School of Education; and to Virginia Polytechnic Institute and State University for making my dreams come true. Special thanks to Professor David Alexander; Professor Penny Burge; Dr. Daisy Stewart; Kathy Tickle; Pat Bryant and Connie Smith.

I am also indebted to Professor Kerstin Palmerus of Gothenburg University, Sweden, and Professor John Agak, of Maseno University, Kenya, for providing the 'stepping stone' through a joint research project that got me to a doctoral program. I am grateful to Gothenburg University for the invaluable academic, professional, technical and cultural experiences that were extremely useful in my doctoral degree program.

I owe special tribute to Dr. Reginald Michael and his loving wife, Annette for their love and emotional support throughout this academic undertaking. I also thank Dr. Sylvanus Nacheri; his wife Syvil and their children, Lucy, Thomas and Deborah for their friendship. Much thanks to Dr. Yoshi Sigusawa for providing some of the technical support I needed to complete the dissertation.

Last but not least, I am grateful to the graduate students on the Educational Research and Evaluation Program at Virginia Tech. All of them were very inspiring and supportive as we worked together toward a common goal. Thanks to you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

Introduction

The concept of power in statistical theory is defined as the probability of the null hypothesis given that the null hypothesis is false. Sample size is known to be positively related to the statistical power of differential item functioning (DIF) procedures (Narayanan & Swaminathan, 1996). Empirical studies based on the mean and covariance structure analysis (MACS) model applied to DIF on graded response items (e.g., Kaplan & George, 1995) using the mean of factor loadings as the estimate; and simulation study results (e.g, Chan, 2000; González-Romá, Tomas, Ferreres, & Hernandez, 2005; González-Romá, Hernández, & Gómez-Benito, 2006; Wasti, Bergman, Glomb, & Drasgow, 2000) using the maximum likelihood (ML) estimation procedure show that unequal sample sizes across groups decrease the statistical power of DIF detection indices.

Kaplan and George's (1995) DIF study that examined the power of Wald test within the multi-group confirmatory factor analysis (CFA) method using an approach similar to that of Hotelling's T^2 found that power was most affected by the degree of factor true mean differences. With small inequalities in sample sizes large changes in the power of the test were observed even under conditions of factorial invariance. González-Romá et al., (2006) observed that power level was inadequate to detect DIF of medium magnitude when the focal group had as small a sample size as100 but when the focal group size was 200, there was acceptable power level to accurately detect DIF of medium magnitude when they employed the ML DIF estimation method.

In the Chan (2000) and González-Romá et al., (2005) studies the largest modification index (MI) associated to the factor loading estimates was evaluated to determine statistical significance of the estimated DIF. A MI shows the reduction in the model chi-square value, if the implied constrained parameter is freely estimated because the chi-square difference is distributed with one degree of freedom. González-Romá et al., (2006) suggested that the inadequacy of statistical power for DIF detection when the focal group's sample size is very small may be due to poorly estimated item parameters, especially if the parameters were not first constrained to be equal across groups.

While Kim, Cohen and Kim (1994) suggest that a minimum of 800 examinees is required for the 3PLM to generate data sets that can provide accurate item parameter estimates with reduced error rates in DIF detection, González-Romá et al., (2006) reported that a minimum of focal group sample size as small as 200 resulted in adequate detection of

DIF of medium magnitude when they used MACS model. In Roussos and Stout's (1996) simulation study in which they examined the effect of small sample sizes and studied item parameters on Simultaneous Item Bias Test (SIBTEST) and Mantel-Haenszel (M-H) with equal sample sizes in the focal as in the reference groups, they reported that the amount of statistical bias due to item parameter variations remained approximately constant with sample size, while the associated false rejection rates decreased with decreasing sample size due to reduced statistical power to detect bias.

González-Romá et al., (2006) conducted a simulation study to investigate power and Type I error rate of a procedure based on the MACS model to detect uniform and non-uniform DIF when the percentage of DIF item is large and sample sizes are equal and unequal. The results of their study indicate that statistical power increases as sample size increases, and the procedure showed acceptable power levels (≥70%) for detecting uniform and non-uniform DIF of medium magnitude (.25) when 0.10 was regarded as low, 0 .25 medium and 0.50 as large with as low as 200/200, and 400/200 sample sizes in the reference and focal groups respectively.

From the results of the previous empirical and simulation studies on the power of procedures for detecting DIF it is not clear how large or how small the difference in sample sizes for the reference and focal groups should be to ensure the appropriate statistical power for DIF detection. This simulation study is designed to address this problem with the intention of extending the research on the power of procedures for detection of DIF and to provide information that may be used by researchers who use empirical data to provide score validity evidence for all examinees across groups.

Purpose of Study

The purpose of this simulation study was to determine the effect of unequal sample sizes in the reference and focal groups on the statistical power of DIF detection using an IRT-based Monte Carlo procedure and MULTISIM for data simulation, and SIBTEST and M-H for DIF detection. The study also aimed at examining the influence of different DIF magnitudes on the performance of the DIF detection procedures. The two procedures were compared on the basis of their statistical power to detect DIF items and on their Type I error rate control.

Research Questions

This study was aimed at addressing two major questions:

1. What is the minimum sample size in the focal group that can provide adequate statistical power for accurate DIF detection?
2. What is the effect of unequal sample sizes on SIBTEST and M-H error rate under uniform DIF of moderate and large magnitudes?

Rationale For and Justification of the Study

In order to enhance equity and improve assessment, it is important that the assessment instruments be unbiased. With increased immigration and settlement of people from diverse ethnic backgrounds, psychometricians have the responsibility of ensuring that accurate assessment occurs for all examinees. Thus, for example, items intended to measure reading proficiency must be valid for use with students from diverse groups (e.g., ethnicity, gender, special education status) for meaningful score interpretation (Finch & French, 2007).

It is evident that DIF continues to receive attention both in applied and methodological studies (Finch and French, 2007). Because DIF can be an indicator of construct irrelevant variance that can influence test scores, continuing to evaluate and improve the accuracy of detection methods is an essential step in gathering score validity evidence. Besides, additional information on DIF detection is necessary because, as Finch and French have stated, highly discriminating items are particularly vulnerable to false flagging for DIF: a situation that leads to disproportionate removal of most informative items from the test. For the organizations that create assessment instruments, false flagging of items is not a simple matter because test preparation is an expensive and time consuming enterprise. Therefore discarding perfectly good items from a test because of inaccurate functioning of a DIF detection procedure has serious educational and economic implications. Finch and French also expressed their suspicion that the low power for DIF detection that is usually observed for most DIF detection procedures is more likely than not a function of contaminated subtest items. This Monte Carlo study was designed to control for the subtest item contamination while determining the statistical power of the procedure for DIF detection.

Several Monte Carlo DIF detection studies have focused on the influence of sample size on DIF detection to determine sample size that results in minimal variance and least error rates with varied DIF detection procedures (e.g., González-Romá et al., 2006; Kim, et al., 1994; Roussos & Stout, 1996b). However, studies that are focused on the effect of unequal sample sizes that mimic the reality that exists in empirical data, especially testing accommodation data, are still limited. This study was designed and undertaken to add to the

limited existing literature on the effect of unequal sample sizes on the statistical power of DIF detection using SIBTEST and M-H procedures.

Overview of the Methodology

The purpose of this study was to determine the effect of unequal sample sizes across groups on the statistical power for DIF detection using an IRT-based Monte Carlo procedure and MULTISIM for data simulation and SIBTEST and M-H for analysis of the generated data. The study also aimed at examining the influence of different DIF magnitudes on the performance of the DIF detection procedures. The independent variables in this study were sample size and DIF magnitude. The dependent variables were the percent of DIF detected and power corresponding to a significance level of .05 for the alternative hypothesis of DIF against the focal group. IRT-based Monte Carlo procedure was used to generate dichotomously scored response data using MULTISIM.

Although the assessment of educational outcomes have in many instances taken the form of open-ended, constructed-response instruments or instruments that combine objective items with performance tasks (Ankenmann, Witt, & Dunbar, 1999), this study used the dichotomous items, first, because they were still widely used; second, because there was sound evidence that multiple choice tests have defined the standards of assessment in the past (Zwick, Donoghue, & Grima, 1993), yet even with the dichotomous items, the problem of item and test bias have not been resolved.

Chapter Two of this study is the review of empirical and simulation studies on the effect of sample size and DIF magnitude on the statistical power of DIF detection and error rate behavior for SIBTEST and M-H. Chapter Three is the presentation of methodology of parameter estimation, data simulation and the models used in the generation of item and ability parameter estimates. The chapter also contains procedures for simulating data and for data analysis. In addition, the chapter provides justification for the choices of procedures and models used for simulation and analysis of the data.

Manuscripts Developed from the Study

Two articles were developed from this study to meet the minimum requirement of the manuscript option to an alternative to the traditional dissertation. Brief descriptions of the contents of the articles are provided in the next section.

Manuscript 1

The first manuscript was a Monte Carlo simulation study designed to determine the effects of unequal sample sizes on the statistical power of the SIBTEST with regression correction procedure to detect DIF of different magnitudes. Many studies have relied on large differences in the total mean scores across groups to indicate the presence of bias. However, the use of mean differences as indication of bias has been challenged that a group mean difference is not sufficient evidence of bias because it may reflect some valid group differences. SIBTEST DIF detection methodology does not rely on group mean differences based on observed scores. SIBTEST, being a latent trait model controls for ability while detecting items that exaggerate the ability difference across groups of examinees. The aim of the study was to determine how small a sample in the focal group would be required to ensure adequate statistical power for the SIBTEST procedure to detect uniform DIF of medium and large magnitudes.

Manuscript 2

The second manuscript was a comparison of the performance of SIBTEST with that of M-H in terms of their error rates in detecting uniform DIF of different magnitudes with unequal sample sizes. The aim of the study was to determine if the effect of unequal sample sizes was dependent on the procedure for DIF detection, considering that the concept of DIF was perceived differently for the two DIF detection procedures. With SIBTEST, if the null hypothesis is rejected, members of the reference and focal groups with the same underlying trait are considered to differ in their probability of a correct response to that item on the test. Hence, the item is identified as exhibiting DIF. On the other hand, with M-H an item is considered as exhibiting DIF when the members of the reference group that are identical in overall ability differ in their mean performance on the studied item.

Definition of Terms

The following terms, as used in this study were defined as follows:

- Monte Carlo Method – A technique that involves using random numbers and probability to generate data.
- Differential item functioning (DIF*)* – A manifestation of bias observed when examinees from different groups have different probability or likelihood of answering an item correctly, after controlling for ability.

- Uniform DIF – The kind of DIF that exists when the statistical relationship between item response and group is constant for all levels of ability.
- Empirical data – Is taken to mean the data collected through observation and are used to derive some conclusions.
- 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.
- Test dimension – Test item characteristic that can affect the probability of correct response.
- Wald test – A statistical test typically used to test whether an independent variable has a statistically significant effect on the dependent variable.
- Testing accommodation – A change in how a test is administered to provide equal opportunity to all test takers to demonstrate their knowledge, without substantially altering what the test is intended to measure (Tindal & Fuchs, 1999).
- 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 provides evidence that supports that the interpretations of the scores are correct and that the manner in which the interpretations are used is appropriate (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).
- Hotelling's T^2 – a statistic for a multivariate test differences between the mean values of two groups.

CHAPTER TWO

Literature Review

This chapter is a review of simulation and empirical studies on DIF detection procedures. The chapter contains literature on the effect of sample sizes on DIF detection and related data simulation models. Most of the reviewed pieces of literature in this section were on the statistical power of SIBTEST and M-H procedures for DIF detection.

A variety of procedures for detecting possible item bias through DIF have been developed for dichotomous items. Potenza and Dorans (1995) classified the most widely used procedures for DIF detection in two major categories: (a) the way in which the matching variable is obtained (observed score versus an estimate of the latent variable presumed to underlie test performance) and (b) whether an assumption is made about the form of relationship between an item score and the matching variable (parametric if a particular form for the item response function is assumed versus nonparametric if such assumption is not made). Under Potenza and Dorans' classification scheme the M-H (1959) and standardization (Dorans & Holland, 1993) procedures are considered observed-score/nonparametric method. SIBTEST procedure (Shealy & Stout, 1993) fits in the latent-trait/nonparametric category. Procedures based on item response theory (IRT; Thissen, Steinberg, and Weiner, 1993) are latent-trait/parametric methods. On the basis of this classification, M-H procedure is considered to be nonparametric because no particular form for item response function is assumed (Akenmann, et al., 1999).

DIF Detection

In the process of establishing the validity of specific score inferences, Tan and Gierl, (2005) noted that one important step is to assess the fairness of the test through the analysis of DIF. DIF analysis is a procedure used to determine if test items are fair and appropriate for assessing the knowledge in a specific subject area across similar groups of examinees. The analysis is based on the assumption that test takers who have similar knowledge should perform in similar ways on individual test items regardless of their sex, race, or ethnicity (Tan & Gierl, 2005). Tan and Gierl are in agreement with previous researchers that DIF analysis should usually be conducted to detect group differences by assessing the probability that individuals of equal ability answer an item correctly. The presence of DIF, therefore, signals that factors related to group membership affect the probability of correct response, thus threaten fair assessment (Pae, 2004). Consequently, procedures have been developed for

DIF detection in an attempt to improve the equity among examinees with same ability on the latent trait that the test item is intended to measure.

As it has been stated in the previous section, there are two major kinds of DIF detection approaches: a parametric approach, which assumes a specific item response model, and a non-parametric approach, which does not assume a specific item response model. Finch (2006) observed that for detecting DIF in dichotomous items with the nonparametric approach, the M-H procedure (Holland & Thayer, 1988; Mantel & Haenzel, 1959) and logistic regression procedure (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990) were the two most popular ones.

Swaminathan and Rogers (1990) and Rogers and Swaminathan (1993) have classified two kinds of DIF: uniform and non uniform DIF. In their classification they suggest that uniform DIF exist 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, 2006).

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 M-H method began to lose control over the Type I error (Fidalgo, Mellenberg, & Muñiz, 2000; Miller & Oshima, 1992; Narayana & Swaminathan, 1994, 1996; Rogers & Swaminathan, 1993; Uttaro & Millsap, 1994). However, more recent study results 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 (Wang, 2004; Wang & Su, 2004;Wang & Yeh, 2003). These more recent simulation studies confirm Raju's suggestion (Raju, 1988) that it is the difference in abilities between the reference and the focal group that affects the Type I error rather than the percentage of DIF items. The reports of the studies have also shown that as long as ability differences between the reference and focal groups approach zero, the M-H and IRT-based DIF detection method maintain control of their Type I error even when there are 50% DIF items in the test. According to Wang, the difference in ability that exceeds approximately 0.04 is an indication of a wide range between the reference and the focal groups.

Many studies have relied on large differences in mean scores across groups to indicate the presence of bias (e.g., Rosser, 1989). Angoff, (1993); McAllister, (1993); and Camilli and

Shepard (1994) argue that a group mean difference is not sufficient evidence of bias because it may reflect some valid group differences. Camilli and Shepard (1994) state 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 M-H suggests that the item is more likely to be measuring additional constructs that function differently from one group to another (Angoff, 1993; Camilli & Shepard, 1994; Roussos & Stout, 1996; Shealy & Stout, 1993).

Gierl et al., (1993) 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. This study used simulation method that allowed the manipulation of sample sizes in the focal and reference groups so as to determine the effect of unequal sample sizes on the power of a procedure to detect DIF of moderate and large magnitudes.

In another DIF study, Bennett, Rock and Novatkoski (1989) attempted to identify categories of test items that demonstrated DIF in math Scholastic Aptitude Test (SAT) and causes of DIF for those items. As part of the study, Bennett et al., obtained individual items from modified administration and subjected them to DIF analyses. Several suspect items were identified. For each item in a cluster, the M-H statistic was computed (Holland & Thayer, 1988) and converted to difficulty scale (Donlon, 1984). The results of the transformation provided an estimate of DIF on the difficulty scale referred to as Δ_{MH} (Holland & Thayer, 1988). The study further looked at the performance of students who took the test under accommodated administrations to see if particular item characteristics could be associated with the DIF. Some items from the suspected category showed no evidence of differential functioning indicating that there might have been other factors besides the hypothesized ones that contributed to differential behavior of test items across groups (Bennett et al., 1989).

The inability to detect DIF where it is highly suspected or where it is modeled in the study and the mixed results in accommodation studies suggest that there is a gap in the DIF studies that need to be bridged. The effect of unequal sample sizes for the reference and focal

groups on DIF detection has not been studied exhaustively. This study used simulated data to manipulate sample sizes in the reference and focal groups to determine the effect of unequal sample sizes on the statistical power of DIF detection for SIBTEST and M-H procedures.

Monte Carlo Data Simulation

Focal groups in DIF analysis are more often than not very small compared to the reference groups. Monte Carlo technique in which data are simulated has been used in several studies to solve the problem of small sample sizes (Harwell, Stone, Hsu, & Kirisci, 1996). Some researchers regard Monte Carlo studies as statistical sampling experiments with underlying model whose results are used to address research questions that would otherwise be difficult to address (Naylor, Balintfy, Burdick, & Chu, 1968; Spence, 1983) with maximum generalizability and replicability of results. Harwell et al., (1996) consider Monte Carlo studies as mirror images of empirical studies. They however caution that Monte Carlo studies should only be employed if information cannot reasonably be obtained in other ways. An example of appropriate employment of Monte Carlo techniques is in studies that focus on determining test characteristics such as DIF. In cases where focal group sample sizes are very small, Monte Carlo simulation of data would be appropriate as suggested by Harwell et al (1996) to detect manipulated DIF items.

There are many variables that may require manipulation and a researcher may be tempted to include many outcome variables when conducting a Monte Carlo study. With reference to DIF studies, Naylor et al. (1968) cautions that too many outcome measures may decrease efficiency of a study and increase the occurrence of chance differences. Therefore, they suggest that the Monte Carlo study designs should include only few outcome measures to make interpretation of the results fairly accurately manageable.

Gonzalez-Romá, Hernandez and Gomez-Benito (2006) conducted a Monte Carlo 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 (800-800, 400-400, 200-200, 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 and latent trait were equal across groups (Gonzalez-Romá et al., 2006; Hernandez & Gonzalez-Romá, 2003; Narayana & Swaminathan, 1996).

Gierl, Gotzmann, and Boughton (2004) also reported the striking results between the balanced and unbalanced DIF conditions when they manipulated DIF percentages. 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. Although this study did not consider unbalanced DIF, and DIF percentage was not one of the manipulated variables, 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.

Selecting the Number of Replications

According to Harwell, Stone, Hsu, and Kirisci (1996) Monte Carlo simulation is one of the available techniques for reducing the variance of estimated parameters through the replication of data. These authors suggest that the number of replications in IRT-based research is influenced by the purpose of the Monte Carlo study, the desire to minimize the sampling variance of the estimated parameters, and by the need for statistical tests of Monte Carlo results to have adequate power to detect the effects of interest. They further suggest that using the same seed value in the generation of the varying sample sizes reduces noise in

the simulated data and helps minimize the effect of random error on parameter estimates. Other studies have also reported that the number of replications has a direct influence on the precision of the estimated parameters. The studies have also reported that more replications produce parameter estimates with less sampling variance (Harwell, Rubinstein, Hayes, & Olds, 1992). While estimated other parameters may require large numbers of replications, Harwell et al. (1996) propose 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 is debatable because it can be argued that there are many variables that influence the statistical power of a DIF detection procedure. Many studies do not control for most of the variables and therefore may produce results that do not lead to solving the problem of bias against minority groups. Educational assessments depend on total test score in their decision making. It is therefore crucial that studies that inform assessment institutions are as accurate as can possibly be. Millsap and Meredith (1992) recommend that large sample sizes and long tests are more reliable when decisions are based on raw test score.

Several Monte Carlo studies stressed how crucial the choice of the number of replications is in a Monte Carlo study because the results of such studies may vary depending on the number of replications and there is a great danger in using no replications because 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, & Rogers, 1993; Hauck & Anderson, 1984; Stone, 1992, 1993). Unlike Harwell et al., (1996), these authors do not suggest the number of replications that are sufficient for DIF detection.

Naylor et al., (1968) observed that Monte Carlo technique was one way of applying valid IRT-based methods in a study. They noted that most of the then existing IRT studies were devoted to solving statistical rather than measurement problems. Rubinstein (1981) and Ripley (1987) are examples of application of IRT-based simulation studies but they are limited to definition, objectives, and limitations of Monte Carlo techniques (Harwell, et al., 1996).

Type I and Type II Error Rates

The quality of a procedure is assessed in terms of Type I error rate (the proportion of items with DIF falsely identified) and Type II error rate (the proportions of items with DIF not identified). In the DIF detection study, Type II error posses 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. Monte Carlo studies allow researchers to

determine with certainty which techniques are better than others and in what conditions, as the parameters that determine the simulations, such as DIF magnitude, item parameters, ability distribution, item discrimination parameter and sample size are known. However, when empirical data are analyzed, there is no way 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 (Fidlago, 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. They argue that the two stage DIF detection produced better Type I error control in empirical studies (Shealy & Stout, 1993, p. 176). 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 alpha.

The SIBTEST Procedure

The SIBTEST procedure is a nonparametric procedure for detecting DIF. It was developed as an extension of Shealy and Stout's (1993) multidimensional item response theory. Within the framework of SIBTEST, DIF is conceptualized as a difference in the probability of endorsing a keyed item response occurring when individuals in groups having the same levels of the latent attribute of interest possess different amounts of nuisance abilities that influence response.

SIBTEST conducts DIF analyses using original item response data rather than parameter estimates from a program. For example, SIBTEST can be used to identify items that are biased against a particular group or groups of examinees. For a scale containing 'n' items, 'n' single-item DIF statistics must be computed to obtain p-values indicating the significance of the outcomes. The observed *p*-values are then compared to a critical p-value, such as .05 / n, that is adjusted for the number of comparisons made. If an observed p-value is less than the "corrected" critical p-value, then the null hypothesis of no DIF:

$$\boldsymbol{H}_0: \beta_{UNI} = 0 \text{ vs. } H_1: \beta_{UNI} \neq 0,$$

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:

$$\beta_{UNI} = \int B(\Theta) f_F(\Theta) d\Theta,$$

where $d\Theta$ is the differential of theta, $B(\Theta)$ is integrated over Θ to produce $\beta_{UNI,}$ 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. In this study the newer version of SIBTEST (Shealy & Stout, 1993) will 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; Stout, Li, Nandakumar, & Bolt, 1997).

SIBTEST and its Regression Correction

One emphasis in the improvement and evaluation of SIBTEST has been the control of Type I error (false flagging of non-DIF items) inflation and estimation bias. When there is no target ability difference between the reference and focal group, it can be shown that estimated beta is an unbiased estimate of β.

The weighted mean difference between the reference and focal groups on the subtest item across the Θ subgroups is computed with

$$\hat{\beta}_{UNI} = \sum_{k=0}^{k} P_{kdk}$$

which provide an estimate β_{UNI} (Shealy & Stout, 1993).

SIBTEST also yields an overall statistical test for $\hat{\beta}_{UNI}$. The test statistic for evaluating the null hypothesis is

$$SIB = \frac{\beta_{UNI}}{\sigma(\beta_{UNI})}$$

A global index of DIF is defined by Shealy and Stout (1993) for the dichotomous case as

$$\hat{\beta} = \int Bo\ (\theta)\ f_F\ (\theta)\ d,\ \theta)$$

where $f_F\ (\theta)$ denotes the density of θ in the focal group, while the modified SIBTEST for dichotomous items (Chang, Mazzeo, & Roussos, 1996) measures the amount of DIF at θ by

$$Bo\ (\theta) \equiv E_R\ [Y \mid \theta] - E_F[Y \mid \theta].$$

when the target ability distribution for the reference and the focal groups of examinees are different, estimated beta is no longer unbiased. In order to more accurately estimate β, SIBTEST adjusts estimated beta for possible differences in the distribution of $\theta_R \mid X_R = x$ and $\theta_F \mid X_F = x$ by incorporating estimation of the regression of examinees matching subtest true score on observed matching subtest score. Shealy and Stout (1993) refer to this adjustment as the regression correction that can be expressed as:

$$Vg(x) = \sum E\ [P_j\ (\theta_g) \mid X_g = x]$$

where $V_g(x)$ denotes the regression of the matching subtest item; P_j is the item response function (IRF) of the j-th matching subtest item when it is assumed that $V_g(x)$ is linear in x and using the true score T model (the true score model and IRT model are linked by viewing the T as a monotone transformation $[(\theta)$ of $\theta]$ (Jiang & Stout, 1998).

The Mantel-Haenszel Procedure

The M-H procedure (Mantel & Haenszel, 1959) is based on estimating the probability of a member of the reference or the focal group at a certain ability level getting an item correct. It is a DIF detection method that first identifies the reference and focal groups of examinees. The reference group is expected to provide standard performance on the item of interest, and the focal group, whose differential performance, if any, is to be detected and measured. M-H requires that these groups be matched according to relevant stratification using levels of ability. It also requires a 2 x 2 contingency table for each of the levels of ability constructed from the responses to the suspect items by the examinees of each group

(Ryan, 1991). The procedure's estimate, α, of the difference in performance on an item between the reference and focal groups across ability levels are:

$$\alpha_{MH} = \frac{\sum A_j D_j / T_j}{\sum B_j C_j / T_j}$$

where A_j is the observed number of examinees in the reference group at score level j answering the item correctly and Dj is the observed number of examinees in the focal group getting a keyed item wrong (Narayanan & Swaminathan, 1994). M-H calculates two statics, the significance and the size of the difference. With large sample sizes, small differences can be reported as significant, thus leading to inflated Type I error. If α_ *MH* is greater than 1, the studied item is favoring the reference group; on the contrary, an α_ *MH* of less than 1 indicates that the studied item is favoring the focal group. The α_ *MH* statistic is often transformed on the delta scale used by the Educational Testing Service (ETS) to measure item difficulty via

$$\Delta_{MH} = 2.35 \ln(\alpha_{MH})$$

According to the ETS system for categorizing the severity of DIF (Zieky, 1993), a value of | Δ_ |. *MH* ≥1.5 indicates that the item must be carefully reviewed (Fidalgo, 2004).

$$\chi^2_{Mantel} = \frac{\left(\sum_k Fk - \sum_k E(Fk)\right)^2}{\sum \mathrm{var}(Fk)}$$

where F_k represents the sum of scores for the focal group at the k^{th} level of the matching variable (ability). Under the null hypothesis:

$$\boldsymbol{H}_{0:}\, P_{Rj}\,(1\text{-}P_{Rj}) = P_{Fj}\,(1\text{-}P_{fj})$$

for all values P_{Fj} where j represents the total score on the test, P_{Rj} is the probability in the reference group that an examinee with a total score of j on the test will respond to the studied item correctly, and P_{Fj} is the probability in the focal group that an examinee with the total score of j will not respond to the studied item correctly, the M-H statistics has $\chi 2$ distribution with one degree freedom. Rejection of the null would indicate that examinees in the reference and focal groups who are matching on overall ability differ in their mean performance on the studied item and thus shows DIF (Ryan, 1991).

Many practitioners use M-H because it is favorably comparable to latent trait procedures and the users claim that it takes much less time to calculate (Clauser, Mazor, & Hambleton, 1991). M-H has gained a level of acceptance that, Holland and Weiner (1993),

mentioned that it is a standard against which new methods could be judged before adoption by measurement practitioners. However, Whitmore and Schumacker (1999) noted that at some test score levels, even the later versions of M-H procedure can have incomplete contingency table cells that cannot be used, resulting in the loss of data and subsequent reduction in score reliability. The M-H procedure (Mantel & Haenszel, 1959) had been proposed as an alternative procedure to IRT methods in investigating DIF (Holland, 1985; Holland & Thayer, 1986). The procedure has been used to detect DIF on tests of educational achievement (McPeek & Wild, 1986; Zwick & Ericikan, 1989). The Educational Testing Service (ETS) incorporated it as a standard procedure for a number of major testing programs (Anrig, 1987; Kubiak & Colwell, 1990).

Mantel-Haenszel Analyses

The M-H procedure consists of the M-H common odds ratio and the M-H chi-square statistic. Holland (1985) suggests transforming the common odds ratio to the M-H delta difference (M-H D-DIF) and calculating the standard errors of the resulting M-H D-DIF (Phillips & Holland, 1987). The M-H D-DIF is the difference in item difficulty in the ETS delta metric for comparable reference and focal group members (Scheuneman & Gerritz, 1990). In the middle range of the delta scale, an absolute value of 1 for M-H D-DIF is considered to be approximately a 10% difference in item difficulty (Kubiak & Colwell, 1990). A negative value for M-H D-DIF indicates that the item is easier for the reference group than for the focal group and the alternative interpretation is appropriate for positive M-H D-DIF values (Ryan, 1991).

Among the advantages of the M-H procedure is the sample size requirement. Kubiak and Colwell (1990) indicate that for the ETS testing programs a sample size of 500 for the groups combined and a minimum of 100 for the focal group are considered adequate for purposes of test assembly. Hills (1989) suggest 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 M-H statistic (Mantel & Haenszel, 1959), the use of these procedures generally has failed to produce meaningful interpretations of bias (Buhr, 1988; McPeek & Wild, 1986; Zwick & Ercikan, 1989). Mazor, Clauser, and Hambleton (1992) reported relatively poor DIF detection results using the M-H 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 M-H 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).

One potential explanation of these results is that the detection methods are not particularly reliable and that many of the identifications of biased items are statistical fluctuations of the item response data. Some support for this hypothesis has come from Hoover and Kolen (1984) who investigated the reliability of several bias detection techniques across random samples from the same population (standardization population) using the Iowa Tests of Basic Skills. For comparisons based on sex and ethnic groups, they found negligible reliability among the indexes used in the study. However, several recently developed approaches to analysis of bias, including the IRT sum-of-squares and M-H methods have been shown to have improved utility over those used in the Hoover and Kolen study (Skaggs & Lissitz, 1992).

SIBTEST verses Mantel-Haenszel

In this study the performance of SIBTEST was compared with that of M-H because does not require local item independence that latent trait procedures for DIF detection require (Holland and Thayer, 1988). The aim of the comparison of the two procedures is to determine if the effect of unequal sample sizes is dependent on the procedure for DIF detection. With SIBTEST, if the null hypothesis is rejected, members of the reference and focal groups with the same underlying trait are regarded to differ in their probability of a correct response to that item on the test; hence, the item is identified as exhibiting DIF. On the other hand, with M-H an item is considered as exhibiting DIF when the members of the reference group that are identical in overall ability differ in their mean performance on the studied item.

WinGen Computer Software

WinGen2 (Han, 2006) was developed to generate dichotomous and polytomous item response data for IRT models and for many conditions that arise in practice (Han & Hambleton, 2007). WinGen2 generates IRT model parameter values from various distributions for realistic data. It is capable of generating item parameters and sets of examinee theta parameters to create realistic item response data from various kinds of distributions. With WinGen2 a user may chose normal, uniform, or beta distributions for the examinee parameter in a model, and a normal, uniform, beta, or lognormal distributions for

item parameters so that a researcher can conduct a study with more realistic IRT data sets. The software is also convenient to use because it immediately provides IRT plots such as the item characteristic curves (ICC) and histograms of examinee ability parameter that let the researcher know if the generated parameters suit the intended purpose.

CHAPTER THREE

Methodology

This chapter is the presentation of methodology of parameter estimation, data simulation and the models used in the generation of item and theta parameter estimates. The chapter also contains procedures for data replication and analysis. In addition, the chapter provides justification for the choices of procedures and models used for simulation of the data. Also included in this chapter is a presentation of trial study results that guided the choice of the items whose difficulty parameters were manipulated to constitute DIF items with regards to their difficulty parameters.

The IRT-based Monte Carlo procedure was used for item parameter generation because with the IRT model the process of responding to items is inherent in the model therefore it is assumed that the generated values are valid (Harwell & Janosky, 1991). Harwell and Janosky suggest that as long as variances in parameters, such as item discrimination, *a,* item difficulty, *b,* or ability, θ, being estimated remain small; parameter estimates will be fairly stable even with quite small sample sizes of examinees in the reference and focal groups. Harwell and Janosky's study was designed with equal sample sizes in the focal as in the reference group. It is hard to tell from the results of their study how large a difference between the reference and focal group sample sizes may affect the stability in parameter estimates and DIF detection. This study departed from their design by varying the sample sizes in the reference as well as in the focal group to determine how unequal sample sizes affect the statistical power of a DIF detection procedure.

The Study Design

To assess the effect of unequal sample sizes on DIF detection varied sample sizes of examinees were generated, as 1000; 500; and 250 for the reference group. The focal group sample sizes were defined by the combinations ratios of 1; .5; and .1. There were two levels of DIF magnitudes: moderate and large DIF. Thus 18 conditions were created in a 3 X 3 X 2 factorial design with three levels of samples sizes, three sample size combination ratios and two levels of DIF magnitudes as specified in Table 1.

Table 1

Independent Variables Manipulation

Sample Size	Combination Ratios			Moderate DIF Magnitude	Large DIF Magnitude
	1.00	.50	.10	.35	.65
1000	1000/1000	1000/500	1000/100		
500	500/500	500/250	500/50		
250	250/250	250/125	250/25		

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 and Stout, 1996) did not report any significant Type I error inflation with as small a 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. Studies that report the performance of DIF detection procedures with unequal sample sizes in the focal and reference groups are limited. Gierl, Bisanz, Boughton and Khaliq (2001) suggested that the difference in sample sizes for the reference and focal groups should not be large enough to hinder the detection of items that behave differentially for examinees with the same ability levels on the latent trait being measured across the groups of examinees. Gierl et al., recommended that fair tests aught be free from bias because when bias occurs, test score interpretations may result in different meanings for members of different groups of examinees taking the same test. In this study the selection of the sample size in the focal group that was matched to the sample size in the reference group was estimated to cover the range in sample sizes that would likely be observed in most minority group studies.

Item and Person Parameter Estimation

Monte Carlo techniques with an IRT-based model were employed to generate the item and the person parameters in WinGen2 (Han & Hambleton, 2007). WinGen2 provides a user friendly interface and was able to support normal, uniform and lognormal distributions that were needed to generate item parameters for reference and focal groups this study. The item parameters for the focal group were the same as those for the reference group except for three items whose parameters were manipulated to show uniform DIF of medium and large

magnitudes. Item parameter generation and DIF manipulation were done at the beginning of the study and the same parameters were used for all the cells of the design (see Table 1), and for all replications within each cell. MULTISIM was employed to generate person abilities and item response data for each replication within each of the 18 cells of the study design. A new seed was generated for every replication to ensure that person abilities were sampled randomly across cells and across all iterations. As a result, 18,000 person abilities sets were generated. SAS 9.3.1 provided an environment within which the response data were generated and for the replication and analysis of the data by SIBTEST with regression correction and an older SIBTEST version for M-H DIF analysis.

Generating Item Parameters

For this study fifty dichotomously scored items were generated using a 2PL IRT because as Finch and French (2007) stated, DIF is typically investigated for the *a* and *b*-parameters. In some studies where the 3PLM had been used (e.g. Kim, Cohen and Kim, 1994) the pseudo guessing parameter *c* was fixed and therefore had no differential effect on the results of the study. The *a*-parameter estimates were generated to be lognormal, N (0, .2) and the *b*-parameter were generated to have normal distribution, N (0, 1). Then DIF was introduced by altering the item *b*-parameter for the items that were selected to display DIF (e.g. Shepard, Camilli, & Williams, 1985). Three items from the fifty (6%) were modeled to display uniform DIF of moderate and large magnitudes in the focal group's item difficulty parameters. Gierl, et al., (2004) reported that SIBTEST yielded adequate detection rates of uniform DIF with 1,000 sample sizes in the reference and focal groups. In this study, the number of DIF items was not the central focus. However, the selection of three items with varying difficulty was to help in the interpretation of the effect of unequal sample sizes on the statistical power of the DIF detection procedures. In this case where the test was composed entirely of dichotomous items, guidelines for categorizing the magnitude of DIF effect variance was established by considering the interpretation employed by other researchers in making decisions concerning the level of DIF in dichotomous items (e.g., Penfield, 2005; Roussos & Stout, 2004; Zwick, Thayer, & Lewis, 1999).

Manipulation of Sample Sizes

The sample sizes were manipulated beginning with matching equal samples of examinees in the focal group as was in the reference group as a bench mark because it was already known that matching sample sizes in the reference and focal groups provides the best

power for DIF detection. In some empirical DIF studies (e.g. Angoff, 1993; Roussos & Stout, 1996) reference group sample sizes have been manipulated so that the numbers in the reference group and the focal groups were the same based on the assumption that DIF methods are based on comparable examinees. SIBTEST requires that test items be divided into two parallel subtests: a "studied" subtest (a subtest of items believed to exhibit DIF administered to the focal group) and a matching (or "valid") subtest (a subtest of items believed to be DIF-free administered to the reference group). The reference and focal groups of examinees are compared on the basis of their ability on the latent trait that the test is intended to measure. M-H, on the other hand does not require a separate matching subtest. For M-H the test for the reference and focal groups were combined. Simulated data was used in this study because testing accommodation focal groups data are usually very small compared to the samples in the reference groups. The response data obtained with MULTISIM were 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. Monte Carlo procedure made it possible to simulate data with varied sample sizes that were aimed to provide information regarding a desirable level of precision in DIF detection without inflating the Type I and Type II error rates. Data for the reference and focal groups were generated from the 2-PLM. The probability of a correct response for an examinee on items *i* for the 2-PLM is:

$$Pi(\theta) = \frac{\exp[1.7ai(\theta - bi)]}{1 + \exp[1.7ai(\theta - bi)]}$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, θ is the ability level parameter and 1.7 is a scaling factor to transform the logistic metric to an approximate normal metric (Welkenhuysen-Gybels, 2004). Fifty items that were dichotomously scored were generated using a 2PLM because many studies have used dichotomous data in examining the accuracy of DIF procedures in DIF detection but few of the studies were with unequal sample sizes. Besides, the dichotomous data was selected to eliminate the complexity that come with graded responses and polytomously scored items.

Manipulation of DIF

The guidelines for DIF magnitude stemmed from the Educational Testing Service (ETS) classification scheme (Zieky, 1993). In this classification scheme a value of .43 and .64 added to the item b-parameter are regarded as small to moderate DIF which Roussos and

Stout associated with a SIBTEST β-$_{UNI}$ value of 0.04 for small DIF and .088 for moderate DIF; and 1.0 and 1.50 for Δ_{MH} small and moderate DIF magnitudes respectively (Roussos & Stout, 1996). A value greater than .64 is associated with a categorization of a substantial DIF greater than SIBTEST β-$_{UNI}$ value of .088 and Δ_{MH} value greater than 1.50. Not all studies have followed the ETS guidelines for DIF magnitude specifications. Shepard, Camilli and Williams (1985), for example, used a difference of .35 in the b-parameter to represent moderate DIF.

In this study items will be examined to determine if a priori manipulated items demonstrate DIF. Accurate DIF detection was considered to indicate that the difference in sample size was sufficient to guide accurate interpretation of the test results for the examinees in the reference and the focal groups. A difference in sample size between the reference and focal groups that did not lead to the detection of all the manipulated DIF items was considered to be a demonstration that with that sample the DIF detection procedure did not have enough statistical power to detect some DIF items that lead to test bias**.** A test item that demonstrated DIF was considered to introduce a secondary dimension to the measure and hence suggested that the test would not be measuring the same construct. Therefore decisions made on the basis of such test results were more likely than not to be biased.

Pilot Study

For this study fifty dichotomously scored items were generated randomly using a 2PL IRT model. The 2PLM was chosen because Finch and French (2007) stated that DIF is typically investigated for the *a* and *b*- parameters. The *a*-parameters were generated randomly to be lognormal, N (0, .2) and the *b*-parameter were generated randomly to have normal distribution, N (0, 1). The generated item parameters are presented in Table 2. DIF was introduced by altering the item *b*-parameter for the items that were selected to display DIF, using guidelines in Shepard, Camilli, and Williams (1985) study. For the pilot study, three items from the fifty were modeled to display uniform DIF of moderate magnitudes in the focal group's item difficulty parameters by increasing the *b*-parameters of items 1-3 by .35 which Shepard et al., regarded as moderate DIF magnitude in their study. Only one cell in the study design was used with DIF of medium magnitude. Item response data for 1000 examinees in both the reference and focal groups were also generated by WinGen 2. The old SIBTEST with M-H was used for DIF analysis. Data were not replicated prior to analysis because one of the objectives for the pilot study was just to try out the process of data generation and DIF detection with SIBTEST. Another purpose of the pilot study was to help

in establishing guidelines for choosing the items to manipulate for DIF considering that highly discriminating items are likely to demonstrate DIF (Harwell et al., 1996).

Table 2

Item Parameter Distribution

Item no.	Item Parameter		Item no.	Item Parameter	
	a	B		A	b
1	1.10	.06	26	.73	.12
2	.83	-1.80	27	.91	-1.47
3	.79	.22	28	.97	-.74
4	1.25	.23	29	1.27	-.82
5	.84	.54	30	1.15	-.03
6	.86	1.14	31	1.13	-.02
7	1.11	-.62	32	.83	.90
8	.90	.46	33	1.15	.19
9	1.07	.62	34	.91	.78
10	1.44	-.43	35	1.27	-.81
11	1.31	1.60	36	.90	-.56
12	.90	.48	37	1.31	-2.72
13	.84	-.97	38	.80	-1.07
14	.81	-.21	39	.61	-.03
15	.93	-.61	40	1.13	-.34
16	.87	.19	41	.79	.03
17	1.05	.31	42	.86	.80
18	.92	.24	43	.97	.17
19	.75	-.17	44	.86	-.92
20	1.10	.49	45	1.18	-.39
21	.96	-.84	46	1.09	1.07
22	1.50	-1.55	47	1.13	.20
23	.97	1.20	48	1.08	.48
24	.77	1.41	49	1.20	-.25
25	1.00	-.71	50	.80	-.20

Only the results of 10 items including the DIF items were presented in Table 3 because they were sufficient to show the results of DIF analysis of the DIF items and a few non DIF items. The results of the pilot DIF analysis were presented in Table 3. The results show how the studied items display DIF that would be classified as moderate because the *β-uni* values, which are the DIF effect size, were within SIBTEST *β-uni* value of .04 -.088 regarded by Roussos and Stout as moderate DIF, and also close to the 1.0 – 1.5 values for M-H delta (Δ], regarded as moderate DIF (Roussos and Stout, 1996). All the three items that were manipulated to display DIF had a statistically significant SIBTEST *β-uni* and M-H D-DIF indicated by the *p-value* < .05. The non DIF items were not falsely flagged for DIF

suggesting that there were no Type I errors. The results of the pilot study were interpreted with caution because no replication of data was conducted and only equal sample sizes for the reference and focal groups were used.

Table 3

Trial SIBTEST and M-H DIF Analysis Results

		SIBTEST			Mantel-Haenszel		
Run No.	Item No.	β	β-uni	SIB-uni p-value	χ^2	p-value	Delta (D-DIF)
1	1	.60	3.20	.00	13.89	.00	-.98
1	2	.70	3.09	.00	6.00	.01	-.60
1	3	.11	5.32	.00	26.02	.00	-1.37
1	4	-.01	-.49	.63	.15	.69	.12
1	5	.00	.20	.84	.01	.93	.07
1	6	-.01	-.59	.56	1.11	.29	.27
1	7	-.03	-1.41	.16	2.09	.15	.38
1	8	-.02	-.88	.38	1.10	.29	.27
1	9	-.01	-.38	.71	.08	.77	.08
1	10	-.01	.18	.86	.18	.67	.12

Manipulation of DIF in the Main Study

The pilot study results support the findings of Skaggs and Lissitz (1992) that showed that SIBTEST had improved utility. On the other hand, Roussos and Stout (1996) reported inflated Type I error for studied items with high discrimination and low difficulty, (a = 1.0, b = -1.5); (a = 2.5, b = -1.5); (a = 2.5, b = -0.5) with SIBTEST. Consequently, items with either very small or very high a-parameter values were not included among the studied items and the differences in the b-parameters for the studied items were altered as in Miller and Oshima (1992) study in which they specified moderate DIF by increasing the b-parameter by .35 for moderate DIF. However, .65 which fits the ETS specification for large magnitude DIF was added to the b-parameter of the items that were expected to display large DIF. Hence in the focal group test of 50 items, uniform DIF was built into three of the item difficulty parameters. Uniform DIF is said to exist if each i and j the ordinal ratio functions (ORFs) $\theta_{ij}(\theta)$ are equal for all θ (Hanson, 1998; Mellenbergh, 1982; Fischer, 1993, 1995). Hanson and Mellenbergh suggested that when uniform DIF exists, the relationship between item response and group is constant for all levels of the ideal matching variable. Hence:

$$\Theta_{ij}(\theta) = \theta_{ij}$$

Thus for this study the b-parameters were increased as: $b_{iF} = b_{iR} + .35$ (for moderate magnitude); and $b_{iF} = b_{iR} + .65$ (for large magnitude).

Items 1(b = .06, a =1.10); item 13(b = -.97, a = .84) and item 46 (b = 1.09, a = 1.07) included in Table 2 were modeled to display uniform DIF. The items that were manipulated to demonstrate DIF were chosen because they had difficulty that varied from fairly low to fairly high and the discrimination parameters were not very high as Shepard et al., (1985) recommend in their study. The item difficulty ranged from -2.72 to 1.60 with a mean of -.09 and a standard deviation of .70. The discrimination parameter estimates ranged from .61 to 1.50, with a mean of 1.16 and standard deviation of .70

The person abilities were generated randomly from a normal, N (0, 1) distribution using MULTISM. Individual item responses were also generated using the probability of a correct response from the 2PLM and a random uniform distribution between 0 and 1 is less than or equal to the probability of correct response; otherwise the response to the items is considered incorrect, using MULTISM. Thus for 50 examinees with the same θ, if the $P_i(\theta)$ were .5, then 25 of the examinees would likely receive a 1 and the remaining 25 would receive a 0.

It has been noted that highly discriminating items are particularly vulnerable to false flagging which lead to disproportionate removal of most informative items (Harwell et al., 1996). When selecting items to be manipulated for DIF, items with obviously large a-parameters were avoided, but since it was not clear what size of the a-parameter would be treated as large by the DIF detection procedures, flagged non DIF item parameters were examined to check if false flagging was related to discrimination parameters to avoid making recommendations that would lead to disproportionate removal of most informative items. The performance of SIBTEST and M-H procedures with unequal sample sizes was observed at critical values of the procedures' statistics corresponding to a significance level of .05 for the alternative hypothesis of DIF against focal group. Rejection rates were observed for the 18,000 replicated data sets that were randomly assigned to the 18 conditions by 1,000 iterations within each condition (Roussos & Stout, 1996). If the null hypothesis is rejected, members of the reference and focal groups with the same underlying trait differ in their probability of a correct response to that item on the test; hence, the item is identified as exhibiting DIF. In this study, it was expected that only the items that had DIF modeled into them would display DIF in the same direction as modeled in the data. The DIF analysis procedure was non-iterative. As such all items at or below the significance level of .05 will be flagged for DIF.

Assessment of Type I and Type II Error Rates

To assess Type I and Type II error behavior the items that are modeled to display DIF will be observed against a null hypothesis of no DIF. SIBTEST procedure (Chang, Mazzeo, & Roussos, 1993) will be employed in the analysis of the simulated data in this study because the procedure is a popular latent trait/parametric DIF detection method that matches examinees based on latent trait (Hanson, 1998). Welkenhuysen-Gybels (2004) argues that techniques that match examinees on the latent trait of interest rather than on observed scores are preferable because they provide the most proper treatment of the matching variable that the test is supposed to measure such as ability. In Finch and French's study on crossing DIF (CDIF), SIBTEST was the best performer when both Type I error and statistical power were considered. Unlike other power methods, SIBTEST did not appear susceptible to problems caused by group ability differences.

SIBTEST was particularly selected for this study because it has the ability to study the simultaneous effect of several DIF items and also because of its potential to match examinees based on the distribution of the latent trait. An item was considered to display DIF when examinees from reference and focal groups' observed DIF statistic was significantly different across the reference and the focal groups indicating that there were group differences in the distribution of the latent trait in the construct that was being measured. In other words, items that displayed *β-uni* or D-DIF (DIF effect sizes) that were statistically significant at .05 suggested that there were differences in the measurement properties of the test items for the reference and the focal groups in the construct of interest.

Data Analysis

SIBTEST (Shealy & Stout, 1993) provided M-H statistics while the newer version of SIBTEST with regression correction (Roussos & Stout, 1996) provided SIBTEST statistic because of its ability to control Type I error. The latent variable DIF procedures were based on the idea that test scores are a measure of both reliable and unreliable portions of the test score (Potenza & Dorans, 1995), making it possible to assume that error of measurement would not affect DIF detection. Since SIBTEST is a latent variable, but in DIF statistics calculations it has been applied as a nonparametric approach. Hence examinees in the reference and focal groups were matched based on their raw score on the "valid subtest" and the score points were weighted by frequency. In the SIBTEST approach, test items were assigned to two subtests: the matching subtest and the studied subtest (Fidago, Ferreres, &

Muňiz, 2004). The studied subtest consisted of the item being studied while the remaining 49 items constituted the matching subtest.

A parameter for DIF items was simulated and displayed. The generated parameters were studied to determine if the items in which DIF was built display DIF. Any other items that produced unexpected group differences were flagged at .05. The non DIF items flagged for DIF were examined to establish why the items were displaying DIF (Stout & Roussos, 1995; Roussos & Stout, 1996a; Bolt & Stout, 1996). Items that elicited positive values of the SIBTEST DIF statistic were considered to produce DIF that favored the examinees in the reference group and negative values of SIBTEST DIF statistic were considered to produce DIF that favored the focal group. DIF items that produced non-significant SIBTEST DIF statistics at .05 constituted Type II error. DIF items that were flagged for DIF at .05 constituted the statistical power of the SIBTEST procedure. Non-DIF items that produced significant SIBTEST DIF statistics were regarded as false positives that constituted the Type I error. Positive DIF was modeled in this study to mimic the bias against examinees in the focal group that is usually experienced with empirical data (Zieky, 1993). The older version of SIBTEST (Shealy & Stout, 1993) was used to calculate the M-H DIF statistics while the newer version was used to calculate the SIBTEST DIF statistics for each replication. The same flagging criteria were set for the M-H DIF detection procedure.

CHAPTER FOUR

Manuscript 1

Effect of Unequal Sample Sizes on the Statistical Power for DIF Detection: An IRT-Based Monte Carlo Study with SIBTEST Procedure

Abstract

The power of SIBTEST procedure for DIF detection is substantially affected by small differences in parameters under study. Differences in sample sizes have also been reported to affect the power of the procedure for DIF detection. On the basis of the information reported in previous studies regarding the statistical power and Type I error rate of SIBTEST procedure, this simulation study investigated the effect of unequal sample sizes in the reference and focal groups on the detection of a priori determined DIF items using the current SIBTEST with regression correction. Sample sizes (1000; 500; 250) were manipulated in the ratios (1:1; 1:.50; 1:.10) in the reference and focal groups respectively, thus creating 9 cells. The nine cells were studied under large, and then moderate DIF magnitudes, resulting in a total of 18 cells that were studied. Item parameters were generated with 2PLM. The item difficulty parameters were generated from a random, N (0, 1) distribution. The discrimination parameters were generated from a lognormal (0, .2) distribution. Ability parameters were simulated from N (0, 1) distribution using MULTISIM. The MULTISIM was also used to generate response data which were then analyzed using SIBTEST. SAS provided the environment for carrying out the operations of MULTISIM and SIBTEST. The results indicated that large sample sizes in the ratio 1:.10 resulted in inflated Type I error with apparent sufficient statistical power for DIF detection. Inflated Type I error rates were found at the ratios of 1:.10 with 250 and 500 reference group sample sizes.

Key words: Differential item functioning, SIBTEST procedure, DIF magnitude, large sample, small sample, sample size ratio.

With increased immigration and settlement of people from diverse ethnic backgrounds, as well as the requirement to include all students with varying degrees of special education needs in the large scale educational assessments, psychometricians have the responsibility of ensuring that accurate assessment occurs for all examinees. Thus, items intended to measure a specific latent trait, for example, reading proficiency, must be valid for use with students from diverse groups including ethnicity, gender, and special education status, without being characterized by differential item functioning (DIF). DIF can be an

indicator of irrelevant variance that may influence test scores and lead to biased decisions (Finch & French, 2007). Therefore, for meaningful score interpretation, test items that demonstrate DIF should be detected and either studied more closely to inform test score interpretation or be discarded.

Both the empirical and simulation studies show that the statistical power of procedures for detection of DIF is positively related to sample size (Chan, 2000; González-Romá, Hernández, & Gómez-Benito, 2006; González -Romá, Tomas, Ferreres, & Hernandez, 2005; Kaplan & George, 1995; Narayanan & Swaminathan, 1996; Wasti, Bergman, Glomb, & Drasgow, 2000). Kaplan and George's study that examined the power of Wald test within the multi-group confirmatory factor analysis (CFA) reported that with marginal inequalities in sample sizes, large changes in the power of the test were observed even under homogeneity of factorial variance.

Several measurement researchers recommend the use of large sample sizes and long tests to ensure reliability of the results when decision is based on raw test score (Chan, 2000; González-Romá, et al., 2005, 2006; Millsap & Meredith, 1992). In the González-Romá et al., (2006) study, they reported low power for DIF detection when sample sizes in the focal groups were small. They conducted a simulation study to investigate statistical power and Type I error rate of a procedure based on the mean and covariance structure analysis (MACS) model to detect uniform and non-uniform DIF when the percentage of DIF item was large and sample sizes were equal and unequal. The results of their study indicate that power increased as sample size increased. The procedure used showed acceptable statistical power levels ($\geq$ 70%) for detecting uniform and non-uniform DIF of medium magnitude, when an increase in the b-parameter by 0.10 was regarded as low, 0 .25 as medium and 0.50 as large with as low as 200/200, and 400/200 sample sizes in the reference and focal groups respectively. The results of the study also suggest a possibility that when sample sizes in the focal group are small, item parameters that are not constrained to be equal across groups may be poorly estimated, leading to inadequacy in statistical power for DIF detection.

In Roussos and Stout's (1996) simulation study in which they examined the effect of small sample sizes and studied item parameters on Simultaneous Item Bias Test (SIBTEST) and Mantel-Haenszel (M-H) with equal sample sizes in the focal as in the reference groups, they reported that the amount of statistical bias due to item parameter variations remained approximately constant with sample size, while the associated false rejection rates decreased with decreasing sample size due to reduced power to detect the statistical bias.

DIF continues to receive considerable attention in educational measurement. Finch and French (2007) recommended continued evaluation of DIF detection procedures to improve the accuracy of DIF detection methods. They argue that the process undertaken to detect bias is an essential step in gathering test score validity evidence and additional information on DIF detection. They also suggested that the assessment of DIF detection procedures to ensure accurate DIF detection has become more crucial as a result of the realization that highly discriminating items on a test are particularly vulnerable to false flagging for DIF. False flagging of highly discriminating items may lead to disproportionate removal of most informative items. Considering that constructing good test items is a time consuming and costly endeavor, discarding perfectly good items because of inaccurate functioning of a statistical procedure entails very serious educational and economic implications (Finch & French, 2007).

Swaminathan and Rogers (1990) and Rogers and Swaminathan (1993) have classified DIF into two categories: uniform and non uniform DIF. In their classification they suggest that uniform DIF exists when there is no interaction between ability level and group membership, thus, 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. Many studies have relied on large differences in mean scores across groups to indicate the presence of bias (e.g., Rosser, 1989). Angoff, (1993); McAllister, (1993); and Camilli and Shepard (1994) argue that a group mean difference is not sufficient evidence of bias because it may reflect some valid group differences. Camilli and Shepard acknowledged that SIBTEST DIF methodology controls for ability while detecting items that exaggerate the ability difference across groups of examinees. That being the case, a large DIF value obtained with the SIBTEST procedure suggests that the item is more likely to be measuring additional constructs that function differently from one group to another (Angoff, 1993; Camilli & Shepard, 1994; Roussos & Stout, 1996; Shealy & Stout, 1993).

Several Monte Carlo DIF detection studies have looked at the influence of sample size on DIF detection to determine sample size that results in minimal variance and least error rates with varied DIF detection procedures. However, studies that are focused on the effect of unequal sample sizes that mimic the reality that exists in empirical data, especially testing accommodation data, are still limited. The rationale for this study is to add to the limited existing literature on the effect of unequal sample sizes on the statistical power of DIF

detection using SIBTEST procedure specifically with an aim of determining the smallest sample size in the focal and reference groups that would lead to reduced Type I error rates and sufficient statistical power for accurate DIF detection.

The SIBTEST Procedure

The SIBTEST procedure fits in the category of nonparametric procedures for detecting DIF. A nonparametric procedure is based on ranked data and it calculates statistics of interest without reference to specific parameters while making less stringent demands of the data. The SIBTEST therefore, when taking a nonparametric procedure's approach, should not require certain underlying conditions or assumptions that must be met for it to be valid. It was developed as an extension of Shealy and Stout's (1993) multidimensional item response theory. According to Shealy and Stout, within the framework of SIBTEST, DIF is conceptualized as a difference in the probability of endorsing a keyed item response occurring when individuals in groups having the same levels of the latent attribute of interest possess different amounts of nuisance abilities that influence response. With the SIBTEST procedure DIF analyses are conducted using original item response data rather than parameter estimates from a program. For example, SIBTEST can be used to identify items that are biased against a particular group or groups of examinees using item response data. For a scale containing *k* items, *k* single-item DIF statistics are computed to obtain p-values indicating the significance of the outcomes. If an observed p-value is less than the "corrected" critical p-value, then the null hypothesis of no DIF can be rejected:

$$\boldsymbol{H}_0: \beta_{UNI} = 0 \text{ vs. } H_1: \beta_{UNI} \neq 0 \qquad (1)$$

where β_{UNI} is the parameter specifying the amount of DIF for an item (Shealy & Stout, 1993). β_{UNI} is defined as:

$$\beta_{UNI} = \int B(\Theta) f_F(\Theta) d\Theta, \qquad (2)$$

where $B(\Theta)$ 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 at given levels of ability can be expressed as:

$$B(\Theta) = P(\Theta, R) - P(\Theta, F) \qquad (3)$$

SIBTEST detects bias by comparing the responses of examinees in the reference and focal groups that have been allocated to strata within cells using their scores on a "matching subtest" (Stout & Roussos, 1996b). The matching subtest is a subset of items that are known to be unbiased. In most practical applications the user of the procedure does not have accurate

a priori knowledge regarding bias. Simulation studies have shown that the SIBTEST procedure is tolerant of small to moderate amounts of contamination of the matching criterion (Shealy & Stout, 1993). Shealy and Stout refer to DIF items in the matching subtest as contamination. 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.

In this study the newer version of SIBTEST (Shealy & Stout, 1993) was used to compute a weighted mean difference between the reference and focal groups' probability of correct response to an item at a certain ability level. The means in this procedure were 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 while improving the statistical power of the DIF detection procedure (Bolt & Gierl, 2004; Roussos & Stout, 1996b; Shealy & Stout, 1993; Stout, Li, Nandakumar, & Bolt, 1997).

SIBTEST with Regression Correction

The emphasis in the improvement and evaluation of SIBTEST has been placed on the control of Type I error inflation and estimation of bias. When there is no target ability difference between the reference and focal group, it can be shown that estimated beta is an unbiased estimate of beta (β). The weighted mean difference between the reference and focal groups on the subtest item across the Θ subgroups is computed with

$$\hat{\beta}_{UNI} = \sum_{k=0}^{k} P_{kdk} \qquad (4)$$

In this equation, P_k is the proportion of focal group examinees in subgroup k and $d_k = P_{Rk} - P_{Fk}$, which is the difference in the adjusted means of studied subtest item for the reference and focal groups respectively, in each subgroup k. The means on the studied subtest item are adjusted to correct for any differences in the ability distributions of the reference and focal groups using a regression correction (Shealy & Stout, 1993).

SIBTEST also yields an overall statistical test for DIF effect size. The test statistic for evaluating the null hypothesis is

$$SIB = \frac{\beta_{UNI}}{\sigma(\beta_{UNI})} \qquad (5)$$

When the examinees from the reference and the focal groups with the same ability in the latent trait of interest demonstrate a difference in their probability for a correct response to the keyed item then beta would be considered biased. In order to more accurately estimate β, SIBTEST adjusts estimated beta for possible differences in the ability distribution for the reference and focal groups examinees by estimating true matching scores in the matching subtest for the examinees in the reference and focal groups as expressed in the equations 6- 8.

$$\bar{Y}_{g^*}(x) = \bar{Y}_g(x) + \hat{M}_g(x)[\bar{V}(x) - \bar{V}_g(x)] \qquad (6)$$

where

$$\hat{M}_g(x) = \frac{\bar{Y}_g(x+1) - \bar{Y}_g(x-1)}{\bar{V}_g(x+1) - \bar{V}_g(x-1)} \qquad (7)$$

and

$$\bar{V}(x) = \frac{1}{2}[\bar{V}_R(x) + \bar{V}_F(x)] \text{ with } V(x) = \frac{1}{2}[V_R(x) + V_F(x)] \qquad (8)$$

According to Jiang and Stout (1998), the adjusted studied item score $\bar{Y}_{g^*}(x)$ should be viewed as the unbiased estimator of the probability of correct response to the keyed items for an examinee in group *g* with a matching subtest true score of $V(x)$.

Jiang and Stout stated that the true regression estimate, $V(x)$, in equation 8 is assumed to be linear and yet it is not. As such the SIBTEST procedure has displayed inflated Type I error rates and sizable DIF effect sizes suggesting that there could be pseudo sources of the Type I error rate inflation when the linear assumption is used in calculating the bias. Jiang and Stout remarked that with SIBTEST if DIF is truly present the statistical power would be reduced to smaller levels than what would normally be observed with false flagging of items for DIF. Consequently they proposed a newer nonlinear approach to SIBTEST with regression correction. In the new approach, Jiang and Stout obtained point estimates of $V(x)$ for each *x* which they expressed as $Z_g(x)$. The point estimates, they said, were built up from proportions of correct estimators therefore their error variances were inversely proportional to the number of examinees in each stratum. In the newer SIBTEST with regression correction approach, Jiang and Stout used the weighted least squares technique to control for the variations that would otherwise be observed in the strata with less examinees if the ordinary least squares with equal weights were used. In this calculation of DIF, the weight for *x* is the number of examinees with observed matching subtest score *x* (Jiang & Stout, 1998). DeMars

(2008) stated that the SIBTEST procedure with nonlinear regression correction improves the accuracy of the effect size with large samples. Although she still obtained Type I errors with large samples sizes, she remarked that with small sizes, the estimation of the DIF statistics effect size was less accurate and thus resulted in a smaller number of non DIF items flagged for DIF, but those that had DIF were more likely to have large effect sizes than the statistically significant items with larger samples, leading to inflated Type I error rates (DeMars, 2008).

From the DIF studies that were reviewed, there is no doubt that sample size influences the power of DIF detection procedures. However, it is not yet clear how unequal sample sizes in the ratios of 1:.10 that are likely to be encountered with real data, affect the statistical power and Type I error rate control of the SIBTEST procedure. This study seeks to answer two main questions:

1. What sample size ratio combinations would produce sufficient power for DIF detection with SIBTEST DIF detection procedure?
2. What sample size ratio combinations result in adequate control of Type I error rates when the SIBTEST DIF detection procedure is used?

Method

For this study the 2PL IRT-based model was used to generate item parameters. With the IRT model the process of responding to items is inherent in the model therefore the model was considered to generate item parameter values that were assumed to be valid (Harwell & Janosky, 1991). Generating data that perfectly fit the 2PLM would minimize variability of item parameters and thus result in reduced unmodeled error. Harwell and Janosky stated that as long as variances in parameters, such as item discrimination, a, item difficulty, b, or ability, θ, being estimated remain small; parameter estimates remain fairly stable even with quite small sample sizes of examinees in the reference and focal groups. In this study the error rates of SIBTEST DIF detection procedure were observed with varying ratios of sample size in the reference and the focal groups as indicated in the study design displayed in Table 4.

Item and Person Ability Parameter Estimation

The 2PL IRT-based model was employed to generate the item parameters with WinGen2 (Han & Hambleton, 2007). The same item parameters were generated once for the reference and the focal groups. Item difficulty parameters of three of the items for the focal

group examinees were manipulated to show uniform DIF of medium and large magnitudes. The generated item parameters and the descriptive statistics are presented in Table 2. Item parameter generation and DIF manipulation were done at the beginning of the study and the same parameters were used for all the conditions in the study design (see Table 4), and for all the 1000 replications for each of the 18 cells.

The MULTISIM was employed to generate person abilities and item response data for each replication within each of the 18 cells of the study design. The response data obtained with MULTISIM were replicated up to 1,000 times for every condition to reduce chances of sampling error that would result in variance large enough to have a confounding effect. SAS provided an environment within which the response data were generated and replicated and also for analysis of the data by SIBTEST with regression correction.

Study Design

To assess the effect of unequal sample sizes on DIF detection, varied numbers of examinees were generated randomly as, 1000; 500; and 250 for the reference groups. The focal group sample sizes were determined by the combination ratios. The combinations between the reference and focal groups were done in the ratios of 1.00; .50; and .10. Thus, 18 conditions were created in a 3 x 3 x 2 factorial design with three levels of samples sizes, three sample size combination ratios and two levels of DIF magnitudes as specified in Table 4.

Table 4

Independent Variables Manipulation

Sample Size	Combination Ratios			Moderate DIF Magnitude	Large DIF Magnitude
	1.00	.50	.10	.35	.65
1000	1000/1000	1000/500	1000/100		
500	500/500	500/250	500/50		
250	250/250	250/125	250/25		

The sample sizes were selected on the basis of information from previous studies (e.g. Ankenmann, Witt, & Dunbar, 1999) that have reported Type I error inflation in DIF detection with sample sizes as large as 500/500 with Likelihood Ratio Goodness of Fit Statistics (LR) and on other studies (e.g. Roussos & Stout, 1996) that did not report any significant Type I error inflation with as small a 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. Gierl, Bisanz, Boughton, and Khaliq (2001) suggested that the difference in sample sizes for the reference and focal groups should not be large enough to hinder the detection of items that behave differentially for examinees with the same ability levels on the latent trait being measured across the groups of examinees. The study recommended that fair tests must be free from bias because when bias occurs, test score interpretations may result in different meanings for members of different groups of examinees taking the same test. The sample size in the focal group matched with that of the reference group samples were estimated to cover the range of the differences in sample sizes that would be observed in most minority group studies.

Table 5

Generated Item Parameters

Item no.	Item Parameter.		Item no.	Item Parameter.	
	a	b		a	b
1	1.10	06	26	.73	.12
2	.83	-1.80	27	.91	-1.47
3	.79	.22	28	.97	-.74
4	1.25	.23	29	1.27	-.82
5	.84	.54	30	1.15	-.03
6	.86	1.14	31	1.13	-.02
7	1.11	-.62	32	.83	.90
8	.90	.46	33	1.15	.19
9	1.07	.62	34	.91	.78
10	1.44	-.43	35	1.27	-.81
11	1.31	1.60	36	.90	-.56
12	.90	.48	37	1.31	-2.72
13	.84	-.97	38	.80	-1.07
14	.81	-.21	39	.61	-.03
15	.93	-.61	40	1.13	-.34
16	.87	.19	41	.79	.03
17	1.05	.31	42	.86	.80
18	.92	.24	43	.97	.17
19	.75	-.17	44	.86	-.92
20	1.10	.49	45	1.18	-.39
21	.96	-.84	46	1.09	1.07
22	1.50	-1.55	47	1.13	.20
23	.97	1.20	48	1.08	.48
24	.77	1.41	49	1.20	-.25
25	1.00	-.71	50	.80	-.20

Parameters	N	Minimum	Maximum	Mean	SD
A	50	.61	1.50	1.00	.70
b	50	-2.72	1.60	-.09	.85

Fifty dichotomously scored items were generated using a 2PL IRT model. The estimated parameters were presented on Table 5. The a-parameters were generated to be lognormal, N (0, .2) and the b-parameters were generated to be normal, N (0, 1), then DIF was introduced by altering the item b-parameter for the items that were selected to display DIF (Shepard, Camilli, & Williams, 1985).

Uniform DIF of moderate and large magnitudes in the focal group's item difficulty parameters were modeled in three of the 50 items following the guidelines for categorizing the magnitude of DIF effect variance and the interpretation employed by other researchers in making decisions concerning the level of DIF in dichotomous items (Penfield, 2005; Roussos & Stout, 2004; Zwick, Thayer, & Lewis, 1999).

Manipulation of DIF

The guidelines for DIF magnitude stem from the Educational Testing Service (ETS) classification scheme (Zieky, 1993). In this classification scheme a Mantel-Haenszel non significant D-DIF value of less than 1.00 is regarded as negligible DIF, which Roussos and Stout (1996) associated with a value of less than 0.04 SIBTEST β-$_{UNI}$. A significant D-DIF value of between 1.00 and 1.50 is associated with a categorization of moderate DIF that Roussos and Stout estimated in their study to be .088 SIBTEST β-$_{UNI}$. A significant D-DIF value greater than 1.50 is associated with a categorization of a large DIF (see Table 6). DIF magnitude of bias specified by Shepard, et al., (1985) was a difference in the *b*-parameter of .20 for the least detectible DIF and .35 for moderate DIF. These differences in *b*-parameters for biased items in math test were used in the Shepard et al., study because they had been cross-validated and were interpretable as valid differences in math test between white and black examinees.

Table 6

Guidelines for DIF Magnitude and DIF Effect Size for SIBTEST and M-H Procedures

		DIF Effect Size	
DIF Level	DIF Magnitude	SIBTEST Beta-uni	M-H D-DIF
Small	$\geq .20 < .35$	.04	1.00
Moderate	$\geq .35 < .64$	$.04 \leq .088$	$1.0 \leq 1.50$
Large	$>.64$	$> .088$	> 1.50

In this study items were examined to determine if a priori manipulated items were flagged for DIF. Accurate DIF detection was considered to indicate that the difference in sample size was sufficient to guide accurate interpretation of the test results for the examinees in the reference and the focal groups. A difference in sample size between the reference and focal groups that did not lead to the detection of all the manipulated DIF items was considered to be a demonstration that with that matching sample in the focal group the DIF detection procedure did not have enough statistical power to detect some DIF items that would lead to test bias**.** A test item that demonstrated DIF was considered to introduce a secondary dimension to the measure and hence suggested that the test would not be measuring the same construct in the two groups of examinees. Therefore decisions made on the basis of such test results would be considered to be biased.

Roussos and Stout (1996) reported inflated Type I error for studied items with high discrimination (a-parameter) and low difficulty (*b*-parameter), (a = 1.0, b = -1.5); (a = 2.5, b = -1.5); (a = 2.5, b = -0.5) with SIBTEST procedure. The items with high discrimination were not selected for DIF manipulation in this study to be able to study the influence unequal sample sizes and DIF magnitude on DIF detection without confounding effect that would possible result from large a-parameters. Since it was not clear what size of a-parameter would be treated as large by the SIBTEST DIF detection procedure, the a-parameters for items that were flagged for DIF were closely studied to determine the possible reasons for false flagging. The studied DIF items were varied to include items with low, moderate and fairly high difficulty to be able to generalize the finding of the study to the whole range of the test. The differences in the b-parameters for the studied items were altered as in Miller and Oshima (1992) study in which they specified moderate DIF by increasing the b-parameter by .35 for moderate DIF. For large DIF, the b-parameter was increased by .65 which fits the ETS specification for large magnitude DIF. Thus the b-parameters were increased as:

$b_{iF} = b_{iR} + .35$ (for moderate magnitude); and $b_{iF} = b_{iR} + .65$ (for large magnitude).

Uniform DIF was modeled based on the knowledge that the matching variable (θ) was similar in each of the matched subgroups of examinees (Fischer, 1993, 1995; Hanson, 1998; Mellenbergh, 1982). Items 1(a =1.10, b = .06); item 13(a = .84, b = -.97) and item 46 (a = 1.07, b = 1.09,) were modeled to display uniform DIF. The three items consisted of relatively low, moderate, and high degrees of difficulty. The DIF item selection criterion was unlike that used by Shepard et al., (1985) where the selected items for manipulation for DIF had difficulty estimates of all the studied items close to the mean. The effects of the interaction between the different magnitudes of DIF with different conditions of sample size were noted

(Miller & Oshima, 1992). The item difficulty estimates ranged from -2.72 to 1.60 with a mean of -.09 and a standard deviation of .70. The discrimination parameter estimates ranged from .61 to 1.50, with a mean of 1.16 and standard deviation of .70

Individual item responses were generated by MULTISIM using the probability of a correct response to items from the 2PLM and a random uniform distribution between 0 and 1 is less than or equal to the probability of correct response; otherwise the response to the item was considered incorrect. Rejection rates were observed with 1000 replications (Roussos & Stout, 1996b) in every cell with specified sample size ratios and DIF magnitude conditions.

Assessment of Type I and Type II Error Rates

For each condition of sample size and DIF magnitude, the item response data were replicated 1000 times to minimize differences that would probably result from sampling variance. The SIBTEST was run using a default option and SIBTEST was calculated for each of the 50 items in every replication using the other 49 items as the matching subtest. In this way, SIBTEST run an exploratory analysis for each test to determine if the procedure detected the DIF items. Normally with real data an initial exploratory analysis would be run followed by a second analysis that would be run without the flagged items. In this study the DIF items were known a priori. False flagging and non-significant DIF statistics for the DIF items were interpreted as Type I and Type II errors respectively.

To calculate the statistical power of the SIBTEST procedure for detecting the DIF items, frequency analysis was run using SPSS (Statistical Package for Social Sciences) for every data set consisting of 1000 iterations. The average of the percent of the proportions of flagging of the DIF items were calculated to represent the statistical power of the SIBTEST procedure since the SIBTEST DIF statistics are weighted by frequency. When the average percent of flagging the DIF item was less than 5% it was considered that the procedure had committed a statistically significant Type II error. The SIBTEST procedure's statistical power for DIF detection was calculated for all the three DIF items that were known a priori. In the same way, the percent of the proportions of flagging of the non DIF items were obtained by running frequency analyses. When the percent of flagging of non DIF items was greater than 5%, it was recorded that the procedure had committed a statistically significant Type I error. The Type I error rates estimated at .05 were calculated for all the 47 non DIF items under all the sample size and DIF magnitude conditions as was presented in the study design.

Results

As was expected, equal sample sizes of different sizes, 1000, 500, and 250 had the best power for detecting the a priori manipulated items to demonstrate DIF. A summary of the SIBTEST procedure's statistical power for detecting DIF in the studied items are displayed in Table 7.

Table 7

The SIBTEST Power for DIF Detection with Large and Moderate DIF Magnitudes

Sample Size	Item Parameters		Ratios					
	a-parameter	b-parameter	1:1		1:.50		1:.10	
			Large DIF	Mod. DIF	Large DIF	Mod. DIF	Large DIF	Mod. DIF
1000	1.10	.06	100.00	100.00	100.00	96.80	90.00	69.60
	.84	-.97	100.00	100.00	100.00	99.70	83.00	13.70
	1.09	1.07	99.80	97.10	99.90	96.60	91.70	61.90
500	1.10	.06	100.00	94.80	100.00	83.30	67.80	28.10
	.84	-.97	100.00	96.90	99.90	47.80	22.10	4.70
	1.09	1.07	99.60	62.40	98.90	79.60	51.40	4.30
250	1.10	.06	100.00	65.20	97.80	55.80	59.80	44.20
	.84	-.97	100.00	65.10	91.10	24.10	11.10	10.10
	1.09	1.07	86.00	28.10	73.70	35.10	16.90	10.40

As shown in Table 7, the statistical power for DIF detection was remarkably higher for equal samples than for unequal sample sizes as was evidenced by the high detection rates of the DIF items in most of the sample size ratios of 1:1, especially when the DIF magnitude was large. The results from this study with large equal samples sizes support Gierl, et al., (2001) suggestion that the difference in sample size for reference and focal group should not be large enough to hinder the detection of items that behave differentially.

The results with 250 sample sizes in the reference group at the 1:.10 ratios and moderate DIF were not reliable as the SIBTEST procedure appeared to be very unstable. With large DIF magnitude, estimated beta that was correctly detected for DIF ranged from -.07 to -.24. The negative estimated beta indicated that the items favored the reference group as was expected because DIF was manipulated to be unidirectional in favor of the reference group. Figure 1 shows the line graphs of the SIBTEST procedure's statistical power at different levels of sample sizes and combination ratios with large and moderate DIF magnitudes.

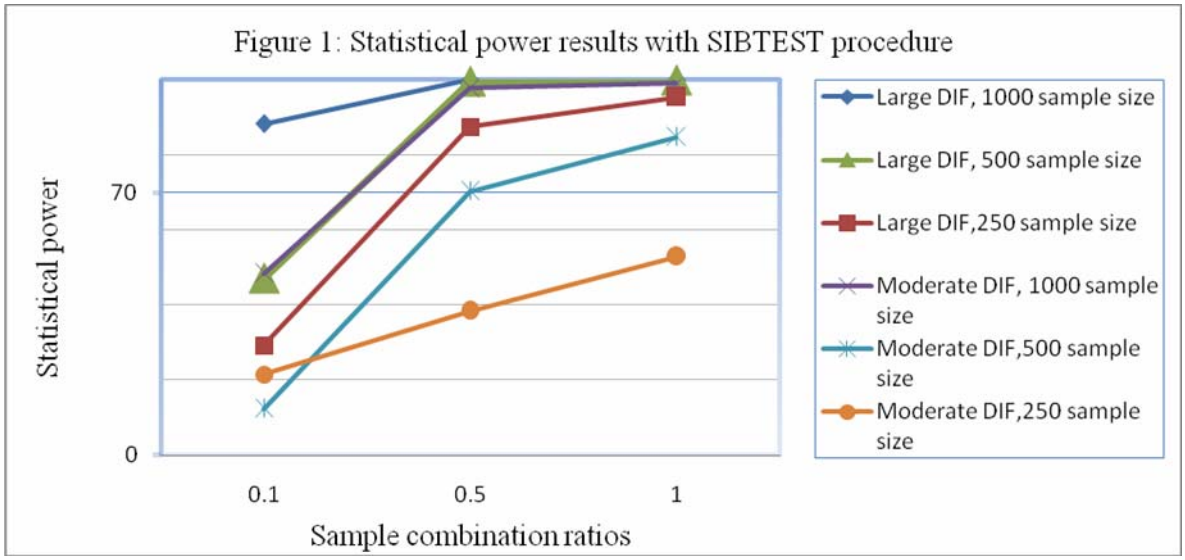


Figure 1. Statistical power results with SIBTEST procedure.

As was expected, equal samples of different sizes and large DIF magnitude conditions produced large statistical power above the 70% power threshold, ranging from 86% -100%, for the detection of the three DIF items. The statistical power for DIF detection was also relatively high when the sample sizes in the reference groups were 1000, 500 250 at the ratios of 1:.50 with large DIF magnitude, ranging from 73.70% - 100%. Acceptable statistical power for DIF detection, > 70%, with sample sizes in the ratios of 1:.10 was only observed when the sample size in the reference group was 1000 with large DIF magnitude (see figure 1). The same flagging rule set at .05 *p-value* for large DIF was also applied in the nine cells with moderate DIF magnitude. With moderate DIF the statistical power for DIF detection was high for 1000 sample size in the reference group at 1:1 and 1:.5 ratios. The SIBTEST procedure's statistical power for DIF detection with 500 and 250 sample sizes in the ratios of 1:.50 in the reference and focal groups respectively were relatively low, and even much lower (below the acceptable power level) in 1:.10 ratios with 250 sample sizes in the reference groups and moderate DIF magnitude. With large DIF magnitude and 1000 sample size at 1:.10 ratio, the statistical power for detecting the three DIF items ranged from 83% - 91.70% while with the same sample size ratios and moderate magnitude the statistical power ranged from 13% - 69.60% for the three DIF items.

It was expected that a statistically significant estimated beta-uni should not be less than .088, the size that Shepard, et al., (1985) equated with moderate DIF when item difficulty parameter was increased by .35. Results of the DIF analyses indicate that when the discrimination parameter was relatively large and the b-parameter was high, as was the case

with item 46, the estimated beta-uni tended to be smaller in value and the item was less frequently flagged for DIF than were the other two DIF items. The results of this study also show that with small and unequal sample sizes in the focal group, the size of estimated beta-uni varied inconsistently and the statistical power for detecting DIF items was smaller than what was observed in the cases with larger unequal sample sizes. Table 8 shows the estimated beta-uni for the studied items and their calculated standard errors. The standard errors were larger when sample sizes were moderate (N=500) and small (N=250) and the sample size combination ratios were 1:.10. Inflated standard errors were observed with small sample sizes in the ratios of 1:.10.

Table 8

Average DIF Effect Sizes for M-H and SIBTEST Procedures with Large and Moderate DIF

	Sample Ratios											
	1:1				1:.5				1:.1			
Sample Size	Large DIF		Moderate DIF		Large DIF		Moderate DIF		Large DIF		Moderate DIF	
	B-uni	SE	B-uni	SE	B-uni	SE	B-uni	SE	B-uni	SE	B-uni	SE
1000	.05	.00	.02	.00	.04	.00	.03	.00	.05	.00	.05	.00
	.05	.00	.02	.00	.04	.00	.03	.00	.06	.00	.05	.00
	.04	.00	.02	.00	.03	.00	.02	.00	.03	.00	.04	.00
500	.03	.00	.03	.00	.04	.00	.04	.00	.04	.02	.09	.02
	.03	.00	.03	.00	.04	.00	.03	.00	.06	.02	.08	.02
	.02	.00	.02	.00	.02	.00	.02	.00	.04	.02	.06	.02
250	.04	.00	.04	.00	.06	.00	.06	.00	.14	.06	.06	.06
	.04	.00	.05	.00	.05	.00	.06	.00	.17	.06	.06	.06
	.03	.00	.04	.00	.04	.00	.04	.00	.11	.06	.06	.06

The estimated beta-uni calculations were not carried out in some replications in the case of 250/25 sample sizes in the reference and focal groups respectively. This meant that the proportions of valid subtests were much smaller in the cases where the DIF statistics were not calculated. The minimum number of examinees had to be two in each stratum (i.e., observed score on matching subtest) to calculate the DIF statistics. Therefore in any stratum where there were less than two examinees at the ability level under study, DIF statistics was not calculated. It is possible that the small proportions of valid strata may have contributed to low statistical power for DIF detection and also resulted in the inconsistency of the SIBTEST procedure in calculating the DIF statistics. The expanded version of the SIBTEST output indicated that some examinees were eliminated from the cells if the number in the stratum

was less than two per the default option of the SIBTEST procedure (see Appendix 1, for example).

It was noted that if the minimum number of examinees required in each stratum to calculate the DIF statistics was increased, even fewer strata would be used and the DIF statistics would be calculated even less often. To reduce the minimum number of examinees required in each stratum at a specified ability level to one, would mean basing the calculation of strata means to one person, and that would not be helpful either. Stout and Roussos (1994) suggested that elimination of the cells from calculation is done to maximize statistical power of the procedure for detecting DIF items. The proportions of strata not used in the DIF analysis could not be observed from the results of the default operations of the SIBTEST procedure.

Therefore one replication was conducted with an expanded output to observe the proportions of the strata not used in the calculation of the estimated beta-uni. It was evident from the expanded output that calculating DIF estimated beta-uni on a small percent of strata leads to an unstable statistic. Flagging of non DIF items for DIF were observed with large magnitude DIF in all the nine cells. The Type I error rate at .05 was more inflated with small sample sizes at 1:.10 ratios with both large and moderate DIF magnitudes (see Table 9).

Table 9

Averages of Type I Error Rates and DIF Effect Size Standard Errors for 47 Non DIF Items with Large and Moderate DIF in the Test

Sample Size	**1:1**				**1:.50**				**1:.10**			
	Large DIF	B-uni	Mod DIF	B-uni	Large DIF	B-uni	Mod DIF	B-uni	Large DIF	B- uni	Mod DIF	B-uni
1000	2.64	.03	1.41	.02	3.26	.03	2.62	.03	4.94	.06	4.79	.05
500	1.68	.03	1.15	.03	2.50	.04	2.13	.04	**8.22**	**.**08	**7.81**	**.12**
250	1.41	.04	1.23	.08	2.85	.07	2.17	.05	**11.75**	.15	**12.54**	**.11**

Note: Bolded values in Table 9 represent Type I error rates > 5%.

Analysis of the Type I error showed that SIBTEST demonstrated better control of the Type I error with equal sample sizes in the focal as in the reference group. The procedure tended to result in inflated Type I error rates when sample sizes were unequal regardless of the sample size. Poorer control of the Type I error was noticed with all the sample sizes in the ratios of 1:.10 under study with moderate DIF magnitude. The average DIF effect size for the SIBTEST procedure increased with decreasing sample size and with increased gap between the sample sizes in the reference and focal groups. Table 9 shows that the value of the beta-uni was larger when the sample ratio combination was 1:.10 for all the sample sizes.

Discussion

Key to this study was the determination of statistical power of the SIBTEST procedure for DIF detection when sample sizes were unequal. In line with Fidalgo et al., (2004), the results of this study showed that the probability of falsely rejecting the null hypothesis of no DIF and hence committing a Type II error is functions of sample size in the reference group, sample size ratio combinations and DIF magnitude. With large sample sizes and large differences between the reference and focal group samples, Type I error inflation rates were observed as had been reported in some DIF detection literature (e.g., Bolt & Gierl, 2004). The results of this study confirmed that large differences in sample size between the reference and focal group examinees diminish the efficiency of the SIBTEST procedure for Type I error rates control. Where sample sizes were equal, the results were consistent with those of the previous studies that had been conducted with equal sample sizes (e.g., Roussos and Stout, 1996) which showed that the SIBTEST procedure was efficient when sample sizes were equal, and (Zheng, Gierl, & Cui, 2007) who reported that SIBTEST flagged more items for DIF than did M-H and Logistic Regression procedures. Zheng, et al., used real data in their study; therefore they did not know the DIF items a priori. It could be argued, on the basis of the current Monte Carlo study results that some of the items the procedure flagged for DIF in the Zheng et al., study were most probably non DIF items, resulting in inflated Type I error rates.

Besides the intended effects in the study, the selection of studied items also demonstrated a notable effect. Items selected had varied difficulty levels. One item was easy (b =-.97; a = .84). The other item had average difficulty (b =.06; a = 1.10) while the last item had a fairly high difficulty level (b-parameter =1.07; a = 1.09). Candell and Hulin (1986) suggested that items to be manipulated for DIF should be selected from close to the mean ability as possible. It was very notable in this study that the item with difficulty level close to the mean ability was detected more accurately in almost every cell in the study design. The results observed in this study indicated that item 46 with fairly high b-parameter value (b = 1.07) was not detected for DIF in some iteration in many data cells of the study design, especially with smaller sample sizes. It is possible that there were many cases where none of the examinees in the focal group got the item right, possibly resulting in many strata where no DIF statistics were calculated. In line with the study by Jiang and Stout (1998) with SIBTEST procedure and a nonlinear regression correction, this study in which SIBTEST procedure with regression correction was used, demonstrated adequate statistical power for

DIF detection with large sample sizes (N = 1000) and large DIF magnitude (i.e., .65) even when the sample sizes were in the ratios of 1:.10. However, the Type I error rates were comparatively higher with sample sizes in the 1:.10 ratios. Since Jiang and Stout used large sample sizes in the ratios of 1:1, 1:.50 and 1:.30 only, the inflated Type I error rates (29.79%) that were observed with 1000 sample in the reference group at the ratio of 1:.10 in this study lacked adequate comparison. The statistical power rates were high in support of Jiang and Stout's suggestion that if DIF was truly present, the statistical power of the procedure would be considerably reduced.

The results of this study should be interpreted in line with the proposal of DeMars (2008) who stated that although she obtained some inflated Type I error rates with SIBTEST with regression correction with large sample sizes, the procedure improves the accuracy of the estimated effect size with large samples. DeMars stated that accurate effect size with larger samples reduces the numbers of non DIF items flagged for DIF while lower accuracy in estimating DIF effect size with small sample sizes results in fewer non DIF items flagged for DIF but the accurately flagged items have larger effect size than with large sample size, thus resulting in inflated Type I error rates. In the current study, with small sample sizes, large proportions of strata were not used in calculating the DIF statistics effect size. Consequently, cumulative effect of the inaccurately flagged items most probably were responsible for the large numbers of items that were flagged for DIF with 250/25 sample size. The results of this study support earlier DIF studies that also indicated that DIF analyses are unstable and unreliable when using small samples (e.g., Meyer, Yuynh, & Seaman, 2004; Parshall, & Miller, 1995; Puhan, Yu, & Dorans, 2007).

Conclusions, Limitations and Recommendations

This study confirmed further that differences in sample sizes affect the statistical power of the SIBTEST procedure for DIF detection. To ensure the fairness of the tests in assessment for all examinees from diverse groups of varying educational needs and ethnic backgrounds, the degree of variance in sample size between reference and focal groups should not be in the ratio smaller than 1:.5 if SIBTEST is to be used in the analysis of bias. Psychometricians who use empirical testing accommodation data to analyze statistical bias ought to insist on re-sampling from large samples in the reference group to reduce problems likely to result from unequal sample sizes if SIBTEST has to be used, regardless of criticisms that re-sampling results in loss of data.

This study is not without limitations. Item parameter generation and DIF manipulation were done at the beginning of the study and the same parameters were used for all the cells of the design and for all replications within each cell making the results of the study only generalizable to data with the same distributions as was in this study. Although using the same item parameters was intended to control for confounding variances, the effect of doing so on the results was not known. Therefore, the item parameter characteristics of the items flagged by SIBTEST procedure for DIF should be studied closely to determine if the items are DIF items or false positives before the items are branded as DIF items, especially when empirical data are used. Secondly, generating uni-dimensional tests was a limitation to the extent to which the results of this study may be generalized considering that many experts have established that most tests have secondary dimensions (Ackerman, 1992; Camilli & Shepard, 1994; Embretson & Reise, 2000; Kok, 1988; Lord, 1980; Oshima et al., 1997; Shealy & Stout, 1993;). The use of unidimensional IRT model with multidimensional test data violates the unidimensionality assumption and poses potentially serious threat to item and examinee ability parameter estimates.

A future study should use a multidimensional IRT model for test item generation. Another study should also be conducted to compare the accuracy of DIF detection with older SIBTEST and SIBTEST with regression correction when sample sizes are unequal in order to establish whether using a separate SIBTEST from the older version that produced the M-H results at the same time for comparison of DIF results is a worthwhile effort. Future Monte Carlo simulation studies should use larger sample sizes in the reference group at the same ratios used in this study with SIBTEST with regression correction to observe the Type I error rate behavior further. The relationship between item difficulty and discrimination parameters should be studied further with unequal samples to be able to explain why non DIF items with low discrimination and high difficulty or high discrimination and low difficulty get flagged for DIF, especially by SIBTEST procedure. Hambleton et al., (1991) and Lord (1980) noted that small sample sizes were not a problem with Rasch model. Future studies should use the Rasch model with unequal sample sizes to determine if using the model instead of the 2PLM for the estimation of item parameters would result in reduced Type I error rates when SIBTEST with regression is used for DIF analysis. With Monte Carlo studies, future DIF studies with the SIBTEST procedure should take a confirmatory approach, thus should not include the DIF items which are known a priori in the matching subtest to avoid the effect of contamination of the matching subtest.

REFERENCES

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.

Angoff, W. H. (1993). Perspectives of Differential Item Functioning methodology. In P W Holland & H. Wainer (Eds.), *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 3-23.

Bolt, D. M., & Gierl, M. J. (2004, April). *Application of a regression correction to three nonparametric test of DIF: Implications for global and local DIF detection*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased items*. Newbury Park, CA: Sage.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaptation-Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*, 169-199.

Chang, H., Mazzeo, J., & Roussos, L. (1993, April). Detecting DIF for polytomously scored items: An adaptation of Shealy-Stout's SIBTEST procedure. Paper presented at the annual Meeting of the American Educational Research Association, Atlanta GA.

DeMars, C. E. (2008). Modification of Mantel_Haenszel and Logistic Regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, XX*(X), xx-xx.

Fidalgo, Á. 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. *The Journal of Experimental Education, 73*(1), 23-39.

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

Fischer, G. H. (1993). Notes on the Mantel–Haenszel procedure and another chi-squared test for the assessment of DIF. *Methodika 7*, 88–100.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika 60*:449–487.

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.

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.

González-Romá, V., Tomás, I., Ferreres, D., & Hernández, A. (2005). Do items that measure self-perceived physical appearance function differentially across gender groups of adolescents? An application of the MACS model. *Structural Equation Modeling, 12*, 157-171.

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. (2007). 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.

Hanson, B. A. (1998). Uniform DIF defined by differences in Item Response Functions. *Journal of Educational and Behavioral Statistics. 23*(3) 244-253.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279-291.

Harwell, M., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics, 17*, 315-339.

Harwell, M., Stone, C. A., Hsu, T. C., Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125.

Hauck, W. W., & Anderson, S. (1984). A survey regarding the reporting of simulation studies. *The American Statistician, 38,* 214-216.

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.

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 2,* 101-118.

McAllister, P. H. (1993). Testing DIF and public policy. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 381-396.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Meyer, J. P.; Huynh, H., & Seaman, M. A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on attitude survey. *Journal of Educational Measurement, 41*(4), 331-344.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16,* 381-388.

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in statistical detection of measurement bias. *Applied Psychological Measurement, 16*, 389-402.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform bias. *Applied Psychological Measurement 20*(3), 257-274.

Naylor, T. H., Balintfy, J. L., Burdick, D. S., & Chu, K. (1968). *Computer Simulation Techniques*. New York: Wiley.

Pae, T. I. (2004). DIF for examinees with different academic backgrounds. *Language Testing 21*, 53-73.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small sample conditions. *Journal of Educational Measurement, 32*(3), 302-316.

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.

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.

Rosser, P. (1989). Gender and testing. ERIC ED 336457. Educational Resources Information Center.

Roussos, L. A., & Stout, W. (2004). Differential item functioning analysis. In D. Kaplan (Eds.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 107-116). Thousand Oaks, CA: Sage:

Roussos, L. A., & Stout, W. F. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

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

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22,* 77-105.

Spence, I. (1983). Monte Carlo simulation studies. *Applied Psychological Measurement, 7*, 405-425.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

Stone, C. A. (1993). The use of multiple replications in IRT based Monte Carlo research. Paper presented at the European Meeting of the Psychometric Society, Barcelona.

Stout, W., Li, H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement, 21*(3), 195-213.

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.

Wasti, S. A., Bergman, M. E., Glomb, T. M., & Drasgo, F. (2000). Test of the cross-cultural generalizability of a model of sexual harassment. *Journal of Applied Psychology, 85*, 766-778.

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.

Zheng, Y., Gierl, M. J., & Cui, Y. (2007, April). *Using real data to compare DIF detection and effect size measures among mantel-Haenszel, SIBTEST and Logistic Regression procedures.* A paper presented at the annual meeting of the National Council on Measurement in Education: Chicago, ILL.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1-28.

CHAPTER FIVE

Manuscript 2

A Comparison of SIBTEST and Mantel-Haenszel Procedures' Statistical Power for DIF Detection when Sample Sizes are Unequal

Abstract

This simulation study focused on determining the effect of unequal sample sizes on statistical power of SIBTEST and Mantel-Haenszel procedures for detection of DIF of moderate and large magnitudes. Item parameters were generated with 2PLM using WinGen2 (Han, 2006). MULTISIM was used to simulate ability estimates and to generate response data that were analyzed by SIBTEST. The current version of the SIBTEST with regression correction was used to calculate the magnitude and the significance of DIF for the SIBTEST procedure. The earlier version of SIBTEST was used to calculate the magnitude and the significance of the DIF for the M-H procedure. SAS provided the environment in which the ability parameters were simulated; response data generated and DIF analyses conducted. Test items were observed to determine if a priori manipulated items demonstrated DIF. The study results indicated that with unequal samples in any ratio, M-H had better error rate control than SIBTEST. The results also indicated that not only the ratios, but also the size of the sample and the magnitude of DIF influenced the behavior of SIBTEST and M-H with regard to error rate control. With small samples and moderate DIF magnitude, Type II error was committed by both M-H and SIBTEST when the sample ratio was 1:.10.

Key words: Differential item functioning, Mantel-Haenszel procedure, SIBTEST procedure, combination ratios, DIF magnitude, item parameter, matching variable

A variety of procedures for detecting possible item bias through differential item functioning (DIF) analysis has been developed for dichotomously scored items. The most widely used procedures for DIF detection have been classified in two major categories depending on: (1) the way in which the matching variable is obtained (i.e., observed score versus an estimate of the latent variable presumed to underlie test performance) and (2) whether an assumption is made about the form of relationship between an item score and the matching variable (i.e., parametric if a particular form for the item response function is assumed versus nonparametric if such assumption is not made (Potenza & Dorans, 1995). Under Potenza and Dorans' classification scheme, the Mantel-Haenszel (M-H) procedure is considered an observed-score/nonparametric method because the procedure does not assume

a particular form for an item response function (Ankenmann, et. al., 1999). The SIBTEST procedure (IRT; Thissen, Steinberg, and Weiner, 1993) is considered to be a latent-trait/parametric method because it assumes a particular item response function. SIBTEST detects bias by comparing the responses of examinees in the reference and focal groups that have been allocated to strata using their scores on a "matching subtest" (Stout & Roussos, 1996b). In a simulation study a matching subtest is a subset of items that are known to be unbiased.

This Monte Carlo study is aimed at comparing the statistical power of SIBTEST and M-H procedures for DIF detection with unequal sample sizes in the reference and focal groups with large and moderate DIF magnitudes. The reference group sample responds to the unbiased subset of items referred to as the "matching subtest". In the reference group no item has a bias associated with target ability. The focal group responds to the subset of items that contain the studied items that have been manipulated to demonstrate large or moderate DIF. Thus in the focal group some a priori known items have bias associated with the target ability.

Roussos and Stout (1996) compared the error rates of SIBTEST and M-H DIF detection procedures with equal sample sizes. They reported comparable results with SIBTEST and M-H with sample sizes of 200 in both the reference and focal groups. Studies that report which of the two DIF detection procedures has better error rate control when sample sizes are unequal are limited. Therefore the results of DIF analyses in this study were expected to inform psychometricians, especially testing accommodation researchers on which of the two DIF detection procedures would produce results which could be considered reliable with small numbers of examinees in the ratios of the reference to focal groups that are frequently encountered in testing accommodation research.

The Mantel-Haenszel Procedure

The M-H procedure (Mantel & Haenszel, 1959) is a DIF detection method that first identifies the reference and focal groups of examinees and then estimates the probability of a member of the reference or the focal group at a certain ability level getting an item correct. The reference group is expected to provide standard performance on the item of interest, and the focal group, whose differential performance, if any, is to be detected and measured. The M-H requires that these groups be matched according to relevant stratification using levels of ability. It therefore requires a 2 x 2 x k contingency table for each of the levels of the matching variable, where *k* is the total number of score levels on the matching variable,

namely the total test score from the responses to the studied items by the examinees of each group (Ryan, 1991). At each score level *j*, where *j*=0,1,2,…*k*, a 2-by-2 contingency table is created for each item as shown in Table 10.

Table 10

The 2x2 Contingency Table at the jth Score Level

	Score on Studied Item		
Group	1	0	Total
Reference	A_j	B_j	N_{Rj}
Focal	C_j	D_j	N_{Fj}
Total	NT_{1j}	NT_{0j}	NT^2_j

Table adopted from Narayanan and Swaminathan (1994)

In Table 10 A_j is the observed number of examinees in the reference group at score level $_j$ answering the item correctly; B_j is the observed number of examinees in the reference group at score level $_j$ providing a non keyed response to the item and N_{Rj} is the total number of examinees in the reference group at score level $_j$. In the same way, C_j is the observed number of examinees in the reference focal group at score level $_j$ answering the item correctly; D_j is observed number of examinees in the focal group at score level $_j$ answering the item wrong and N_{Fj} is the observed total number of examinees in the focal group at score level $_j$. Hence, NT_{1j} is the observed number of examinees in the reference and focal groups at score level j who provided the correct response to the item, while NT_{0j} is the observed total number in the two groups at score level $_j$ who provided a wrong response to the item and NT^2_j is the grand total of the examinees in the reference and focal groups at score level $_j$ (Narayanan & Swaminathan, 1994).

The M-H compares the probabilities of correct response in the focal and reference groups for examinees of the same ability and calculates two statistics, the effect size (α_{MH}) of the standardized mean difference and the statistical significance (*p-value*) of that difference (Dorans & Schmitt, 1991). The α_{MH} is the ratio of the odds that reference group examinees will get the item correct compared to the odds for a matched focal group examinee. With large sample sizes, small differences can be reported as statistically significant, thus leading to inflated Type I error. If α_{MH} is greater than 1, the studied item is considered to be favoring the reference group; on the other hand, α_{MH} less than 1 is taken to indicate that the studied

item is favoring the focal group. An estimate of the common odds ratio α which is the $\alpha_{MH,}$ that also provides the estimate of the DIF effect size, can be expressed as:

$$\alpha_{MH} = \frac{\sum A_j D_j / T_j}{\sum B_j C_j / T_j}$$

where T_j is the true score at level $_{j.}$

The M-H D-DIF statistic, a frequently used measure of DIF, is a rescaling of the natural log of an estimate of α-$_{MH}$, which is the M-H procedure's DIF statistics effect size. The α_{MH} statistic is often transformed on the delta scale used by the Educational Testing Service (ETS) to measure item difficulty via

$$\Delta_{MH} = - 2.35 \ln(\alpha_{MH}).$$

The α_{MH} is usually transformed to the Δ scale to enhance the interpretability of the result. In this study uniform DIF was manipulated in the three DIF items. Uniform DIF occurs when the difference in the probability of a correct response to an item between the two groups is constant across all ability levels.

According to the ETS system for categorizing the severity of DIF (Zieky, 1993), a value of $| \Delta\_ |$ MH ≥1.50 indicates substantial amount of DIF, therefore the item must be carefully reviewed (Fidalgo, Ferreres, & Muniz, 2004). The statistical significance of the difference in performance between the reference and the focal groups is estimated as:

$$\chi^2_{MH} = \frac{\left(\sum Aj - \sum E(Aj) - 0.50 \right)^2}{\sum \text{var}(Aj)}$$

where,

$$E\ (Aj\ ) = \frac{N_{rj}\ T_{lj}}{T_j}$$

and,

$$\text{var}(A_j) = \frac{N_{Rj} N_{Fj} T_{1j} T_{0j}}{T_j^2 (T\text{-}1)}$$

while, T is the true score

where A_j is the observed number of examinees in the reference group at score level j providing a correct response to the item under the null hypothesis:

$$\boldsymbol{H}_{0:}\ \pi_{Rj}\ (1\text{-}\pi_{Rj}) = \pi_{Fj}\ (1\text{-}\pi_{fj})$$

for all values π_{Rj} where T^2_j represents the total score on the test, π_{Fj} is the probability in the reference group that an examinee with a total score of j on the test will respond to the studied item correctly, and π_{Fj} is the probability in the focal group that an examinee with the total score of j will not respond to the studied item correctly, the M-H statistics has $\chi 2$ distribution with one degree freedom. Rejection of the null hypothesis would indicate that examinees in the reference and focal groups who are matching on overall ability differ in their mean performance on the studied item and thus shows DIF (Ryan, 1991).

The M-H procedure is the DIF detection method used by many practitioners. Clauser, Mazor and Hambleton (1991) observed that M-H is favorably comparable to latent trait procedures and that the users of the procedure claim that it takes much less time to calculate. With the desirable characteristics that have been accorded to M-H, this study aimed at determining which of the procedures had better control of the error rates with sample sizes in the ratios of 1:.50 and 1: .10 with the smaller sample size being in the focal group of simulated examinees.

The newer version of SIBTEST (Shealy & Stout, 1993) was used to compute a weighted mean difference between the reference and focal groups' probability of correct response to an item at a certain ability level. The procedure adjusts the means 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 and hence reduced Type I error rates (Jiang & Stout, 1998). The weighted mean difference between the reference and focal groups on the subtest item across the Θ subgroups is computed with

$$\hat{\beta}_{\text{UNI}} = \sum_{k=0}^{k} P_{kdk}$$

In this equation, P_k is the proportion of focal group examinees in subgroup k and $d_k = P_{Rk} - P_{Fk}$, which is the difference in the adjusted means of studied subtest item for the reference and focal groups respectively, in each subgroup k. The means on the studied subtest item are adjusted to correct for any differences in the ability distributions of the reference and focal groups using a regression correction (Shealy & Stout, 1993).

SIBTEST also yields an overall statistical test for DIF effect size. The test statistic for evaluating the null hypothesis is:

$$SIB = \frac{\beta_{UNI}}{\sigma(\beta_{UNI})}$$

When the examinees from the reference and the focal groups with the same ability in the latent trait of interest demonstrate a difference in their probability for a correct response to the keyed item then beta (SIBTEST DIF statistic) would be considered biased.

The SIBTEST procedure can be used to detect DIF at either the item or testlet level (a testlet is a bundle of items). The procedure conducts DIF analyses using original item response data rather than parameter estimates from a program, such as WinGen2. The emphasis in the improvement and evaluation of SIBTEST has been placed on the control of Type I error inflation and statistical power for detection of DIF. According to Shealy and Stout, when there is no target ability difference between the reference and focal group, it can be said that the test is not biased against or for a particular group.

Interpretation of Bias

The most popular methods for the analysis of bias are those based on item response theory (IRT) and chi-square distributions, most notably the M-H statistic (Mantel & Haenszel, 1959). However, several measurement researchers have reported that the use of these procedures has generally failed to produce meaningful interpretations of bias (Buhr, 1988; McPeek & Wild, 1986; Zwick & Ercikan, 1989). Critics of the DIF procedures (e.g., Hoover & Kolen, 1984) argue that the procedures lack significant reliability. The DIF detection procedures have also been faulted for statistical fluctuations that have been noticed in the item response data. Hoover and Kolen based their criticisms on the results of a study they conducted to investigate the reliability of several bias detection techniques across random samples from the same population. They used the Iowa Tests of Basic Skills (ITBS) for comparisons based on sex and ethnic groups. The results of their study showed negligible reliability among the indexes they used.

Although several approaches developed later for analysis of bias, including the IRT sum-of-squares and M-H methods, are said to have demonstrated improved utility over those used in the Hoover and Kolen (1984) study as was observed by Skaggs and Lissitz (1992), the performance of SIBTEST and M-H in terms of their error rate behavior when sample sizes are unequal need to be studied further.

SIBTEST and Mantel-Haenszel Procedures

In this study, the performance of SIBTEST was compared with that of M-H when sample sizes were unequal. The M-H has been accepted as a benchmark against which procedures for DIF detection are judged before adoption by measurement practitioners

(Holland and Weiner, 1993). It is on the basis of this available information that comparing the performance of M-H with that of SIBTEST on error rate control when sample sizes are unequal between the reference and the focal groups was considered. A study by Whitmore and Schumacker (1999) reported that at some test score levels, even the later versions of M-H procedure could have incomplete contingency table cells that could not be used, resulting in the loss of data and subsequent reduction in M-H reliability. Another reason for comparing the performance of M-H and SIBTEST is that M-H treats continuous data like discrete data, and it does not require local item independence that latent trait procedures for DIF detection require (Holland &Thayer, 1988).

In this simulation study, the performance of SIBTEST was compared with that of M-H in terms of their error rates in detecting uniform DIF of moderate and large magnitudes with unequal sample sizes across groups of simulated examinees. The aim of the study was to determine if the effect of unequal sample sizes was dependent on sample sizes, sample combination ratios, as well as on the procedure for DIF detection. With SIBTEST, if the null hypothesis of no DIF is rejected, members of the reference and focal groups with the same underlying trait differ in their probability of a correct response to that item on the test; hence, the item is identified as exhibiting DIF. On the other hand, with M-H an item is considered as exhibiting DIF when the members of the reference group that are identical in overall ability differ in their mean performance on the studied item.

Simulated data has been used to assess the quality of DIF detection procedures in terms of Type I error rate (proportion of non DIF items falsely flagged for DIF) and Type II error rates (proportion of DIF items not identified). Monte Carlo studies allow researchers to determine with certainty which DIF analyses techniques are better than others and in what conditions as parameters that determine the simulations (e.g., DIF magnitude, item parameters, ability distribution) are known (Fidalgo, Ferreres, & Muniz, 2004). In this study the error rate control of SIBTEST and M-H were compared when sample sizes were unequal.

Several studies recommend the application of multiple procedures of DIF detection to be fairly certain that the item identified with DIF is a correct detection and not a false positive. The use of multiple DIF detection procedures also increases the level of certainty on how many items with DIF have failed to be indentified (Hambleton & Jones, 1994; Kim & Cohen, 1998; Shealy & Stout, 1993). Kim and Cohen pointed out that given the available knowledge, “there seems little to justify the use of a single statistics in any DIF study for DIF detection” (p. 310). However, it is of significant importance that the selection of the procedures should be well informed based on the sample sizes under consideration.

The logic Kim and Cohen (1998) provided for use of multiple DIF statistics was that the more the agreement between procedures of DIF detection, the greater the certainty that the items detected were items that certainly functioned differentially, thus justifying that neither Type 1 nor Type II error was likely to be committed (Fidalgo, et al., 2004). The utility of SIBTEST and M-H has been explored in previous studies (Fidalgo, 2004; Holland & Thayer, 1988). These studies compared the two DIF detection procedures using Rasch model with equal sample sizes.

Although it has been argued that M-H DIF statistics takes less time to calculate, as compared to the latent trait DIF analysis procedures, Mazor, Clauser and Hambleton(1992) observed that with sample sizes of 500, the M-H detection rate was only 50% for the item that was known to have DIF. It is therefore necessary to identify which procedure would be better to use with varying sample sizes in the focal group, especially small sample sizes as is often the case in the test results data where testing accommodation is provided. It has been indicated that small sample sizes are not problematic with Rasch model (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980) but it is not yet clear how unequal sample sizes would affect the performance of M-H and SIBTEST in their error rate behavior when items are generated with 2PLM.

Apparently, Roussos and Stout (1996) proposed a solution to the problem of empty contingency cells with the Chi-Square ($\chi 2$) bias analysis procedures noted by Whitmore and Schumacker (1999). Roussos and Stout stated that different suspect items have their score cells weighted differently. They also said that estimated delta gives zero weight to cells for which either no reference group examinees got the items wrong or no focal group examinees got the items right. Nevertheless, because M-H does not require local item independence that latent trait procedures for DIF detection require, it was still important to compare the DIF detection results of M-H with that of SIBTEST (Holland & Thayer, 1988). This study used simulated data to manipulate sample sizes in the focal group to determine the effect of unequal sample sizes on DIF detection with SIBTEST and M-H procedures.

Method

Simulation

Item parameters for 50 items were generated using a 2-PL unidimensional model in WinGen2 (Han & Hambleton, 2007). The item difficulty parameters were generated with a normal distribution N (0, 1) while the discrimination parameters were generated to be

lognormal, N (0,.2). The generated item parameters used in this study were presented in Table 11.

Table 11

Generated Item Parameters

Item no.	Item Parameter		Item no.	Item Parameter	
	a	b		a	b
1	1.10	.06	26	.73	.12
2	.83	-1.80	27	.91	-1.47
3	.79	.22	28	.97	-.74
4	1.25	.23	29	1.27	-.82
5	.84	.54	30	1.15	-.03
6	.86	1.14	31	1.13	-.02
7	1.11	-.62	32	.83	.90
8	.90	.46	33	1.15	.19
9	1.07	.62	34	.91	.78
10	1.44	-.43	35	1.27	-.81
11	1.31	1.60	36	.90	-.56
12	.90	.48	37	1.31	-2.72
13	.84	-.97	38	.80	-1.07
14	.81	-.21	39	.61	-.03
15	.93	-.61	40	1.13	-.34
16	.87	.19	41	.79	.03
17	1.05	.31	42	.86	.80
18	.92	.24	43	.97	.17
19	.75	-.17	44	.86	-.92
20	1.10	.49	45	1.18	-.39
21	.96	-.84	46	1.09	1.07
22	1.50	1.55	47	1.13	.20
23	.97	1.20	48	1.08	.48
24	.77	1.41	49	1.20	-.25
25	1.00	-.71	50	.80	-.20

The ability (θ) parameters and item response data were generated with MULTISIM. The ability values were randomly selected from normal theta distribution N (0, 1). The 2-PLM was selected to determine the effect of unequal sample sizes on the M-H and SIBTEST procedures' error rate behavior when samples are small considering that Lord (1980) and Hambleton, Swaminathan, and Rogers (1991) reported that small sample sizes were not problematic with Rasch model. The same item parameters were generated for the reference as for the focal group. The parameters were simulated once at the beginning of the study. The item difficulty, b-parameters for three items, 1, 13, and 46, in the focal group set of items were altered by adding .35 and .65 to the b-parameter so that the items would display DIF of

moderate and large magnitudes respectively, before the data were replicated 1,000 times (e.g., Miller & Oshima, 1992; Shepard, Camilli, & Williams, 1985). The items that were manipulated to demonstrate DIF were chosen because they had varied item difficulty estimates ranging from fairly low to fairly high difficulty, to allow for generalization of the results to the whole range of a test. The item difficulty estimates ranged from -2.72 to 1.60 with a mean of -.09 and a standard deviation of .70. The discrimination parameter estimates ranged from .61 to 1.50, with a mean of 1.16 and standard deviation of .70. The abilities were sampled in every cell by iteration combination rather than using the same abilities across the cells for a single iteration.

Study Design

A design of (3 X 3 X 2) was chosen with three different sample sizes, three different ratios, and two varying DIF magnitudes as shown in Table 12. The same conditions used with M-H were applied to the SIBTEST procedure to allow for meaningful comparison between the two DIF detection methods.

Table 12

Manipulation of Sample Size and DIF Magnitudes

No. of Simulated Examinees in ref. group	Combination Ratios			Medium DIF Magnitude	Large DIF Magnitude
	1:1	1:.50	1:.10	.35	.65
1000	1000/1000	1000/500	1000/100		
500	500/500	500/250	500/50		
250	250/250	250/125	250/25		

The number of simulated examinees was varied to study the effect of sample size when the reference and focal groups were equally large (e.g., when $N_1 = N_2$) versus when focal group sample size was small using the sample sizes and sample size ratios shown in Table 12. The ratios were 1.00, .50 and .10 and reference group sample sizes were 1000, 500 and 250.

Manipulation of Bias

The magnitude of bias was specified as in Shepard, et al., (1985) study where an increase of .35 in the b-parameter was considered to constitute moderate DIF. An increase in the b-parameter of above .64 had been specified by the Educational Testing Services as constituting large DIF. Thus b-parameters of the studied items that were to demonstrate DIF

of large magnitude were increased by .65. The same differences in the b-parameter that were used with M-H were also used with SIBTEST. Again this was done to allow for comparison between the DIF detection error rates of SIBTEST and M-H procedures.

DIF was introduced into the *b*-parameters in only one direction to make the study comparable with previous DIF studies on item bias (e.g., Shepard et al., 1985). Unidirectional DIF is specifically important in a comparative study with varying sample sizes. It has been noted that cumulative effect of item bias on the test response functions for the subpopulations can be nonexistent when different items differ only in the a-parameters (Stout & Shealy, 1991). Consequently, the consistency in the direction of DIF modeled in this study presupposed that the response function was also expected to differ by the different ratio combinations of sample size in the reference and focal group (Miller & Oshima, 1992).

This study aimed at comparing M-H and SIBTEST to determine which one of the two procedures had better Type 1 error rate control and sufficient statistical power for detection of moderate and large DIF magnitudes when sample sizes were unequal. It was considered that unequal sample sizes in varying ratios portray the situations normally observed by psychometricians engaged in DIF analyses using empirical data. Although DIF detection with equal sample sizes has been well documented, the results with equal sample sizes, 1000/1000; 500/500 and 250/250 were also displayed in Tables 13 - 17 to serve as benchmark for interpreting the results with other sample size ratio combinations. A critical *p*-value was set at .05 for determining significant proportions of flagging that indicated a demonstration of DIF.

Results

The results of the 1000 sample size in the reference group with large DIF magnitude indicated that both SIBTEST and M-H had sufficient statistical power to detect the three DIF items, 1, 13, and 46 as indicated in Tables 13 and 14. The results also indicated that M-H had better Type I error control than SIBTEST as shown in Table 15. The items that had fairly high discrimination and low ability parameters were more often falsely flagged for DIF by SIBTEST procedure than by the M-H procedure. The flagging of those items by SIBTEST and not by M-H tended to suggest that SIBTEST was more sensitive to the differences between discrimination and difficulty parameters as had been noted by Finch and French (2007), who had reported from the results of their study that highly discriminating items were more vulnerable to false flagging for DIF. When correlation analyses were run to determine if there were any relationships between item difficulty and item discrimination, no statistically

significant correlations were observed, suggesting that if the item parameters influenced the detection rate or Type I error rate, then they did so without interacting with each other.

It was evident from the results with 1000 sample size that equal samples in both the reference and focal groups demonstrated adequate statistical power for 100% detection rate of the three DIF items by both M-H and SIBTEST when the DIF magnitude was large ($b_{iF} = b_{iR} + .65$). DIF items with fairly high difficulty were flagged less times for DIF compared to items that had difficulty close to the mean. Bennett, Rock and Novatkoski (1989) proposed that when items suspected to be DIF items or items that are known a priori to be DIF items show no evidence of differential functioning then it could be deduced that perhaps there are other factors besides the hypothesized ones contributing to differential behavior of the test item across groups.

The statistical power for detection in 1:.50 ratio combinations with large DIF magnitude in the 1000 sample size category was comparable to the 1:1 ratio except that more non DIF items were flagged for DIF when the SIBTEST procedure was used. The detection rate was consistently high with 1000 sample size even when the ratio of the reference group sample size to focal group was 1:.10. The results displayed in Table 12 show that the rate of correct detection ranged between 83.00 to 100 % with large sample size (N=1000) and large DIF magnitude ($b_{iF} = b_{iR} + .65$) for the two procedures, with the lowest detection rate being recorded in the 1:.10 sample size ratio by the SIBTEST procedure.

The proportions of non DIF items flagged for DIF at all levels of the sample size ratios when the reference group sample was large (N=1000) and DIF magnitude was large ($b_{iF} = b_{iR} + .65$) differed depending on whether M-H or SIBTEST had been employed in the detection procedure. At the ratio of 1:.50 M-H had 5/47 items falsely flagged for DIF, (10.64% Type I error rates at .05) compared to 9/47 (19.15% Type I error rates at .05) for SIBTEST. When the sample ratio was 1:.10 the M-H maintained the same error rate (5/47 = 10.64% Type I error rates at .05) while SIBTEST had the increased (14/47 =29.79% Type I error rates at .05). The SIBTEST procedure also falsely flagged items with *b* and *a*-parameters that were in close resemblance to each other for DIF, hence, the inflated Type I error rate at .05. The results of the analyses when sample size and DIF magnitude were large showed that SIBTEST tended to show a poorer control of Type I error than M-H procedure with all sample size ratios.

With moderate sample size (N=500) and large DIF magnitude at the sample size ratio combination of 1:1 neither M-H nor SIBTEST procedure had any Type I errors at .05. However, with 500 sample size in the reference group at the ratio of 1:.5, some non DIF

items were falsely flagged for DIF. At the same sample ratio M-H showed that 4% of the non DIF items had been erroneously flagged for DIF while SIBTEST procedure had 8.5% of non DIF items flagged in the same sample size category. The non DIF items identified as false positives tended to show some similarities in their parameters. The discrimination parameters of the items were fairly high while the difficulty parameters were fairly far from the mean of zero. Both the M-H and SIBTEST had sufficient statistical power for 100% detection rate of the DIF items when sample size was 500 in the 1:1 ratio in the reference and focal groups respectively, with large DIF magnitude.

In the 1:.10 ratios M-H had strata within some of the replications that were not used in the calculation of the DIF statistics. In spite of the DIF statistics that could not be calculated the DIF procedure still demonstrated a fairly good control of the Type I error rate at .05. Only 2/47 (4%) of the non DIF items had been flagged for DIF. The SIBTEST on the other hand had a Type I error rate >.05 in 33/47 (70%) of the non DIF items at 1:.10 ratios when the reference group sample size was 500. No missing data were displayed by SIBTEST in this sample size and DIF magnitude category. As was expected, power for DIF detection decreased as sample size in the focal group decreased as evidenced by the results presented in Tables 14 and 15 as well as in figures 2 and 3. The results support González-Romá et al., (2006) who stated that as sample sizes increase, so do the DIF detection rates. With large sample size (N= 1000) M-H had sufficient power level for detecting all the three DIF items with large DIF With moderate DIF the procedure did not attain acceptable power levels for DIF detection except for item 1 (see Table 13). With small sample size (N= 250) M-H only had sufficient power for DIF detection at 1:1 and 1:.50 combination ratio with large DIF as shown in Table 13. With moderate DIF the procedure did not attain acceptable level of statistical power for DIF detection with small sample size at all sample size ratio combinations (see Figure 2).

Table 13
Statistical Power Results with M-H

Sample Size	Item Parameters		Ratios					
	a-parameter	b-limit	1:1		1:.50		1:.10	
			Large DIF	Mod. DIF	Large DIF	Mod. DIF	Large DIF	Mod. DIF
1000	1.10	.06	100.00	100.00	100.00	98.90	90.00	78.80
	.84	-.97	100.00	100.00	100.00	99.70	83.00	15.40
	1.09	1.07	100.00	99.90	100.00	96.60	96.30	60.90
500	1.10	.06	100.00	95.80	100.00	86.80	81.70	20.60
	.84	-.97	100.00	97.30	100.00	51.30	35.30	1.60
	1.09	1.07	99.60	83.00	98.90	88.40	55.40	.20
250	1.10	.06	100.00	66.70	98.90	59.60	84.80	52.60
	.84	-.97	100.00	60.10	95.90	24.20	10.50	.50
	1.09	1.07	93.10	38.00	84.50	33.90	.80	.20

Table 14

Statistical Power Results for SIBTEST with Regression Correction

Sample Size	Item Parameters		Ratios					
	a-parameter	b-parameter	1:1		1:.50		1:.10	
			Large DIF	Mod. DIF	Large DIF	Mod. DIF	Large DIF	Mod. DIF
1000	1.10	.06	100.00	100.00	100.00	96.80	90.00	69.60
	.84	-.97	100.00	100.00	100.00	99.70	83.00	13.70
	1.09	1.07	99.80	97.10	99.90	96.60	91.70	61.90
500	1.10	.06	100.00	94.80	100.00	83.30	67.80	28.10
	.84	-.97	100.00	96.90	99.90	47.80	22.10	4.70
	1.09	1.07	99.60	62.40	98.90	79.60	51.40	4.30
250	1.10	.06	100.00	65.20	97.80	55.80	59.80	44.20
	.84	-.97	100.00	65.10	91.10	24.10	11.10	10.10
	1.09	1.07	86.00	28.10	73.70	35.10	16.90	10.40

Both the M-H and SIBTEST had sufficient statistical power (> 70%) for detecting DIF of large magnitude with 250 sample sizes in the ratios of 1:1 and 1:.50 (see Table 13 and Table 14). Table 15 shows the DIF effect size for the SIBTEST and M-H procedures.

Table 15

Average DIF Effect Sizes for M-H and SIBTEST Procedures with Large and Moderate DIF

	Sample Ratios											
	1:1				1:.5				1:.1			
Sample Size	Large DIF		Moderate DIF		Large DIF		Moderate DIF		Large DIF		Moderate DIF	
	D-DIF	B-uni	D-DIF	B-uni	D-DIF	B-uni	D-DIF	B-uni	D-DIF	B-uni	D-DIF	B-uni
1000	3.67	.05	1.30	.02	1.37	.04	1.22	.03	2.95	.05	1.75	.05
	2. 89	.05	1.38	.02	1.08	.04	1.34	.03	2.11	.06	.88	.05
	2.59	.04	1.33	.02	1.75	.03	1.36	.02	3.10	.03	1.95	.04
500	2.43	.03	1.31	.03	2.56	.04	1.37	.04	1.57	.04	1.47	.09
	2.67	.03	1.42	.03	2.40	.04	1.08	.03	1.40	.06	.49	.08
	2.37	.02	1.28	.02	2.82	.02	1.75	.02	1.68	.04	2.15	.06
250	2.53	.04	1.30	.04	2.77	.06	1.57	.06	5.25	.14	1.86	.06
	2.80	.04	1.27	.05	2.54	.05	1.27	.06	1.88	.17	1.45	.06
	2.41	.03	-.01	.04	2.80	.04	1.67	.05	1.98	.11	1.38	.06

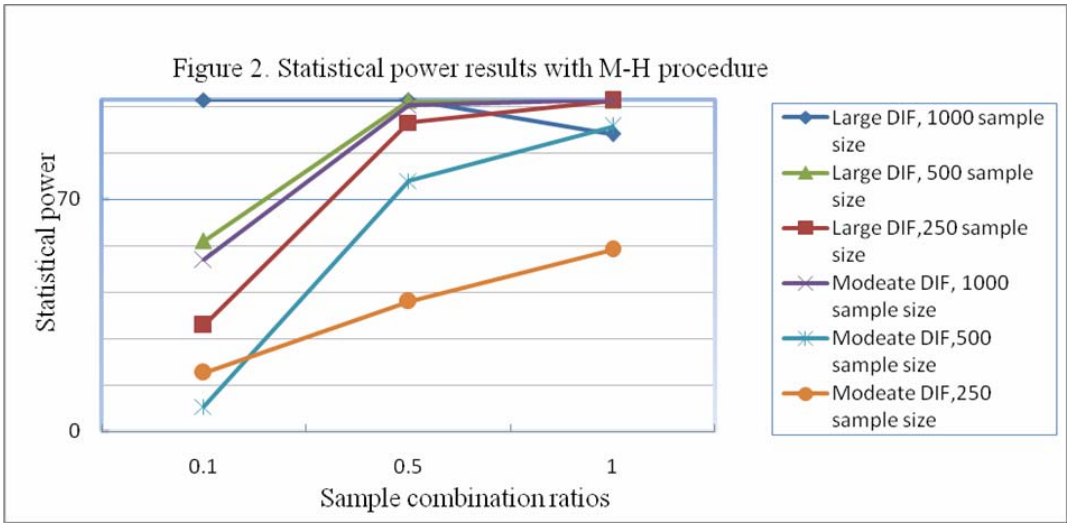
The DIF effect size was largest when the sample ratio combinations were 1:.10 for all sample sizes for both M-H and SIBTEST, hence the inflated Type 1 error rates and reduced statistical power for accurate DIF detection when the sample ratios were 1:.10. SIBTEST, however, flagged 10.60% of the non DIF items for DIF in the 1:.50 sample size ratios indicating a poorer Type I error rate control than M-H procedure that did not flag any non DIF items at the same sample size ratio combination. The SIBTEST procedure had even a higher Type I error rate at .05 and reduced power when the sample size ratios were 1:.10. The M-H on the other hand had lower statistical power for accurate detection of items 13 and 46 that were DIF items, but maintained adequate control of its Type I error rate at less than .05. Average of the statistical power rates for each of the procedures were calculated at each sample size, sample size ratio combinations and DIF magnitude levels. Figures 2 shows the M-H average power rates for detecting the three DIF items, while Figure 3 presents the average power rates for detecting the DIF items by the SIBTEST with regression correction procedure. The average DIF effect size for the DIF items were calculated for each cell as shown in Table 15. The standard errors of the DIF effect size for the studied items are shown in Table 16.

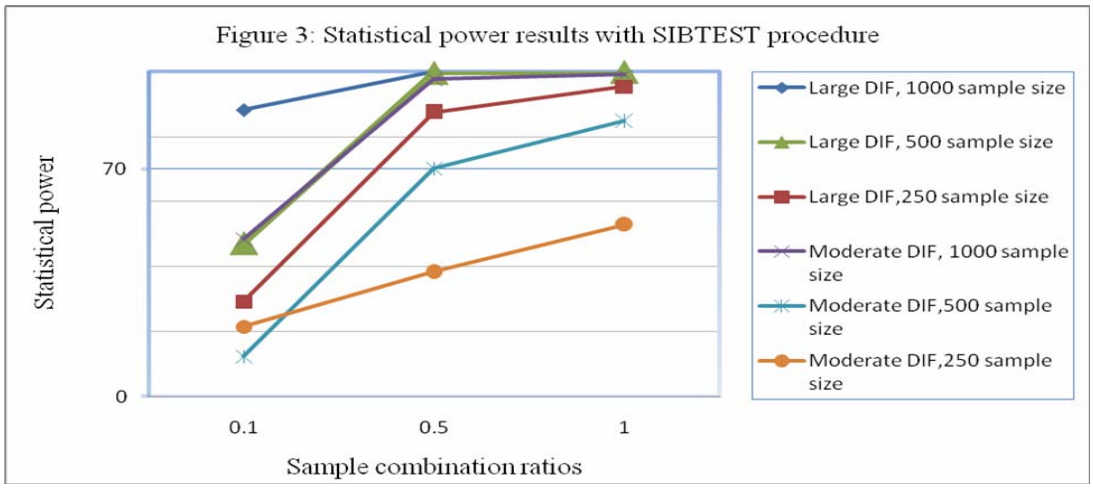
Table 16

Average DIF Effect Size Standard Errors for M-H and SIBTEST Procedures with Large and Moderate DIF

	Sample Ratios											
	1:1				1:.5				1:.1			
Sample Size	Large DIF		Moderate DIF		Large DIF		Moderate DIF		Large DIF		Moderate DIF	
	D-DIF SE	B-uni SE	D-DIF SE	B-uni SE	D-DIF SE	B-uni SE	D-DIF SE	B-uni SE	D-DIF SE	B-uni SE	D-DIF SE	B-uni SE
1000	.24	.00	.20	.00	.35	.00	.25	.00	.53	.00	.50	.00
	.28	.00	.20	.00	.38	.00	.25	.00	.48	.00	.49	.00
	.32	.00	.25	.00	.46	.00	.31	.00	.80	.00	.66	.01
500	.30	.00	.30	.00	.37	.00	.35	.00	.75	.00	.74	.02
	.30	.00	.31	.00	.35	.00	.35	.00	.73	.00	.77	.02
	.41	.00	.36	.00	.53	.00	.38	.00	1.27	.00	1.04	.02
250	.45	.04	.43	.00	.58	.00	.54	.00	2.96	.06	1.48	No SE
	.48	.04	.53	.00	.59	.00	.60	.00	1.24	.06	1.40	.06
	.61	.03	.73	.00	.83	.00	.72	.00	1.50	.06	No SE	No SE

As shown in Table 16 the standard errors for the DIF effect sizes tended to increase with a decrease in sample size. Item 46 which was harder had a larger standard error for all sample sizes in all sample combination ratios as indicated by the D-DIF standard errors. The beta-uni standard errors were zero except for the medium and small sample sizes in the ratios of 1:.10. The SIBTEST procedure did not show much difference in the standard errors among the studied items, as shown in Table 16. Some standard errors were not calculated when sample size was small at the ratio of 1:.10 as shown in Table 16.





As has been afore mentioned, the two procedures were also compared when DIF magnitude was moderate ($b_{iF} = b_{iR} + .35$). The results from the analyses with large sample size (N=1000) with moderate DIF showed that M-H and SIBTEST procedures had sufficient statistical power (≥ 70%) to detect DIF of moderate magnitude when the sample ratios were 1:1 and 1:.50. Lower detection rates of the three DIF items were observed for SIBTEST than for M-H when the sample size in the reference group was moderate (N= 500) with moderate DIF magnitude, indicating that with smaller sample sizes the SIBTEST procedure's power for detecting DIF items was reduced more than was the case for the M-H procedure. Again the observed results from the analyses were in support of González-Romá et al., (2006), as sample size decreased, so did the power for DIF detection (see Figures 2 and 3). Both M-H and SIBTEST procedures committed Type II error by failing to flag items 13 and 46 for bias more than 5% of the time when sample sizes were in the ratios of 1:.10.

The M-H and SIBTEST procedures demonstrated comparable Type I error control with moderate DIF and large sample size (N=1000) to the large DIF and medium sample size, as was indicated in Figures 2 and 3. The M-H procedure had no items with Type I error rates at .05 with samples in the ratio of 1:1 in the 1000 sample size category with large and medium DIF magnitudes. With SIBTEST two non DIF items out of 47 were falsely flagged for large DIF at .05 and non for moderate DIF with same sample size and sample size ratios. With the M-H and SIBTEST procedures the Type I error rate at .05 increased with increased range in sample sizes between the reference and the focal groups. There were fewer Type I errors at .05 with 1:1 ratio than with 1:.50 ratios and there were more of the errors at 1:.10 than 1:.50 ratios with large and moderate DIF magnitudes with 1000 and 500 sample sizes in the reference group. Type I error rates at .05 reported with 250 sample size in the reference group at the ratio of 1:.10 was low for M-H and very high (> .05) for SIBTEST (see Table 14). The Type I error rates at .05 did not consistently decrease with a decrease in sample size ratio for M-H.

A closer study of the expanded output of some cells with small sample sizes in the focal groups for both the M-H and SIBTEST procedures showed that a minimum of two simulated examinees were required in each stratum for the DIF statistics to be calculated. When sample sizes in the focal group was small as were the cases with sample sizes in the ratios of 1:.10 several of the contingency table cells were either incomplete or empty. The empty contingency table cells or strata were represented by zero as was observed when frequency analyses were run. These empty cells represented by zeros in the strata included cases where either 100% of the examinees in reference group at the ability level got the item right or no one in the focal group got the item right at the same ability level. The contingency cells that had one simulated examinee at the ability level were represented by dots as were seen in the frequency analyses output. All the empty and incomplete strata were not regarded as valid cases. The SIBTEST and M-H DIF analyses output show that several cells which had 1 or 0 had only small proportions of valid data that accounted for the DIF statistics, the effect size and the *p-value* of the DIF effect size. It was noted that M-H was stable even when the proportions of valid data used in the analyses were small. However, SIBTEST became unstable with small sample sizes that led to even fewer strata with valid data in which DIF statistics were calculated. Consequently, inflated Type I error rates at .05 were observed with SIBTEST.

Stout and Roussos (1995) suggested that inflated Type I error rates observed with small sample sizes in the focal groups are due to erroneous estimates of the true valid subtest

score (T) with systematic errors being different for the reference and the focal group of examinees. Regression correction was meant to correct for the systematic errors of beta estimates in the trait the test is designed to measure. In this study SIBTEST with regression correction was used. With expanded SIBTEST it was observed that with small sample sizes in the focal group as was the case in the 1:.10 ratios, very small proportions of the examinees with the same valid subtest score formed valid subtests that were included in the calculation of the DIF statistics. Stout and Roussos recommended that instead of taking weighted beta estimates obtained by taking a weighted sum of the DIF effect sizes across the strata or calculation cells, the reference and focal group examinees should be pooled and the proportion of the pooled group should be used to solve the problem of possible inflated Type I error rates. The same procedure as proposed by Stout and Roussos was used in this study. One weighting obtained by pooling the two groups was examined through an expanded DIF analysis. The expanded analysis showed that with small sample sizes in the ratios of 1:.10 valid cases that were used in the calculation of the DIF statistics were very few.

The authors of this paper consider that ability differences is not accountable for the inflated Type I error rates observed at .05. It is possible that when sample sizes are small and are in the ratios of 1:.10 then SIBTEST with regression correction behaves in the same way it would behave if estimated betas in the trait the test was designed to measure were biased (see Table 17. It is apparent that regression correction mainly corrects for ability estimates and not other inequalities across the group samples. The power for falsely detecting the item for DIF tended to be influenced more by the interaction of the item with sample size than with the DIF magnitude. Considering that the items had different item response functions (IRFs), it is possible that there were interactions between the IRFs and sample sizes and sample size ratios. It is also possible that the fit of data to the 2PLM was not perfect. The effect of using a unidimensional model with a multidimensional test could be responsible for the high levels of flagging with medium sample size. It is also possible that the matching subtests were contaminated by the DIF items, hence the inflated Type I error rates, particularly by the SIBTEST procedure. With small sample size (N=250) the statistical power for DIF detection decreased for both M-H and SIBTEST procedures as was manifested by lower detection rates of the three DIF items. The M-H still showed well controlled Type I error rate at .05 with all sample sizes in the 1:1, 1:.50 and 1:.10 ratios while SIBTEST had high Type I error rates at .05 with 250 sample size in the ratio of 1:.10.

Table 17

Means of Type I Error Rates Per Cell with Large and Moderate DIF Magnitudes

Sample size	Sample Ratios and DIF Detection Procedures											
	1:1				1:.50				1:.10			
	M-H		SIB		M-H		SIB		M-H		SIB	
	Large DIF	Mod DIF	Large DIF	Mod DIF	Large DIF	Mod DIF	Large DIF	mod DIF	Large DIF	Mod DIF	Large DIF	Mod DIF
1000	2.17	1.04	2.64	1.41	2.71	1.59	3.26	2.26	2.17	1.93	4.94	4.79
500	1.21	.66	1.68	1.15	1.52	1.22	2.50	2.13	2.39	1.41	**8.22**	**7.81**
250	.82	.57	1.41	1.23	1.22	1.09	2.85	2.71	1.07	.67	**11.75**	**12.54**

Note:

1. SIB in Table 17 is SIBTEST procedure
2. Mod DIF is moderate DIF magnitude
3. Bold numbers are SIBTEST average Type I error rates > .05 for non DIF items.

The M-H procedure had lower Type I error rates at .05 than SIBTEST with the same sample size, DIF magnitude and sample size ratios. Whitmore and Schumaker (1999) suggested that incomplete contingency table cells are responsible for reduced reliability of the calculated DIF statistics. On the basis of expanded analysis results it was evident that despite the proportions of the replications in which the DIF statistics were not calculated, the M-H procedure was still able to maintain control of its Type I error rates because the DIF statistics was calculated on basis of the valid cells within the contingency table. The expanded version of SIBTEST provided evidence that the DIF statistics were calculated only in the valid strata in which the number of examinees was ≥ 2. These results support Roussos and Stout's (1996) statement that M-H places zero weight on empty contingency cells. The inflated Type I error rates at .05 with SIBTEST were attributed to the unused strata in which the DIF statistics could not be calculated because the number of examinees was < 2 which was a default option. The DIF statistics that were calculated were done based on valid strata that were used. The means of all the Type I error rates at .05 were computed using SPSS to obtain more succinct results that would show a closer and clearer comparison of Type I error rates at the

same significant level, sample size and sample size ratio combinations for M-H and SIBTEST procedures as presented in Table 17.

The average Type I error rates at .05 with regards to sample sizes were consistent for 1:1 and 1:.50 ratios, with larger values observed in the 1:.5 ratio category. The trend changed with sample sizes in the ratios of 1:.10. For the M-H procedure a larger mean Type I error value was observed in the 1:.10 sample size ratios (2.39) than with 1:.50 sample size ratios (1.52). For the SIBTEST procedure, all the mean Type I error rates at .05 in the 1:.10 were higher than the means with the ratios in 1:1 and 1:.50 sample size categories. High mean Type I error rates > .05 observed with SIBTEST procedure was most likely due to the effect of sample size and sample size ratio as well as the magnitude of DIF. However, the mean Type I error rates > .05 for non DIF items shown in Table 17 for sample sizes 500 and 250 at 1:.10 ratios would also be considered to be an indication of how unstable the SIBTEST procedure gets when sample sizes are small and unequal. It was also worth noting that the high mean Type I error rates were due to only a few items for the M-H procedure. For example, with 250 sample size and moderate DIF, one non DIF item had Type I error rate of 43.70% at .05 level of significance while all the other non DIF items in the same sample size and ratio category had Type I error rates of < 5%, resulting in an overall mean Type I error rate of < .05 for the M-H procedure. With SIBTEST there were more non DIF items that were flagged for DIF at .05. Generally, the results for SIBTEST procedure showed more inflated mean Type I error rates at .05 than the M-H procedure.

Discussion, Conclusions, Recommendations and Limitations

The results of this study demonstrated that sample size ratios between the reference and focal groups had a significant effect on the statistical power of a DIF detection procedure. The results also indicated that not only the ratios, but also the size of the sample and the magnitude of DIF influenced the behavior of SIBTEST and M-H with regard to their error rate control. Even after controlling for ability differences Type I error was not controlled with some sample size ratios when SIBTEST procedure was used. With small samples and moderate DIF magnitude, Type II error was also committed by both M-H and SIBTEST procedures.

The results of the DIF analyses in this study provided useful indications that empirical sample sizes in the ratios of less than 1:.5 would not likely provide sufficient statistical power for reliable detection of moderate DIF magnitudes if M-H or SIBTEST procedure was to be used. With small sample sizes (N=250) the two procedures failed to attain the statistical

power threshold of 70% that Kaplan and George (1995) and González-Romá et al., (2006) considered to be acceptable level of DIF detection rate. With large sample size (N= 1000), and large DIF, both the M-H and the SIBTEST procedures showed sufficient statistical power, ≥ 70%.

On the basis of the results displayed in Tables 13-17 and in Figures 2 and 3 it could be argued that with unequal sample sizes in any ratio, M-H has a general trend of better performance than the SIBTEST with regression correction procedure, but on the average the two procedures had comparable results. However, the results of SIBTEST with unequal sample sizes in the ratios of 1:.10 were unreliable because the procedure was very unstable and yielded statistical power rates and Type I error rates that were not reliable. The results presented show that both procedures should not be used when sample sizes are in the ratios of 1:.10 because of lack of sufficient power to detect the DIF items of moderate magnitude and large standard errors of the DIF statistics.

On the basis of Type I error rate control, the M-H performed better even with small samples at 1:.10 ratios. With studies that involve empirical data where the DIF items are unknown, the M-H procedure would be a better choice if sample sizes are unequal. Based on the results of this study, the SIBTEST procedure should not be used when sample sizes are small and in the ratios of 1:.10 for the reference and focal groups; especially with empirical data. Holland and Weiner (1993) stated that the M-H procedure is a standard DIF detection procedure against which new methods should be judged. This study supports the position in which the M-H has been ranked in the past with regards to Type I error rates control. The results obtained from this study also indicated that even with regression correction, SIBTEST procedure still appeared to be inferior to M-H in its ability to control Type I error rates with unequal sample sizes in the reference and the focal groups at 1:.50 or 1:.10 ratios, especially with medium (N=500) and small (N= 250) sample sizes in the reference group with moderate DIF magnitude.

Hoover and Kolen (1984) criticized DIF detection procedures for statistical fluctuations and claimed that the procedures portrayed insignificant reliability. In the context of the results presented in Tables 13-17, M-H demonstrated a consistent behavior in terms of the statistical power and Type 1 error rate control with all sample size and all sample size ratios under study, indicating that the DIF procedure is significantly reliable. As Shealy and Stout (1993) stated, the quality of a statistical procedure can be indicated by its robustness and higher statistical power. The M-H showed the robustness indicative of a quality statistical procedure and sufficient power for DIF detection of moderate and large magnitudes. Cases of

isolated incidences with specific items call for a closer study for possible sampling fluctuation or item parameter characteristics, to determine the possible reason why the items were flagged for DIF even to levels higher than the flagging rates for the actual DIF items.

The possible performance outcome with SIBTEST DIF detection procedure may be predicted if item parameter characteristics, reference and focal group sample sizes, and DIF magnitude are known because the procedure performed fairly well when item difficulty was close to the mean, sample sizes were equal and DIF magnitudes were either moderate or large. Apparent statistical fluctuations observed by Hoover and Kolen may have been due to unequal sample sizes and characteristics of studied items. Simulation study results reported by Raju (1988) and Wang and Yeh (2003) suggested that the difference in abilities between the reference and the focal group is what affects the Type I error. This study controlled for ability differences and yet, Type I error was not adequately controlled at the cells with some sample size ratios in the reference to focal group by the SIBTEST procedure. The results of this study provided evidence that not only ability differences would affect a procedure's Type I error but sample size ratios between the references and focal groups also have significant effect on the statistical power of DIF detection procedures. With large DIF magnitude, it was evident that M-H procedure outperformed SIBTEST in all categories of sample sizes ratio combinations in terms of their Type I error rates at .05.

The authors of this paper view committing of Type I and Type II errors as very critical problems in assessment. Testing corporations depend on reliable DIF detection procedures that would not flag non DIF items. If non DIF items were falsely identified as biased with empirical data, the normal procedure would be to study the items closely and determine possible reasons for flagging. There could be times when the reasons for false flagging might not be determined and the items would have to be discarded. Finch and French (2007) stated that committing Type I error has serious economic and educational implications. As was seen in this study, non DIF items that had high discrimination and low difficulty or high difficulty and low discrimination were more likely to be falsely flagged for DIF, particularly by the SIBTEST procedure. If good items that discriminate between ability groups for fair decisions based on test results are vulnerable for flagging by SIBTEST when sample sizes are small and unequal, then it is important that SIBTEST procedure is avoided under those conditions to reduce the chances for possible economic and educational consequences of false flagging of items for bias.

In the view of the authors of this study, of more serious, and weightier, significance is committing Type II error due to lack of sufficient power of a DIF procedure to detect DIF

items. When Type II error is committed, examinees that would be branded as failures as a result of DIF items not identified present ethical, political and economic problems. The problems would be considered to be ethical in that the educational opportunities of the individuals affected by the decisions might never be redeemed. The issue is political in that policy decisions that may be made based on the biased test results may deprive the individual of career opportunities for life and hence lead to subsequent economic deprivation for the individual.

In cases of testing accommodation, fair interpretations of the results rely on the power of a DIF procedure to accurately and reliably detect items that demonstrate DIF. Based on the afore presented arguments and the results of this study the M-H procedure would be a better choice to use with empirical data of equal and unequal sample sizes in the reference and focal groups for most sample size ratios for equity in assessment.

This study, like all other simulation studies, was not without limitations. First, Monte Carlo studies, highly dependent on how realistic the conditions such as parameter distribution and data generation are modeled was taken to be a limitation, considering that model specifications and fit to data can be a real challenge. Due to the challenge, the results of Monte Carlo studies had to be interpreted with much caution. Another limitation of this study was that only sample size and DIF magnitude were manipulated in the data. DIF type was modeled as uniform and the effect of unequal sample sizes on non-uniform DIF was not estimated. Again, ability was modeled with normal distributions with a variance of 1.0 for all proficiency levels, while in studies with empirical data this would not be the case.

The seed value was arbitrarily selected and how arbitrary seed selection influenced the study was not examined. Another limitation was that the number of replications was not manipulated as one of the independent variables although it was possible that Monte Carlo results could vary with varying number of replications. The effects of test length and item response function were not estimated in this study although several studies report that they influence DIF detection and behavior of error rates. Since item parameter generation and DIF manipulation were done at the beginning of the study and the same parameters were used for all the cells of the design and for all replications within each cell, the results of the study may only be generalized to data with similar distributions.

Finally, Item Response Theory assumes that tests are unidimensional. However, many experts agree that most achievement and aptitude tests are actually multidimensional with perhaps a dominant primary dimension and several secondary dimensions (e.g., Ackerman, 1992; Camilli & Shepard, 1994; Embretson & Reise, 2000; Kok, 1988; Lord, 1980; ; Oshima et al.,

1997; Shealy & Stout, 1993a). Hence, the use of unidimensional 2PL IRT model with multidimensional test data possibly violated the unidimensionality assumption and posed serious threat to item and examinee parameter estimation, although it was assumed that taking a non parametric approach, SIBTEST assumptions were not necessarily binding.

Future studies should use large sample sizes comparable to those encountered in testing accommodation studies in the ratios that were used in this study to determine the performance of the M-H and the SIBTEST procedures with large unequal sample sizes. Item parameters used in future Monte Carlo studies should be generated using multidimensional models. Future Monte Carlo studies should not include DIF items in the matching subtest when using SIBTEST, instead the studies should take a confirmatory approach that does not require purification steps since DIF items are known a priori.

REFERENCES

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.

Bennett, R. E., Rock, D. A., & Novatkoski, I. (1989). Differential item functioning on the SAT-M Braille edition. *Journal of Educational Measurement, 26*(1), 67-79.

Buhr, D. C. (1988, April). Use of the multiple-category scoring procedure to investigate item bias. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*.Newbury Park, CA: Sage.

Clauser, B., Mazor, K. M., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31*, 67-78.

Dorans, N, J., & Schmitt, A. P. (1991). Constructed-response and differential item functioning: A pragmatic approach (ETS Research Report No. 91 -47). Prinstone, NJ: Educational Testing Service .

Fidalgo, Á. M., Ferreres, D., & Muñiz, J. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning with small sample sizes. *The Journal of Experimental Education, 73*(1), 23-39.

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

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.

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.

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.

Hambleton, R. K., Swaminathan, H., Rogers, J. H. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Han, K. T., & Hambleton, R. K. (2007). 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.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In Wainer & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Hoover, H. D., Kolen, M. J. (1984). The reliability of six item bias indices, Applied Psychological Measurement, 8, 173-181.

Hulin, C. L., Lissack, R. I., & Drasgow, F. (1982). Recovery of two-and three parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. Applied Psychological Measurement, 22, 345-355.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K.(1992)The effect of sample size on the function of Mantel-Haenszel statistics. *Educational and Psychological Measurement, 52*, 443-451.

McPeek, W. M., & Wild, C. L. (1986, April). *Performance of the ManteI-Haenszel statistic in a varie O' of situations.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16,* 381-388.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small sample conditions. *Journal of Educational Measurement, 32*(3), 302-316.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

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

Ryan, K. E. (1991) The performance of the Mantel-Haenszel procedure across samples and matching criteria. *Journal of Educational Measurement, 28*(4), 325-337.

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

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22,* 77-105.

Skaggs, G.; & Lissitz, R. W. (1992). The consistency of detecting item bias across independent samples: Implications of another failure. *Journal of Educational Measurement, 29*, 227-242.

Stout, W., & Roussos, L. (1995). *SIBTEST user manual.* Urbana: University of Illinois.

Thissen, D., Steinberg, L., & Weiner, H. (1993). Detection of differential functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

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,* 479-498.

Whitmore, M. L., & Schumacker, R. E. (1999).A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement, 59*, 910-927.

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.

Zwick, R., & Ericikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26,* 55-66.

CHAPTER SIX

Discussion, Conclusions, Recommendations and Limitations

The results of this study demonstrated that sample size ratios between the reference and focal groups had a significant effect on the statistical power of the SIBTEST and the M-H DIF detection procedures. The results also indicated that not only the ratios, but also the size of the sample and the magnitude of DIF influenced the behavior of the DIF procedures with regard to their Type I error rates control. Even after controlling for ability differences, inflated Type I error rates were still observed with small sample sizes at 1:.10 ratios when the SIBTEST procedure was used. With small samples and moderate DIF magnitude, the statistical power of both DIF detection procedures was remarkably reduced, resulting in Type II error. The results of the DIF analyses in this study provided useful indications that empirical sample sizes in the ratios of less than 1:.5 would not likely provide sufficient statistical power for reliable detection of moderate and large DIF magnitudes if SIBTEST procedure was to be used.

On the basis of the results of the two studies presented here it could be argued that with unequal sample sizes in any ratio, M-H would be a better DIF detection procedure choice. The credibility of the performance of the SIBTEST with regression correction should be judged against the M-H procedure in line with Holland and Weiner (1993) recommendation. In this Monte Carlo study the M-H procedure was a more efficient DIF detection procedure than SIBTEST when sample sizes were unequal, resulting in higher statistical power for detecting the a priori known DIF items and less Type I errors. In the context of the results of the studies presented here M-H demonstrated a consistently high statistical power and lower Type I error rate with most sample size ratios, to disapprove Hoover and Kolen's (1984) conclusion that DIF detection procedures portrayed insignificant reliability. However, variations were observed with SIBTEST procedure whose performance was more affected by unequal samples and sample ratios as well as item difficulty and discrimination. As Shealy and Stout (1993) stated, the quality of a statistical procedure can be indicated by its robustness and higher statistical power. The M-H showed the robustness indicative of a quality statistical procedure and sufficient power for DIF detection of moderate and large DIF magnitudes. Cases of isolated incidences like that with item 19 and 14 call for a closer study for possible sampling fluctuations or item parameter characteristics, to be able to determine the possible reason why the items were flagged for DIF to high proportions.

The SIBTEST procedure, however, was comparable to M-H when sample sizes were equal. When sample sizes were in the ratios of 1:.5, SIBTEST had higher Type I error rates at a critical level of .05 than M-H. The possible performance outcome with SIBTEST DIF detection procedure may be predicted if item parameter characteristics, reference and focal group sample sizes, and DIF magnitude are known because the procedure performed fairly well when item difficulty was close to the mean, sample sizes were equal and DIF magnitudes were either moderate or large. Apparent statistical fluctuations observed by Hoover and Kolen might have been due to unequal sample sizes and characteristics of studied items. Simulation study results reported by Raju (1988) and Wang and Yeh (2003) suggested that the difference in abilities between the reference and the focal group is what affects the Type I error. This study controlled for ability differences and yet, Type I error was not adequately controlled at the cells with small sample sizes in the focal group by the SIBTEST procedure. The results of this study provide evidence that not only ability differences would affect a procedure's Type I error but sample size ratios between the references and focal groups also have significant effect on the statistical power of DIF detection procedures. With large DIF magnitude, it was evident that M-H procedure outperformed SIBTEST in all categories of sample sizes ratio combinations in terms of their Type I error rates at .05.

Committing of Type I and Type II errors is a very critical problem in assessment. Testing corporations depend on reliable DIF detection procedures that would not flag non DIF items. If non DIF items were falsely identified as biased with empirical data, the normal procedure would be to study the items closely and determine possible reasons for false flagging. There could be times when the reasons for false flagging might not be determined and the items would have to be discarded. As was seen in this study, non DIF items that had high discrimination and low difficulty or high difficulty and low discrimination were more likely to be falsely flagged for DIF, particularly by the SIBTEST procedure. If good items that discriminate between ability groups for fair decisions based on test results are vulnerable for flagging by SIBTEST when sample sizes are small and unequal, then it is important that SIBTEST procedure is avoided under those conditions to reduce the chances for possible economic and educational consequences of false flagging of items for bias.

It can be argued that of more serious and weightier significance is committing Type II error due to insufficient statistical power for DIF procedure to accurately detect DIF items. When Type II error is committed, examinees that would be branded as failures as a result of DIF items not identified present ethical, political and economic problem. In cases of testing accommodation, fair interpretations of the results rely on the power of a DIF procedure to

accurately and reliably detect items that demonstrate DIF. Based on the afore presented arguments and the results of the studies M-H procedure would be a better choice to use with empirical data of equal and unequal sample sizes in the reference and focal groups for most sample size ratios for equity in assessment.

Future studies should compare the performance of SIBTEST with regression correction and the old SIBTEST DIF detection with unequal sample sizes to determine if it is worthwhile employing the newer version of SIBTEST. The relationship between item difficulty and discrimination parameters should be studied further with unequal samples to be able to explain why non DIF items with low discrimination and high difficulty or high discrimination and low difficulty get flagged for DIF, especially by SIBTEST procedure when DIF analysis show that the items have very small DIF statistics. A DIF study with same unequal sample sizes and combination ratios should be conducted again with SIBTEST using the confirmatory approach. In the study, the matching subtest should exclude the DIF items to determine the influence of the a priori known DIF items on the performance of the SIBTEST procedure.

Hambleton et al., (1991) and Lord (1980) noted that small sample sizes were not a problem with Rasch model. Future studies should use the Rasch model with unequal sample sizes to determine if using the model instead of the 2PLM for the estimation of item parameters would result in reduced Type I error rates when SIBTEST with regression is used for DIF analysis.

This study, like all other simulation studies, was not without limitations. First, Monte Carlo studies highly dependent on how realistic the conditions such as parameter distribution and data generation are modeled was taken to be a limitation, considering that model specifications and fit to data can be a real challenge. Due to the challenge, the results of Monte Carlo studies had to be interpreted with much caution. Another limitation of this study was that only sample size and DIF magnitude were manipulated in the data. DIF type was modeled as uniform and the effect of unequal sample sizes on non-uniform DIF was not estimated. Another limitation of this study was that ability was modeled with normal distributions with a variance of 1.0 for all proficiency levels, while in studies with empirical data this would not be the case.

The seed value was arbitrarily selected and how arbitrary seed selection influenced the study was not examined. Another limitation was that the number of replications was not manipulated as one of the independent variables although it was possible that Monte Carlo results could vary with varying number of replications. A large number (1000) of replications

was used to reduce chances of variance large enough to increase error rates that would be contributed by the number of replications. The effects of test length and item response function were not estimated in this study although several studies report that they influence DIF detection and behavior of error rates. Since item parameters generation and DIF manipulation were done at the beginning of the study and the same parameters were used for all the cells of the design and for all replications within each cell, the results of the study may only be generalized to similar data with the similar distributions.

Finally, Item Response Theory assumes that tests are unidimensional. However, many experts agree that most achievement and aptitude tests are actually multidimensional with perhaps a dominant primary dimension and several secondary dimensions (e.g., Ackerman, 1992; Camilli & Shepard, 1994; Embretson & Reise, 2000; Kok, 1988; Lord, 1980; Oshima et al., 1997; Shealy & Stout, 1993a). Hence, the use of unidimensional 2PL IRT model with multidimensional test data possibly violated the unidimensionality assumption and posed serious threat to item and examinee parameter estimation.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

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.

Angoff, W. H. (1993). Perspectives of differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 3-23.

Anrig, G. R. (1987). ETS on the golden rule. *Educational Measurement: Issues and Practice, 6*(3), 24--27.

Baker, F. B. (1985). *The basics of item response theory*. Portsmouth NH: Heinemann.

Bennett, R. E., Rock, D. A., & Novatkoski, I. (1989). Differential item functioning on the SAT-M Braille edition. *Journal of Educational Measurement, 26*(1), 67-79.

Bolt, D. M., & Gierl, M. J. (2004, April). *Application of a regression correction to three nonparametric test of DIF: Implications for global and local DIF detection*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased items*. Newbury Park, CA: Sage.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaptation-Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*, 169-199.

Chang, H., Mazzeo, J., & Roussos, L. (1993, April). Detecting DIF for polytomously scored items: An adaptation of Shealy-Stout's SIBTEST procedure. Paper presented at the annual Meeting of the American Educational Research Association, Atlanta GA.

Clauser, B., Mazor, K. M., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31*, 67-78.

DeMars, C. E. (2008). Modification of Mantel_Haenszel and Logistic Regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, XX*(X), xx-xx.

Donlon, T. F. (Ed.). (1984). *The college board technical handbook for the scholastic aptitude test and achievement tests*. New York: College Entrance Examination Board.

Dorans, N. J., & Holland, P. W. (1993). DIF detection description: Mantel-Haenszel and standardization. In P. W. Holland & H. Weiner (Eds.), Differential Item Functioning (pp. 35-66). Hillsdale, NJ. Erlbaum

Fidalgo, A. M., Mellenbergh, G. J., & Muñniz, J. (2000). Effects of amount of DIF, test length, and puri.cation type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5*, 43-53.

Fidalgo, Á. 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. *The Journal of Experimental Education, 73*(1), 23-39.

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.

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

Fischer, G. H. (1993). Notes on the Mantel–Haenszel procedure and another chi-squared test for the assessment of DIF. *Methodika 7*, 88–100.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika 60*:449–487.

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.

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.

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.

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.

González-Romá, V., Tomás, I., Ferreres, D., & Hernández, A. (2005). Do items that measure self-perceived physical appearance function differentially across gender groups of adolescents? An application of the MACS model. *Structural Equation Modeling, 12*, 157-171.

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.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*(3), 211-235.

Han, K. T., & Hambleton, R. K. (2007). 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.

Hanson, B. A. (1998). Uniform DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics. 23*(3) 244-253.

Harwell, M., R. (1991, April). Analyzing and reporting the results of Monte Carlo studies in educational and psychological research. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Harwell, M., R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279-291.

Harwell, M., R. (1997). Analyzing Monte Carlo results in item response theory. *Educational and Psychological Measurement, 57*(2), 266-279.

Harwell, M., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics, 17*, 315-339.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125.

Hauck, W. W., & Anderson, S. (1984). A survey regarding the reporting of simulation studies. *The American Statistician, 38,* 214-216.

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.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In Wainer & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Hoover, H. D., & Kolen, M. J. (1984). The reliability of six item bias indices, *Applied Psychological Measurement, 8*, 173-181.

Hulin, C. L., Lissack, R. I., & Drasgow, F. (1982). Recovery of two-and three parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Jiang, H., & Stout, W. (1998). Improve Type 1 error control and reduce estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics, 23*(4) 291-322.

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 2,* 101-118.

Kim, S. H., Cohen, A. S., & Kim, H. O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement, 18*(3), 217-228.

Kubiak, A. T., & Colwell, W. R. (1990, April). *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.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

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.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K.(1992)The effect of sample size on the function of Mantel-Haenszel statistics. *Educational and Psychological Measurement, 52*, 443-451.

McAllister, P. H. (1993). Testing DIF and public policy. In P.W. Holland and H. Wainer (Eds.), *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 381-396.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100-117.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillside, NJ: Erlbaum.

McPeek, W. M., & Wild, C. L. (1986, April). *Performance of the ManteI-Haenszel statistic in a varie O' of situations.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Meredith, W. & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*, 289-311.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on attitude survey. *Journal of Educational Measurement, 41*(4), 331-344.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16,* 381-388.

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

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in statistical detection of measurement bias. *Applied Psychological Measurement, 16*, 389-402.

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.

Narayanan, P. & Swaminathan, H. (1996). Identification of items that show nonuniform bias. *Applied Psychological Measurement 20*(3), 257-274.

Naylor, T. H., Balintfy, J. L., Burdick, D. S., & Chu, K. (1968). *Computer Simulation Techniques*. New York: Wiley.

Pae, T. I. (2004). DIF for examinees with different academic backgrounds. *Language Testing 21*, 53-73.

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.

Phillips, A., & Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics, 43,* 425-43 I.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

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

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.

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.

Rosser, P. (1989). Gender and testing. ERIC ED 336457. Educational Resources Information Center.

Roussos, L., & Stout, W. F. (1996a). A multidimensionality-based DIF analysis paradigm. A*pplied Psychological Measurement, 20*, 355–371.

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

Roussos, L. A., & Stout, W. (2004). Differential item functioning analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 107-116). Thousand Oaks, CA: Sage:

Ryan, K. E. (1991) the performance of the Mantel-Haenszel procedure across samples and matching criteria. *Journal of Educational Measurement, 28*(4), 325-337.

Rubinstein, R. V. (1981). *Simulation and the Monte Carlo Method*. New York: Wiley.

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

Sheng, Y. (2005, July). Baysian analysis of hierarchical IRT models: Comparing and combining the unidimensional and the multi-dimensional IRT models. Dissertation published online: http://edt.missouri.edu/summer 2005/dissertation/shengY-071905-D2504/research.pdf

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22,* 77-105.

Spence, I. (1983). Monte Carlo simulation studies. *Applied Psychological Measurement, 7*, 405-425.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

Stone, C. A. (1993). The use of multiple replications in IRT based Monte Carlo research. Paper presented at the European Meeting of the Psychometric Society, Barcelona.

Stout, W., Li, H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement, 21*(3), 195-213.

Stout, W., & Roussos, L. (1995). *SIBTEST user manual*. Urbana: University of Illinois.

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

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.

Thissen, D., Steinberg, L., & Weiner, H. (1993). Detection of differential functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Tindal, G., & Fuchs, L. (1999). *A summary of research on test changes: An empirical basis for defining accommodations*. Lexington, KY: University of Kentucky, Mid-South Regional Resource Center.

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.

Von Neuman, J. (1951). Various techniques used in connection with random digits. *Standards Applied Mathematics Series, 12*, 36-38.

Van der Flier, H., Mellenbergh, G. J., Ader, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21*, 131-145.

Wasti, S. A., Bergman, M. E., Glomb, T. M., & Drasgo, F. (2000). Test of the cross-cultural generalizability of a model of sexual harassment. *Journal of Applied Psychology, 85*, 766-778.

Welkenhuysen-Gybels, J. (2004). The performance of some observed and unobserved conditional invariance techniques for the detection of differential item functioning. *Quality and quantity, 38*, 681-708.

Welkenhuysen-Gybels, J., & Billiet, J. (2002). A comparison of techniques for detecting cross-cultural inequivalence at the item level. Quality and Quantity, 36: (In Press).

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.

Zheng, Y., Gierl, M. J., & Cui, Y. (2007, April). *Using real data to compare DIF detection and effect size measures among mantel-Haenszel, SIBTEST and Logistic Regression procedures.* A paper presented at the annual meeting of the National Council on Measurement in Education: Chicago, ILL.

Zwick, R., & Ericikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26,* 55-66.

Zwick, R., Donoghue, J. R., & Girma, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Measurement, 30*, 233-251.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1-28.

APPENDIX 1

EXTENDED SIBTEST RUN

OUTPUT FOR RUN NUMBER 13 500/50 moderate

Suspect subtest items:

13

Matching subtest items:

1	2	3	4	5	6	7	8	9	10
11	12	14	15	16	17	18	19	20	21
22	23	24	25	26	27	28	29	30	31
32	33	34	35	36	37	38	39	40	41
42	43	44	45	46	47	48	49	50	

estimate of guessing on this matching subtest = 0.20

Matching Subtest Score	NR	NF	ybar-R	ybar-F	Adj. ybar-R	Adj. ybar-F	D	D*wt-p
0	2	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	4	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	6	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	7	0	0.1429	0.0000	0.1429	0.0000	0.0000	0.0000
4	7	1	0.1429	0.0000	0.1429	0.0000	0.0000	0.0000
5	7	3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	7	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	8	0	0.1250	0.0000	0.1250	0.0000	0.0000	0.0000
8	7	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	13	1	0.0769	0.0000	0.0769	0.0000	0.0000	0.0000
10	13	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
11	13	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
12	13	1	0.0769	0.0000	0.0769	0.0000	0.0000	0.0000
13	18	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
14	10	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15	12	2	0.0833	0.5000	0.0833	0.4704	-0.3871	-0.0319
16	12	3	0.0833	0.0000	0.0822	0.0000	0.0822	0.0073
17	12	2	0.1667	0.0000	0.1669	0.0000	0.1669	0.0137
18	16	3	0.0625	0.0000	0.0623	0.0000	0.0623	0.0070
19	12	1	0.1667	1.0000	0.1667	1.0000	0.0000	0.0000
20	16	2	0.1875	0.0000	0.1872	0.0000	0.1872	0.0198
21	14	0	0.0714	0.0000	0.0714	0.0000	0.0000	0.0000
22	11	1	0.1818	0.0000	0.1818	0.0000	0.0000	0.0000
23	12	1	0.1667	0.0000	0.1667	0.0000	0.0000	0.0000
24	14	3	0.1429	0.0000	0.1431	0.0000	0.1431	0.0143
25	9	0	0.1111	0.0000	0.1111	0.0000	0.0000	0.0000
26	19	1	0.2632	0.0000	0.2632	0.0000	0.0000	0.0000
27	12	1	0.2500	0.0000	0.2500	0.0000	0.0000	0.0000
28	16	0	0.3750	0.0000	0.3750	0.0000	0.0000	0.0000
29	13	1	0.0769	1.0000	0.0769	1.0000	0.0000	0.0000
30	13	0	0.3077	0.0000	0.3077	0.0000	0.0000	0.0000
31	6	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
32	10	2	0.1000	0.5000	0.1004	0.4991	-0.3987	-0.0281
33	9	0	0.4444	0.0000	0.4444	0.0000	0.0000	0.0000
34	9	1	0.2222	0.0000	0.2222	0.0000	0.0000	0.0000
35	15	2	0.4000	1.0000	0.4005	0.9988	-0.5983	-0.0598
36	13	3	0.3077	1.0000	0.3076	1.0000	-0.6924	-0.0652
37	8	0	0.3750	0.0000	0.3750	0.0000	0.0000	0.0000
38	12	0	0.4167	0.0000	0.4167	0.0000	0.0000	0.0000
39	8	1	0.3750	1.0000	0.3750	1.0000	0.0000	0.0000

```
   40     9   0  0.1111  0.0000  0.1111  0.0000   0.0000   0.0000
   41     9   2  0.3333  0.0000  0.3348  0.0018   0.3330   0.0215
   42     7   2  0.7143  0.5000  0.7143  0.4769   0.2374   0.0126
   43    12   0  0.3333  0.0000  0.3333  0.0000   0.0000   0.0000
   44     6   1  0.8333  0.0000  0.8333  0.0000   0.0000   0.0000
   45     6   0  0.8333  0.0000  0.8333  0.0000   0.0000   0.0000
   46     7   1  0.7143  0.0000  0.7143  0.0000   0.0000   0.0000
   47     1   0  1.0000  0.0000  1.0000  0.0000   0.0000   0.0000
   48     2   0  1.0000  0.0000  1.0000  0.0000   0.0000   0.0000
   49     3   0  0.0000  0.0000  0.0000  0.0000   0.0000   0.0000

no. of possible usable cells =  48
proportion of cells used = 0.250

proportion of Ref. grp. examinees eliminated =  0.716
proportion of Focal grp. examinees eliminated =  0.440
KR-20 for Ref. grp. =   0.961
KR-20 for Foc. grp. =   0.964

Matching Subtest Summary Statistics
    Reference Group:              Mean = 23.69
                     Standard deviation = 12.15
    Focal Group:                  Mean = 23.00
                     Standard deviation = 12.07

    Standardized Score Difference = 0.06


              SIBTEST-pooled weighting Results
                                          p-value
 Beta estimate   standard error   for DIF against either grp.
    -0.089           0.068                0.194002

                   SIBTEST error flag = 0
   No errors. This SIBTEST run had a normal successful completion.
```

APPENDIX 2

EXPANDED SIBTEST RUN WITH ITEM 46

```
Suspect subtest items: 46


 Valid subtest items:
  1  2  3  4  5  6  7  8  9 10
 11 12 13 14 15 16 17 18 19 20
 21 22 23 24 25 26 27 28 29 30
 31 32 33 34 35 36 37 38 39 40
 41 42 43 44 45 47 48 49 50

  Valid
 Subtest                          Adj.  Adj.
  Score   NR  NF ybar-R ybar-F ybar-R ybar-F     D    D*wt
     0     2   1  .0000  .0000  .0000  .0000   .0000   .0000
     1     5   0  .2000  .0000  .2000  .0000   .0000   .0000
     2     6   0  .1667  .0000  .1667  .0000   .0000   .0000
     3     8   1  .3750 1.0000  .3750 1.0000   .0000   .0000
     4     7   2  .4286 1.0000  .4286 1.0000   .0000   .0000
     5     7   1  .2857  .0000  .2857  .0000   .0000   .0000
     6     7   0  .2857  .0000  .2857  .0000   .0000   .0000
     7     9   0  .4444  .0000  .4444  .0000   .0000   .0000
     8     8   1  .5000  .0000  .5000  .0000   .0000   .0000
     9    15   3  .4667  .6667  .4667  .6667   .0000   .0000
    10    13   0  .4615  .0000  .4615  .0000   .0000   .0000
    11    15   2  .5333  .5000  .5333  .4616   .0717   .0084
    12    10   0  .6000  .0000  .6028  .0093   .0000   .0000
    13    21   1  .3810 1.0000  .3775 1.0000   .0000   .0000
    14    10   1  .8000 1.0000  .7939 1.0000   .0000   .0000
    15    12   3  .7500 1.0000  .7508  .9949  -.2441  -.0253
    16    12   3  .7500  .6667  .7486  .6667   .0820   .0085
    17    13   2  .8462 1.0000  .8475  .9912  -.1437  -.0149
    18    17   1  .6471  .0000  .6452  .0000   .0000   .0000
    19    13   3 1.0000 1.0000  .9978 1.0000  -.0022  -.0002
    20    13   0  .8462  .0000  .8495  .0000   .0000   .0000
    21    15   1  .6667 1.0000  .6653 1.0000   .0000   .0000
    22    10   1 1.0000 1.0000  .9984 1.0000   .0000   .0000
    23    15   2  .8667 1.0000  .8675  .9933  -.1258  -.0148
    24     8   1  .8750  .0000  .8752  .0000   .0000   .0000
    25    17   1  .8235 1.0000  .8233 1.0000   .0000   .0000
    26    15   1  .9333 1.0000  .9327  .9966   .0000   .0000
    27    13   0 1.0000  .0000  .9998  .0000   .0000   .0000
    28    16   0 1.0000  .0000 1.0000  .0013   .0000   .0000
    29    11   1  .8182 1.0000  .8182 1.0000   .0000   .0000
    30    10   2  .9000 1.0000  .9001 1.0000  -.0999  -.0083
    31     9   1  .8889 1.0000  .8888 1.0000   .0000   .0000
    32     7   1  .8571 1.0000  .8575 1.0000   .0000   .0000
    33    11   1 1.0000 1.0000 1.0000 1.0000   .0000   .0000
    34    11   0 1.0000  .0000 1.0000  .0000   .0000   .0000
    35    15   2 1.0000 1.0000 1.0000  .9937   .0063   .0007
    36     9   3 1.0000 1.0000 1.0000 1.0000   .0000   .0000
    37    10   0 1.0000  .0000 1.0000  .0084   .0000   .0000
    38    10   0 1.0000  .0000 1.0000  .0000   .0000   .0000
    39    11   1 1.0000 1.0000 1.0000  .9894   .0000   .0000
    40     7   2 1.0000 1.0000 1.0000  .9995   .0005   .0000
    41     5   1 1.0000 1.0000 1.0000 1.0000   .0000   .0000
    42    13   1 1.0000 1.0000 1.0000 1.0000   .0000   .0000
```

```
   43     5   1 1.0000 1.0000 1.0000 1.0000   .0000   .0000
   44     6   0 1.0000  .0000 1.0000  .0000   .0000   .0000
   45     7   1 1.0000 1.0000 1.0000 1.0000   .0000   .0000
   46     5   0 1.0000  .0000 1.0000  .0000   .0000   .0000
   47     1   0 1.0000  .0000 1.0000  .0000   .0000   .0000
   48     5   0 1.0000  .0000 1.0000  .0000   .0000   .0000
   49     0   0  .0000  .0000  .0000  .0000   .0000   .0000

no. of possible usable cells =  48
proportion of cells used =  .208
proportion of Ref. grp. examinees eliminated =   .758
proportion of Focal grp. examinees eliminated =   .520

                     SIB-uni          Mantel-Haenszel Results
                   p-value for             p-value for
          SIB-uni   DIF against             DIF against
             z      either Ref.      Chi   either Ref.  Delta
Beta-uni statistic  or Foc. grp.     sqr. or Foc. grp. (D-DIF)
  -.046     -.615       .539          .57      .452      1.187
```

APPENDIX 3

PERCENT ITEMS FLAGGED FOR LARGE SAMPLE SIZE

Appendix 3

Percent items flagged for large sample size (N=1000) with large DIF items

Item	1000/1000		1000/500		1000/100	
No.	MH	SIB	MH	SIB	MH	SIB
1	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**90.00**
2	1.50	1.70	1.10	1.50	1.00	3.60
3	1.20	1.50	3.50	3.90	.10	1.30
4	3.70	4.50	.70	1.60	.60	1.90
5	.90	.20	1.60	1.60	**5.40**	**6.40**
6	2.00	2.30	1.50	1.90	.30	2.50
7	4.40	4.50	3.20	4.70	2.10	**6.20**
8	2.00	2.10	.80	1.00	3.60	4.60
9	3.40	2.80	1.20	1.50	1.50	2.70
10	4.80	5.30	**8.30**	10.80	20.10	**25.00**
11	3.80	3.60	**15.40**	9.80	3.10	**5.00**
12	2.60	1.80	**5.50**	6.50	.70	2.70
13	**100.00**	**100.00**	**100.00**	**100.00**	**95.00**	**83.00**
14	1.50	1.60	00	.50	6.70	**10.50**
15	1.40	1.80	5.30	**5.30**	.30	2.40
16	1.50	3.00	1.40	3.80	.10	1.40
17	3.00	3.10	1.60	1.20	.30	2.10
18	1.10	.70	2.20	2.90	.50	1.70
19	1.00	1.60	.10	.70	1.00	2.50
20	2.70	2.40	1.30	1.90	.50	2.90
21	1.00	1.70	.40	.1.10	.20	3.10
22	3.60	4.70	1.80	2.70	2.10	**7.30**
23	2.20	1.80	.80	1.80	.20	3.50
24	1.50	1.10	4.20	3.20	.70	1.90
25	1.40	2.00	.70	1.20	1.80	**6.10**
26	1.50	1.40	1.70	1.70	**13.20**	**12.50**
27	.40	1.30	.70	1.20	2.10	**15.00**
28	1.80	3.00	1.00	1.20	.90	1.50
29	3.10	4.40	.70	1.90	.30	3.00
30	3.20	4.50	3.10	3.60	.20	2.10
31	3.90	**5.20**	.80	1.10	.60	3.00
32	1.90	2.00	.40	1.00	1.60	**7.10**
33	2.40	3.30	1.70	1.60	1.60	2.00
34	2.60	3.70	4.00	**5.10**	.50	1.90
35	3.30	4.70	.40	1.00	1.10	3.80
36	1.60	1.80	.60	1.40	3.10	**11.80**
37	1.70	3.90	2.20	5.60	.50	**8.70**
38	1.00	1.70	4.40	6.30	3.10	3.80
39	.70	1.60	6.80	6.10	1.00	.20
40	2.50	4.20	3.90	5.00	1.50	4.00
41	1.60	1.70	3.80	3.90	.50	2.20
42	1.60	2.00	.30	.60	.30	2.90
43	1.70	2.60	1.80	2.60	.50	2.30
44	1.40	1.60	.50	1.00	.70	**5.10**
45	2.80	3.80	1.40	2.40	.60	2.00
46	**100.00**	**99.80**	**100.00**	**99.90**	**96.30**	**91.70**
47	2.70	3.90	3.40	**5.20**	.80	2.60
48	2.70	2.80	1.60	1.90	1.90	**8.90**
49	2.50	2.30	**19.60**	**19.80**	**11.10**	**14.20**
50	1.10	.90	1.10	2.00	1.50	4.40

APPENDIX 4

PERCENT ITEMS FLAGGED FOR MEDIUM SAMPLE SIZE

Appendix 4

Percent items flagged for medium sample size (N=500) with Large DIF items

Item	500/500		500/250		500/50	
No.	M-H	SIB	M-H	SIB	M-H	SIB
1	**100**	**100**	**100**	**100**	**81.70**	**67.80**
2	.90	2.00	.80	1.30	.20	**12.60**
3	.70	1.20	.40	.90	.20	3.60
4	1.90	3.40	2.30	2.70	1.20	**5.30**
5	1.10	.80	.70	2.40	1.40	4.40
6	.80	.90	.30	2.10	.60	4.70
7	1.20	1.90	.40	1.40	.70	**6.20**
8	.70	1.40	1.00	1.30	4.30	**8.20**
9	1.40	1.30	.50	1.80	.50	4.70
10	1.90	3.60	4.10	4.80	2.10	**6.10**
11	2.90	1.90	4.10	4.10	.50	**7.20**
12	.70	1.60	3.40	**5.20**	.20	4.50
13	**100**	**100**	**100**	**99.90**	**35.30**	**22.10**
14	.40	1.10	1.10	1.90	**8.00**	**19.20**
15	.60	.90	.40	.80	.50	5.00
16	.70	1.20	1.60	2.70	.40	3.90
17	.70	1.80	1.10	2.20	2.00	**9.50**
18	.80	1.70	2.20	3.00	1.60	**5.30**
19	.50	1.40	.90	2.00	**48.90**	**39.90**
20	1.90	1.40	1.00	1.20	.10	3.60
21	.90	1.20	.50	1.10	.80	**6.40**
22	3.00	4.10	**5.20**	**11.00**	2.10	**18.00**
23	1.30	1.90	3.70	2.50	.10	**18.10**
24	.70	.50	.40	1.90	.60	**5.80**
25	.50	1.60	2.50	2.90	.90	**7.50**
26	1.60	.70	2.50	2.90	.50	3.40
27	.70	1.30	1.60	1.30	1.50	**16.20**
28	.70	1.50	3.20	**7.10**	4.70	**5.40**
29	1.20	1.80	1.40	.1.60	1.40	**5.10**
30	1.60	2.20	.70	1.90	.50	4.20
31	2.00	3.00	.70	1.60	.90	**5.20**
32	1.60	2.80	1.30	3.20	2.40	**18.70**
33	1.20	1.30	**7.10**	**9.80**	.50	**5.10**
34	1.90	2.60	.90	1.50	2.10	**12.30**
35	1.60	2.00	1.30	2.50	.70	**5.20**
36	1.10	1.60	.10	.50	2.20	**8.80**
37	1.80	3.00	1.10	3.20	.50	**11.60**
38	.10	.20	2.20	2.90	5.00	**5.20**
39	1.00	.70	.40	.90	.20	3.20
40	1.50	2.50	.50	1.40	.70	**6.40**
41	.60	1.30	.40	.80	00	3.20
42	1.00	1.10	.60	1.50	.20	**5.60**
43	.70	1.00	1.10	1.90	1.20 (7)	**7.30**
44	1.00	1.10	1.10	3.30	3.50 (7)	**13.60**
45	1.30	1.20	.50	1.10	.50 (7)	**5.50**
46	**100**	**99.60**	**100**	**98.90**	**55.90** (8)	**51.40**
47	2.20	2.00	.60	1.60	.80 (8)	**5.40**
48	1.30	1.70	1.10	1.90	3.30 (8)	**10.90**
49	1.90	2.20	2.20	2.90	.80 (8)	4.10
50	.90	1.20	.10	.30	.40 (8)	4.90

APPENDIX 5

PERCENT ITEMS FLAGGED FOR SMALL SAMPLE SIZE

Appendix 5

Percent items Flagged for small sample size (N=250) with Large DIF items

Item	250/250		250/125		250/25			
no.	M-H *p*-value	SIB p-value	M-H *p*-value	SIB *p*-value	M-H *p*-value	M-H % missing	SIB *p*-value	SIB % missing
1	**100**	**100**	**98.90**	**97.80**	**84.80**	**27**	**59.80**	**27**
2	.90	1.10	1.90	**6.10**	.00	117	**8.40**	108
3	.10	.70	.20	1.40	.90	118	**11.90**	22
4	1.50	2.40	1.10	2.20	4.10	119	**11.30**	15
5	1.00	1.00	1.00	1.10	1.70	120	**9.50**	25
6	.80	1.70	.50	1.80	.60	164	**8.90**	71
7	.40	1.40	.70	3.10	.50	166	**11.40**	00
8	.50	.60	.40	.50	1.90	167	**10.90**	25
9	1.10	1.90	.70	2.30	.20	172	**10.40**	27
10	1.60	2.40	4.20	**6.20**	1.20	177	**11.00**	28
11	1.40	1.90	.80	3.10	1.90	352	**22.20**	141
12	.50	1.00	2.60	3.90	1.50	354	**15.20**	43
13	**100**	**100**	**95.90**	**91.10**	**10.50**	**359**	**11.10**	**26**
14	.40	.70	3.60	**6.50**	10.50	360	**21.80**	23
15	.70	.70	1.90	4.40	.50	362	**8.40**	27
16	.30	1.50	.60	2.40	.20	363	**7.80**	21
17	1.60	1.80	2.40	3.50	2.70	365	**14.30**	34
18	.50	.90	.20	1.30	.00	366	**8.90**	23
19	.40	1.20	1.50	1.60	**5.40**	366	**15.50**	27
20	.90	1.30	2.50	3.40	.30	369	**10.20**	25
21	.50	1.20	.20	1.70	.20	370	**10.70**	30
22	1.50	2.50	3.00	**5.50**	1.20	40.30	**17.80**	8.70
23	.60	1.20	2.40	3.60	.30	42.70	**13.8**	28.2
24	.70	1.30	.20	2.00	.40	45.60	**6.90**	115.20
25	.60	1.20	4.10	**5.90**	.00	45.70	**10.00**	30
26	.60	1.30	.90	1.70	.60	48.00	**7.80**	6.20
27	.80	1.00	2.20	**6.60**	.10	48.10	**16.60**	26
28	1.30	1.90	.80	2.50	1.30	48.40	**10.70**	3.40
29	.80	1.50	.20	2.30	.70	48.50	**11.30**	30
30	1.20	1.60	1.20	1.20	.10	48.60	**9.70**	21
31	.80	2.70	.80	2.60	.40	49.40	**10.90**	59
32	.10	1.00	.30	1.80	.20	49.5	**16.70**	27
33	1.30	1.70	.40	2.30	.50	49.60	**10.80**	42
34	.70	1.40	.60	2.60	.30	49.70	**16.50**	25
35	1.60	2.50	1.00	2.00	.70	49.80	**8.20**	22
36	.90	1.00	1.30	3.50	.00	60.90	**8.10**	24.20
37	1.40	2.20	.60	4.20	.20	60.90	2.80	30
38	.30	.80	.50	1.60	.20	60.90	**13.50**	26
39	.60	.40	1.50	3.70	.20	61.00	**6.90**	28
40	.30	1.00	.50	1.40	.10	61.10	**7.80**	28
41	.70	1.10	.60	2.20	.10	61.20	**9.30**	44
42	.70	1.40	1.20	1.70	.20	61.20	**12.20**	27
43	.60	1.90	1.20	2.70	.20	61.40	**10.30**	35
44	.40	.70	.30	.90	1.10	61.40	**19.00**	23
45	.90	1.60	1.00	2.10	1.50	64.00	**15.30**	23
46	**93.10**	**86.00**	**84.50**	**73.70**	**.80**	**64.10**	**16.90**	**89**
47	1.80	2.00	.30	1.70	.40	64.20	**9.70**	19
48	1.10	1.40	.80	3.50	.60	64.20	**11.30**	32
49	.70	1.60	1.60	2.80	1.70	64.20	**14.20**	29
50	.30	.80	.70	2.70	2.50	64.20	**15.30**	29

APPENDIX 6

PERCENT ITEMS FLAGGED FOR LARGE SAMPLE SIZE (N=1000) WITH MODERATE DIF ITEMS

Item	1000/1000		1000/500		1000/100	
No.	M-H	SIB	M-H	SIB	M-H	SIB
1	**100**	**100**	**98.90**	**96.80**	**78.80**	**69.60**
2	.60	1.20	3.10	2.70	2.40	2.90
3	.70	.70	1.30	2.20	00	1.30
4	1.20	1.70	1.40	1.40	.60	2.30
5	.60	1.10	.40	.70	4.60	**5.60**
6	.60	.80	1.10	1.70	.40	3.30
7	1.80	2.40	1.90	3.40	2.10	**6.00**
8	.60	1.10	.70	1.20	2.60	4.40
9	.50	1.60	1.00	1.90	.40	3.20
10	2.00	1.70	4.60	**6.30**	**12.40**	**19.70**
11	3.10	1.90	**6.70**	**5.20**	1.40	4.90
12	.70	1.00	3.40	**5.60**	.50	2.80
13	**100**	**100**	**99.70**	**99.60**	**15.40**	**13.70**
14	.40	.40	.50	.50	**5.60**	**9.40**
15	.60	.80	2.80	3.80	.30	3.30
16	1.00	1.10	.70	1.20	.20	2.60
17	1.40	1.90	.40	.70	.80	3.10
18	1.10	1.30	2.20	4.90	.30	2.40
19	.20	.60	.60	1.30	.80	2.70
20	1.90	2.40	.80	1.60	.60	3.10
21	1.70	1.00	1.50	2.00	.10	3.00
22	2.60	3.80	1.60	2.40	2.10	**7.60**
23	1.20	1.50	1.80	1.40	.40	2.60
24	1.50	.70	.80	2.20	.70	1.80
25	1.00	1.10	1.60	1.30	.80	4.50
26	.30	.40	.90	1.10	**9.30**	**11.70**
27	.30	.70	1.30	1.70	3.30	**14.40**
28	.80	1.10	.90	1.10	.90	2.80
29	1.70	2.30	1.00	1.40	1.20	2.30
30	.70	2.10	1.10	1.50	.40	2.30
31	1.30	2.10	1.00	1.40	1.20	3.60
32	.60	1.00	.70	1.50	2.10	**8.60**
33	.80	1.50	.50	.80	.50	2.20
34	1.00	1.30	2.80	3.20	.20	2.50
35	1.40	1.60	.40	.90	.70	4.10
36	.90	1.50	2.20	5.00	2.90	**9.40**
37	1.50	2.60	3.60	**6.10**	.90	**5.40**
38	.30	.80	4.40	5.00	3.60	3.00
39	.10	.90	2.40	3.70	00	1.60
40	1.20	1.20	1.80	1.90	1.40	3.20
41	.50	1.20	.90	1.50	.10	2.10
42	.80	1.10	1.40	2.40	.30	2.50
43	1.60	1.90	.50	1.20	.40	2.70
44	1.10	1.50	1.00	1.70	1.10	5.10
45	1.30	1.90	1.00	1.70	1.60	3.60
46	**99.90**	**97.10**	**96.60**	**99.60**	**60.90**	**61.90**
47	.70	2.00	1.90	2.70	.90	3.20
48	.70	1.10	.60	1.30	**5.80**	**10.30**
49	1.40	1.50	.80	.70	**8.90**	**11.00**
50	.90	1.30	.50	1.30	3.00	**5.10**

APPENDIX 7

PERCENT ITEMS FLAGGED FOR MEDIUM SAMPLE SIZE (N=500) WITH MODERATE DIF ITEMS

Item	500/500		500/250		500/50	
no.	M-H	SIB	M-H	SIB	M-H	SIB
1	**95.80**	**94.80**	**86.80**	**83.30**	**20.6**	**28.10**
2	.60	1.00	.90	1.50	.40	**10.40**
3	00	.60	.30	.80	.30	3.30
4	.80	.90	1.00	1.90	1.00	4.00
5	.50	.70	.40	1.00	.90	4.00
6	.60	.90	.50	.90	1.30	**5.20**
7	.30	.70	1.10	1.70	.40	**5.80**
8	1.00	.80	.70	1.80	2.50	**7.30**
9	.20	1.40	.20	1.30	.30	**6.30**
10	1.10	2.00	2.30	3.50	.70	**7.20**
11	1.10	1.60	3.00	**5.30**	.50	**7.70**
12	.70	1.10	2.20	3.60	.50	**5.90**
13	**97.30**	**96.90**	**51.30**	**47.80**	**1.60**	**4.70**
14	.40	1.00	.40	1.70	**6.60**	**14.90**
15	.70	1.20	.50	1.20	1.80	**6.40**
16	.50	1.20	1.40	2.30	00	3.90
17	.70	1.20	1.30	1.30	3.30	**9.00**
18	.60	1.70	1.40	3.20	1.20	4.80
19	.20	.30	.50	1.00	**47.00**	**39.10**
20	.90	1.20	.30	.70	.40	**5.10**
21	.60	1.00	.20	1.00	.90	**5.90**
22	1.00	2.10	4.80	**7.90**	2.00	**17.10**
23	.30	1.00	1.80	1.40	.40	**7.20**
24	.50	.70	.20	2.50	.50	**5.20**
25	.30	.50	.90	1.10	.30	**5.20**
26	.30	.90	1.50	2.30	.50	**5.30**
27	.50	1.40	.70	1.40	1.50	**15.90**
28	.50	1.10	1.10	2.50	**6.60**	**7.40**
29	1.40	1.60	1.40	1.70	1.20	**5.50**
30	.80	.80	.50	.60	.50	4.50
31	.90	1.80	1.20	1.50	.50	**6.60**
32	1.20	1.00	2.00	3.20	2.10	**17.70**
33	.90	1.20	**6.40**	**7.80**	.30	**5.50**
34	.60	1.30	.70	1.80	3.40	**15.50**
35	.80	1.10	1.80	2.50	1.00	4.70
36	.90	1.90	.20	1.10	1.80	**7.40**
37	1.60	2.90	1.30	3.10	.50	**11.40**
38	.50	.80	2.80	3.00	**5.60**	**6.20**
39	.20	.40	.80	1.70	.80	**5.90**
40	.30	.80	.10	1.40	.10	3.90
41	.50	1.20	1.80	3.70	.20	**5.10**
42	.30	.80	1.00	1.50	1.20	**6.30**
43	1.40	1.30	.50	1.70	2.80 (7)	**13.20**
44	.50	.70	.40	1.00	.80	3.70
45	.80	1.40	.50	.70	**22.20** (8)	**31.20**
46	**83.00**	**62.40**	**88.40**	**79.60**	**.20**	**4.30**
47	1.00	1.70	1.30	2.40	4.00 (8)	**11.20**
48	.80	1.70	1.40	1.80	4.00 (8)	3.6
49	.50	.70	1.50	2.60	.50 (8)	**5.40**
50	.30	.60	00	.30	.20 (8)	4.40

APPENDIX 8

PERCENT ITEMS FLAGGED FOR MODERATE DIF

Item	250/250		250/125			250/25		
n o.	M - H	SIB	M - H	SIB	M - H	M - H % rep. DIF stats not calculated	SIB	SIBTEST % rep. DIF stats not calculated
1	**66.70**	**65.20**	**59.60**	**55.80**	**52.60**	1.00	**44.20**	3 .0 0
2	.90	2.00	.80	3.30	.10	13.00	**8.80**	0.00
3	.10	1.00	.10	1.50	.40	13. 0 0	**11.50**	0.00
4	.90	1.80	1.20	1.60	3.60	13. 0 0	**10.90**	0.00
5	.30	.90	1.20	2.00	1.70	13. 0 0	**11.90**	3.00
6	.10	1.10	.40	1.50	.50	17. 0 0	**17.30**	9.00
7	.90	1.50	.90	2.70	.40	17. 0 0	**8.40**	30.00
8	.30	1.00	.50	1.70	1.40	1 8 .0 0	**19.00**	31.00
9	.40	1.20	.50	1.10	.30	1 9 .0 0	**15.80**	5.00
10	1.30	2.20	2.20	4.80	.70	19.00	**11.00**	24.00
11	.30	1.20	1.30	3.70	1.00	3 8 .0 0	**22.20**	148.00
12	.40	.90	3.00	4.30	.80	38.00	**14.00**	32.00
13	**60.10**	**65.10**	**24.20**	**24.10**	**.50**	3 9 .0 0	**10.10**	37 .00
14	00	.70	3.30	**6.20**	**6.80**	3 9 .0 0	**2 2 .6 0**	19 .00
15	.10	.80	2.70	3.70	.10	39.00	**9.50**	35 .00
16	.40	.80	.50	2.50	.10	39.00	**8.30**	26 .00
17	.60	1.40	3.90	**5.80**	2.50	40 .0 0	**17.50**	27 .00
18	.50	.60	.10	1.50	.20	39.50	**7.90**	26 .00
19	.60	.70	.80	1.60	3.50	40 .0 0	**17.00**	21 .00
20	.60	1.40	1.20	1.90	.30	10 .0 0	**8.20**	35.00
21	.20	.90	.20	2.50	.00	40 .0 0	**9.00**	27 .00
22	.90	1.30	1.60	5.00	.60	43. 0 0	**14.50**	10.00
23	.80	1.30	1.50	2.50	.20	45.00	**11.20**	82 .00
24	.90	.90	.50	1.20	.60	49. 0 0	**4.80**	13. 0 0
25	1.20	1.70	2.80	**5.70**	.20	49. 0 0	**10.50**	30 .00
26	.20	1.00	.20	1.70	.40	49. 0 0	**9.40**	29 .00
27	.60	1.60	2.40	7.40	.10	5 1 .0 0	**8.90**	57 .00
28	.50	1.40	.60	2.30	.90	5 1 .0 0	**8.60**	7.00
29	.70	2.50	.90	2.60	.20	1 .0 0	**9.30**	30 .00
30	.40	1.40	.60	1.70	.40	51.00	**10.10**	32 .00
31	1.00	2.00	.80	2.50	.40	51.00	**10.10**	28 .00
32	.50	.70	.10	1.90	.30	5 2 .0 0	**13.80**	61 .00
33	.90	1.50	.80	1.80	.10	54.00	**9.10**	27 .00
34	.70	1.20	.40	1.70	.70	52. 0 0	**16.20**	4. 0 0
35	1.10	1.20	1.20	1.90	.00	52. 0 0	**8.50**	27 .00
36	.30	.90	.50	2.60	.10	52. 0 0	**9.20**	46 .00
37	1.10	3.10	.50	4.30	.10	63. 0 0	**1.90**	27.00
38	.20	.30	.10	.80	.20.	6 4 .0 0	**11.70**	3. 0 0
39	.10	.30	.70	3.00	.10	6 4 .0 0	**7.10**	36 .00
40	.70	.80	.30	1.20	.10	6 4 .0 0	**8.20**	4.00
41	.50	1.10	.20	1.60	.10	6 4 .0 0	**6.40**	3. 0 0
42	.30	1.30	.80	2.40	.50	6 4 .0 0	**10.20**	42 .00
43	.20	.80	1.7	2.60	.10	64. 0 0	**9.90**	21 .00
44	.80	.40	.30	1.20	.90	64. 0 0	**19.70**	44 .00
45	.80	1.60	1.50	2.20	1.10	64. 0 0	**12.50**	24 .00
46	**38.00**	**28.10**	**33.90**	**35.10**	**.20**	65. 0 0	**1 2 .5 0**	7. 0 0
47	.80	1.70	.70	2.60	.10	65. 0 0	**10.60**	28 .00
48	.90	1.10	1.70	4.40	.60	6 6 .0 0	**11.90**	29 .00
49	.40	1.40	.60	1.50	2.30	6 6 .0 0	**13.90**	29 .00
50	.40	.70	2.50	3.20	1.70	6 6 .0 0	**16.90**	3 .0 0