Item Discrimination and Type I Error Rates in DIF Detection Using the Mantel-Haenszel and Logistic Regression Procedures

A dissertation presented to

the faculty of

The Patton College of Education

of Ohio University

In partial fulfillment

of the requirements for the degree

Doctor of Philosophy

Yanju Li

August 2012

© 2012 Yanju Li. All Rights Reserved.
This dissertation titled
Item Discrimination and Type I Error Rates in DIF Detection Using the Mantel-Haenszel
and Logistic Regression Procedures

by

YANJU LI

has been approved for
the Department of Educational Studies
and The Patton College of Education by

______________________________

Gordon P. Brooks
Associated Professor of Educational Studies

______________________________

Renée A. Middleton
Dean, The Patton College of Education
Abstract

LI, YANJU, Ph.D., August 2012, Educational Research and Evaluation

Item Discrimination and Type I Error Rates in DIF Detection Using the Mantel-Haenszel and Logistic Regression Procedures (132 pp.)

Director of Dissertation: Gordon P. Brooks

The inflation of Type I error rates can have damaging effects in DIF identification. This study primarily aimed to examine the performance of the Type I error rates in DIF analysis when using the Mantel-Haenszel (MH) and logistic regression (LR) procedures by simulating data based on two-parameter logistic (2PL) and three-parameter logistic (3PL) item response theory (IRT) models. Specifically, the focus of this study was to explore how item discrimination parameters affect the Type I error rates in both MH and LR procedures when other influencing factors such as, sample size, group mean difference, and matching method were manipulated. Several Monte Carlo simulation studies were conducted. The patterns of the false rejection rates under various conditions were displayed and the effects of influencing factors were evaluated.

The findings suggested that under thin matching, a small range of discrimination parameters for all items resulted in very little Type I error rate inflation for both MH and LR procedures, even with large sample sizes and large group mean differences. The results also indicated that when all items have relatively high discrimination parameters, there is less Type I error inflation regardless of the range of discrimination parameters for all items when using thin matching and deciles thick matching. Additionally, for the condition where the non-studied items did not include weak items, the false rejection
rates were controlled fairly well when the studied item had a relatively larger discrimination value. When data were generated with a 3PL IRT model, the results confirmed that guessing was a nuisance determinant on the inflation of Type I error rates. This study also concluded that thin matching was preferable in controlling Type I error rates, deciles thick matching was acceptable in most circumstances, and quintiles thick matching was poor.

Approved: _____________________________________________________________

Gordon P. Brooks
Associate Professor of Educational Studies
Acknowledgments

I would like to acknowledge some important people through the completion of the dissertation. I want to first express my greatest gratitude to my research committee who provide the vital advice and suggestions in the process of the doctorate journey. My advisor, Gordon Brooks, has guided me with insightful comments and constructive recommendations which helped me grow the academic knowledge. He encouraged me participating national conference to expand my academic zone. He has always devoted his time and efforts when I needed it the most and I truly appreciated that.

Dr. George Johanson provides his expertise on my dissertation. He has patiently delivered me detail comments within each chapter and carefully reviewed my work especially helping my professional writing skills. His encouragement and patience has enlightened my world.

I also want to thank Dr. Valerie Conley and Dr. Adah Ward Randolph who are valuable members of my committee. They put their time and efforts to provide instructive comments on my work during their busy schedule. I had broadened my academic work through their lectures.

I also want to thank my parents who have always believed in me the whole time. Their encouragement is always the vital motivation to keep me going. I thank my parents and brother fly over to visit and keep my company for a short period of time, and that healed the homesick.

My good friends, Hsunli Hsu, Qian An, Hui Guan, Lei Huang and Yang Yang gave me companionship and moral support through some tough time. I highly appreciate
their friendship over years. I also need to thank academic cohorts, including upper and lower classmates. We often exchange our thoughts and academic work. They help me to find new ideas and concepts.
Table of Contents

Abstract ............................................................................................................................... 3
Acknowledgments ............................................................................................................... 5
List of Tables .................................................................................................................... 10
List of Figures ................................................................................................................... 11
Chapter 1: Introduction ..................................................................................................... 12
  Background of the Study .............................................................................................. 12
  Purposes of the Study ................................................................................................... 15
  Statement of Problem .................................................................................................... 16
    Research question 1 (RQ1) ....................................................................................... 17
    Research question 2 (RQ2) ....................................................................................... 17
    Research question 3 (RQ3) ....................................................................................... 17
  Significance of the Study .............................................................................................. 17
The Description of Statistical Procedures ....................................................................... 19
  Mantel-Haenszel procedure ...................................................................................... 19
  Logistic-Regression procedure ................................................................................. 21
Monte Carlo Simulation ................................................................................................. 21
Delimitations and Limitations of the Study .................................................................... 23
Definition of Terms ...................................................................................................... 25
  Differential item functioning (DIF) ........................................................................... 25
  Item bias .................................................................................................................... 25
  Reference and focal group ........................................................................................ 26
  Impact ........................................................................................................................ 26
  Classical test theory .................................................................................................. 26
  Item response theory ................................................................................................. 26
  IRT parameters ......................................................................................................... 26
  Uniform DIF and nonuniform DIF ........................................................................... 27
  Thin matching and thick matching ......................................................................... 29
  P-value ...................................................................................................................... 30
List of Tables

Table 1: 2×2 Contingency Table for Mantel-Haenszel Procedure at jth Score Level ....19
Table 2: Type I Error Rates for the MH Procedure with Sample Size of 300 ............67
Table 3: Type I Error Rates for the MH Procedure with Sample of Size 650 ..........67
Table 4: Type I error Rates for the MH Procedure with Sample Size of 1000 .........68
Table 5: Type I Error Rates for the LR Procedure with Sample Size of 300 ..........70
Table 6: Type I Error Rates for the LR Procedure with Sample Size of 650 ..........71
Table 7: Type I Error Rates for the LR Procedure with Sample Size of 1000 .......71
Table 8: Type I Error Rates for the MH Procedure with Varied Ranges of Discrimination Parameters for All Items .................................................................................................75
Table 9: Type I Error Rates for the LR Procedure with Varied Ranges of Discrimination Parameters for All Items .................................................................................................81
Table 10: Type I Error Rates When Non-studied Items Include Low Discrimination Parameters in the 2PL IRT Model .................................................................84
Table 11: Type I Error Rates When Non-studied Items Exclude Low Discrimination Parameters in the 2PL IRT Model .................................................................87
Table 12: Type I Error Rates When Non-studied Items Include Low Discrimination Parameters in the 3PL IRT Model .................................................................89
Table 13: Type I Error Rates When Non-studied Items Exclude Low Discrimination Parameters in the 3PL IRT Model .................................................................91
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Uniform DIF</td>
<td>28</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Nonuniform DIF</td>
<td>29</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Type I error rates for the MH procedure</td>
<td>69</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Type I error rates for the LR procedure</td>
<td>72</td>
</tr>
<tr>
<td>Figure 5</td>
<td>The MH odds ratio with impact of 0.0</td>
<td>78</td>
</tr>
<tr>
<td>Figure 6</td>
<td>The MH odds ratio with impact of 1.0</td>
<td>79</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Type I error rates when non-studied items include low discrimination parameters in the 2PL IRT model</td>
<td>85</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Type I error rates when non-studied items exclude low discrimination parameters in the 2PL IRT model</td>
<td>88</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Type I error rates when non-studied items include low discrimination parameters in the 3PL IRT model</td>
<td>90</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Type I error rates when non-studied items exclude low discrimination parameters in the 3PL IRT model</td>
<td>92</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

Background of the Study

Dating back to the 1960s, the investigation of item bias began to attract the attention of researchers. Early studies on item bias focused on identifying the differences of mean item scores between groups (Angoff, 1982; Cleary & Hilton, 1968). With more emphasis on fairness and validity of testing and assessment, the studies on detecting item bias were booming because scholars realized that item bias had possible risk for the equality and validity of a test (Clauser & Mazor, 1998).

In the 1980s, the original definition of differential item functioning (DIF) based on the framework of item response theory (IRT) model was proposed by Lord (1980) as “If an item has a different item response function for one group than for another, it is clear that the item is biased” (p. 212). Holland and Thayer (1988) pointed out that the term DIF is one of the “more neutral terms” than the term bias because in quite a few examples exhibiting DIF, bias “does not accurately describe the situation” (p. 129). Clauser and Mazor (1998) mentioned that “DIF is a necessary, but not sufficient, condition for item bias” (p. 31). Whitmore and Schumacker (1999) also stated that “Although bias and DIF terms are used interchangeably, differential item functioning is more meaningful and neutral terminology than item bias when statistical properties are studied” (p. 911). After addressing the conception of DIF, considerable DIF analyses have been conducted in a variety of fields. For instance, 676 entries were found when using differential item functioning as key words solely searching on Education Resources Information Center (ERIC) electronic database.
Conceptually, if a studied item in a test displays a completely identical item response function (IRF) in different groups, then examinees who are members of these different groups will have the same chance to answer this item correctly if they have the same proficiency (e.g., Lord, 1980). DIF occurs when examinees in two groups, traditionally defined as reference group and focal group, exhibit different performance on the studied item after controlling for their overall skills (Dorans & Holland, 1993; Narayanan & Swaminathan, 1996). The total raw score is often used to measure examinee’s overall skill; this is referred to as ‘thin’ matching. In some applications, score groups are used to measure overall skill; this is referred to as ‘thick’ matching. Thick matching is often necessary because several DIF researchers reported that thin matching does not work well at extremes, especially when group ability differences exist (e.g., Donoghue & Allen, 1993). Item performance is the proportion of correct responses on the studied item.

DIF detection has a role in both theoretical research and applied testing. A large number of researchers have focused on exploring various techniques about the detection of DIF, ranging from traditional simple statistical techniques to more modern complex procedures. For example, in 1980, item response theory was proposed by Lord (1980) in order to detect item bias. Additional procedures include the Mantel-Haenszel (MH) common odds ratio (Dorans & Holland, 1993; Holland & Thayer, 1988; Mantel & Haenszel, 1959), logistic regression (LR) (Swaminathan & Rogers, 1990), simultaneous item bias test (SIBTEST) (Shealy & Stout, 1993), methods based on analysis of variance (ANOVA) (Whitmore & Schumacker, 1999) and multi-level modeling (Luppescu, 2002).
Among these approaches, the MH and LR procedures are widely used in DIF detection because of their several advantages. The MH procedure is applied through evaluating the ratio of the probability that reference group examinees respond to the studied item successfully to the probability that focal group examinees respond to the studied item successfully, based on a series of 2×2 contingency tables (Holland & Thayer, 1988; Mantel & Haenszel, 1959). Narayanan and Swaminathan (1996) stated that the MH procedure was affordable to access, simple to calculate, and less strict on sample size requirements. Moreover, the MH procedure provides an associated significance test (Swaminathan & Rogers, 1990). The LR procedure is used to examine DIF by estimating the odds of answering the given item correctly on the basis of a LR model which includes several variables such as total score, group membership, interaction between ability and group membership and a variety of corresponding coefficients (Swaminathan & Rogers, 1990). The LR can be considered as “an extension of the MH procedure that is effective in detecting both uniform and nonuniform DIF” (Rogers & Swaminathan, 1993, p. 105). A main advantage of LR is that “it is a model-based procedure with the ability variable treated as continuous” (Narayanan & Swaminathan, 1996, p. 257-258). It can also be easily applied (Swaminathan & Rogers, 1990).

In recent studies of DIF investigation involving the MH and LR procedures, researchers have noticed that Type I error inflation tends to be a big concern in DIF analysis using the MH and LR procedures in some situations (e. g., DeMars, 2010; Uttaro & Millsap, 1994; Zwick, 1990). As a matter of fact, inflated Type I error can have damaging effects. For instance, Roussos and Stout (1996) summarized three main
negative effects on testing if a high Type I error rate existed in DIF investigation. First of all, inflated Type I error would lead test developers to remove some good items, which results in the waste of human and financial resources. Second, the inflation of Type I error might prevent researchers from exploring the primary causes of DIF. Roussos and Stout explained that “Inflated Type I error is one plausible explanation of why testing companies have to date had little success in explaining why so many seemingly innocuous items have been statistically flagged as displaying DIF” (p. 215).

Consequently, the progress of DIF investigation might be impacted if scholars simply imputed DIF occurrence to inflated Type I error. Third, items with a high discrimination parameter often contain more information but have a higher possibility of being rejected falsely. More informative items might be discarded because of high Type I error rates, which could alter the nature of the test. Previous studies showed that several factors influenced the inflation of Type I error rates. These include test length, sample size, item indices, and group mean differences, which is also termed impact (DeMars 2010; Roussos & Stout, 1996). Currently, more and more studies are focused on the description, explanation and demonstration of inflated Type I error rates in DIF detection because several issues related to DIF remain largely unresolved.

**Purposes of the Study**

One purpose of this study was to display the patterns of Type I error rates of DIF detection using the MH and LR procedures through conducting Monte Carlo simulations. In this study, Type I error rates were examined across group mean ability differences varying from zero to one standard deviation using two-parameter logistic (2PL) and
three-parameter logistic (3PL) IRT models to simulate data. The second purpose was to more fully understand the Type I error rates in 2PL model data and under varying item discriminations. The focus of this study is on how the item discrimination parameter and group mean ability difference combine to impact the inflation of Type I error rates in the MH and LR procedures. Finally, this study was developed to provide DIF researchers and practitioners with guidance when using the MH and LR procedures.

**Statement of Problem**

Factors that impact DIF detection have been discussed in earlier studies of DIF analysis (e.g., Güler & Penfield, 2009; Jodoin & Gierl, 2001; Mazor, Clauser, & Hambleton, 1992; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Sample size, test length, item discrimination, and the distribution of ability differences were considered significant factors that can have effects on DIF detection. The issue of inflated Type I error rates has also attracted researchers’ attention when the MH and LR procedures were applied to assess DIF, especially when group mean ability differences exist (DeMars, 2009; Meredith & Millsap, 1992; Roussos & Stout, 1996; Uttaro & Millsap, 1994; Zwick, 1990).

The current study aimed to investigate the patterns of Type I error rates in DIF analysis when using the MH and LR procedures based on simulated data from 2PL and 3PL IRT models. The final problem involved in this study was how item discrimination impacted Type I error rates when sample size, thick matching, and group mean abilities varied. Several Monte Carlo simulations were conducted to obtain the Type I error rates
in different combinations of situations. Three primary research questions are more specifically as follows:

**Research question 1 (RQ1).** What is the pattern of the Type I error rates displayed in the MH and LR procedures under thin matching, deciles thick matching, and quintiles thick matching with different sample sizes and varied group ability differences with data generated based on 2PL IRT models?

**Research question 2 (RQ2).** What is the pattern of the Type I error rates for the MH and LR procedures displayed when the range of discrimination parameters for all items varies from very small to relatively large under large sample size (e.g., 1,000) and large group ability difference (e.g., $\mu_d=1$) using thin matching, deciles thick matching, and quintiles thick matching with data simulated according to 2PL IRT models?

**Research question 3 (RQ3).** What is the pattern of the Type I error rates for the MH and LR procedures displayed when the discrimination parameter for the studied item varies systematically from low (e.g., 0.2) to high (e.g., 2.0) and the non-studied items include and exclude items with low discrimination parameters under large sample size (e.g., 1,000) and large group ability difference (e.g., $\mu_d=1$) using thin matching and deciles thick matching where data are simulated using both 2PL and 3PL IRT models?

**Significance of the Study**

The study of DIF is an important area for researchers and test developers. A variety of statistical procedures have been proposed for identifying DIF, however, among these techniques, the MH and LR procedures are two of the most often used methods. Swaminathan and Rogers (1990) stated that “The Mantel-Haenszel procedure is
particularly attractive because it is easy to implement and has an associated test of significance” (p. 361). The LR procedure is widely used because it is sensitive to both uniform and nonuniform DIF (Rogers & Swaminathan, 1993). Therefore, it is useful to examine issues such as the inflation of Type I error when the MH and LR procedures are used in the detection of DIF.

In early studies, researchers put great effort into exploring the factors which could impact DIF detection rates. Sample size, length of test, group mean difference, DIF effect size and item discrimination are regarded as chief elements (Narayanan & Swaminathan, 1996). Fixed high discriminations for items were chosen in many previous research studies because researchers believed that higher discriminations for items could result in larger Type I error inflation (e.g., DeMars, 2009). However, no previous studies illustrating how item discrimination impacts Type I error rates when the other factors varied were uncovered. In other words, the influence of item discrimination has been considered, but has not been studied and reported systematically by previous scholars. Consequently, more research on item discrimination and Type I error rates using the MH and LR procedures is necessary.

Thin matching, deciles thick matching, and quintiles thick matching are used in this research. Donoghue and Allen (1993) stated that only two published studies focused on the effects of thin and thick matching at that time. They also proposed that “The use of total score as the matching variable (thin matching) for the MH is a natural choice, but it is also somewhat arbitrary” (p. 134). The use of thick matching may well have an effect on the Type I error rates. Until now, no studies specifically explored how the item
discrimination parameters affect the Type I error rates using the MH and LR procedures under deciles and quintiles thick matching, which will be examined in this study.

The Description of Statistical Procedures

Mantel-Haenszel procedure. Based on a series of $2 \times 2$ contingency tables, the MH common-odds ratio procedure (Holland & Thayer, 1988; Mantel & Haenszel, 1959) is used to estimate the ratio of the probability of correct answers for reference group members to the probability of correct answers for the focal group members. Total scores for examinees, representing item performance, are divided into $j$ score levels. At each score level, a $2 \times 2$ contingency table is created according to the number of examinees who succeed and fail in answering the studied item correctly across given groups. The contingency table for the MH procedure at $j^{th}$ score level is displayed in Table 1.

Table 1

<table>
<thead>
<tr>
<th>Group</th>
<th>Success</th>
<th>Failure</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>$R_{1j}$</td>
<td>$R_{0j}$</td>
<td>$N_{rj}$</td>
</tr>
<tr>
<td>Focal</td>
<td>$F_{1j}$</td>
<td>$F_{0j}$</td>
<td>$N_{fj}$</td>
</tr>
<tr>
<td>Total</td>
<td>$T_{1j}$</td>
<td>$T_{0j}$</td>
<td>$T_{j}$</td>
</tr>
</tbody>
</table>

Note. $R_{1j}$ and $R_{0j}$ denote the number of reference group members in the $j^{th}$ score level who have correct and incorrect responses for the studied item, respectively; $F_{1j}$ and $F_{0j}$ denote the number of focal group members in the $j^{th}$ score level who have correct and incorrect responses for the studied item, respectively; $T_{1j}$ and $T_{0j}$ denote the total number of both reference and focal group members in the $j^{th}$ score level who have correct and incorrect responses for the studied item, respectively; $N_{rj}$ and $N_{fj}$ denote the total number of reference and focal group members in the $j^{th}$ score level, respectively; $T_{j}$ indicates the total number of reference and focal group members in the $j^{th}$ score level.
The pooled odds ratio can be calculated as the odds that reference group members answer the given item correctly to the odds that focal group members answer the given item correctly. Using the same notation provided in Table 1, the equation used to compute the MH odds ratio, is given by

$$\alpha_{MH} = \frac{\sum_j R_{ij} F_{ij} / T_j}{\sum_j R_{ij} F_{ij} / T_j}$$  \hspace{1cm} (1)

The null hypothesis for the MH statistic is $\alpha_{MH} = 1$, which means the given item will favor reference group, which is the group with higher ability, if $\alpha_{MH} > 1$ and the given item will favor focal group, which is the group with lower ability, if $\alpha_{MH} < 1$ (Dorans, 1989). Moreover, the MH statistic test follows a $\chi^2$ distribution with 1 degree of freedom.

The calculation of $\chi^2$ test is given by

$$\chi^2_{MH} = \frac{\left(\sum_j R_{ij} - \sum_j E(R_{ij}) - 0.5\right)^2}{\sum_j \text{Var}(R_{ij})}$$  \hspace{1cm} (2)

where $R_{ij}$ denotes the number of reference group members in the jth score level who have answered the studied item correctly and where $E(R_{ij})$ and $\text{Var}(R_{ij})$ indicate the expectation and the variance of $R_{ij}$, respectively, given by

$$E(R_{ij}) = \frac{N_{ij} T_{ij}}{T_j}$$  \hspace{1cm} (3)

$$\text{Var}(R_{ij}) = \frac{N_{ij} N_{ij} T_{ij} T_{ij}}{(T_j)^2 (T_j - 1)}$$  \hspace{1cm} (4)

The common odds ratio estimate can be transformed into an index under the delta scale ($\hat{\Delta}_{MH}$), typically used at Educational Testing Service (ETS) to measure the performance of DIF (Zwick & Ercikan, 1989).

$$\hat{\Delta}_{MH} = -2.35 \ln(\alpha_{MH})$$  \hspace{1cm} (5)
The null hypothesis for the Delta-MH statistic is $\Delta_{\text{MH}} = 0$. Items with a negative $\Delta_{\text{MH}}$ will favor the reference group, while items with a positive $\Delta_{\text{MH}}$ will favor the focal group (Zwick & Ercikan, 1989).

**Logistic-Regression procedure.** The probability of responding to the studied item correctly using the LR model can be expressed as

$$ p(Y = 1|t, g) = \frac{e^{\beta_0 + \beta_1 t + \beta_2 g + \beta_3 tg}}{1 + e^{\beta_0 + \beta_1 t + \beta_2 g + \beta_3 tg}} \quad (6) $$

where $p(Y = 1)$ is the probability of getting a correct response for the given item, $t$ is the observed ability of a respondent, measured by total score, $g$ is reference and focal group membership, traditionally coded as 1 and 0, $tg$ is the interaction between ability and group membership, $\beta_0$, $\beta_1$, $\beta_2$ and $\beta_3$ are the intercept and the coefficients of the ability, group membership and interaction terms, respectively.

The LR procedure is regarded as an effective method in the detection of DIF because it is sensitive to both uniform DIF and nonuniform DIF. In terms of the above formulation, the null hypothesis of no DIF holds, if both $\beta_2 = 0$ and $\beta_3 = 0$; uniform DIF shows if $\beta_2 \neq 0$ and $\beta_3 = 0$; nonuniform DIF occurs if $\beta_3 \neq 0$ independent of the value of $\beta_2$ (Güler & Penfield, 2009; Jodoin & Gierl, 2001; Swaminathan & Rogers, 1990).

**Monte Carlo Simulation**

Mathematical analyses are commonly used to provide closed-form or analytical solutions to a variety of research questions; however, they are not feasible in all conditions. For instance, Mooney (1997) stated that there are three particular circumstances where a mathematical analysis should be abandoned. They are if (a)
statistical assumption violations exist; (b) conditions that mathematical theories require are not able to be met, for instance, the null hypothesis is not true; and (c) the sampling distribution is unknown for a statistic. For many statistical issues, certain assumptions are needed. Researchers may worry about the validity of statistical inference when assumptions are not held. Fortunately, Monte Carlo simulation studies allow researchers to assess relevant risks if they continue their research even when assumption violations exist.

Mooney (1997) proposed that Monte Carlo simulation provides researchers with “an alternative to analytical mathematics for understanding a statistic’s sampling distribution and evaluating its behavior in random samples” (p. 2). Using Monte Carlo simulation, researchers can simulate the random sampling of data through a large number of replications with the aid of specific computer software rather than actually collect samples from a real population (Fan, Felsővályi, Stephen & Keenan, 2002). Monte Carlo methods can be used to solve mathematically intractable problems, such as “when the sampling distribution is unknown or the null hypothesis is not true” (Brooks, 1998, p. 125). Fan et al. (2002) also stated that conducting a Monte Carlo study would be a good choice when the assumptions of a statistical theory may not hold, or when statistical theory is either weak or nonexistent.

Additionally, data is an essential part in DIF investigation. There are two types of data: real data and simulated data. Some early scholars examined DIF by collecting real data through questionnaires. For example, a test question about a weight of newborn human baby was constructed by Driana (2007) in her dissertation, aiming to investigate
gender DIF based for ninth-grade students. The question is: “A newborn human baby would probably weigh” and the four options are: “8 tons”, “8 ounces”, “8 pounds” and “8 cubic centimeters” (p. 183). Driana (2007) reported that one reason this item showed large DIF might be because girls had more knowledge about the characteristics of newborn babies than boys.

Nevertheless, in contrast to gathering real data, Monte Carlo simulation is regarded as an effective and convenient way to obtain data for DIF analysis. Simulating replicated random sampling of data with the aid of computer program is the core of the Monte Carlo method (Brooks, 1998). Monte Carlo simulation is especially appropriate for solving some mathematically intractable problems (Brooks, 1998). Previous studies showed that data simulated through Monte Carlo methods have been widely used by DIF researchers.

**Delimitations and Limitations of the Study**

The current study was delimited in that it primarily examined the effect of the item discrimination parameter on the performance of the Type I error rates under various conditions when using the MH and LR procedures. There are several other specific delimitations and limitations in this research.

First, in this study, discrimination parameters for items were designed from 0.0 to 2.0 and item difficulty parameters were set from -2.0 to 2.0. That is to say, only the effect of a certain range of item parameters was investigated. Additionally, 50 items were generated and examined, which restricted this research to a medium test length. Moreover, previous researchers believed that the percentage of test items with DIF, also
called the level of DIF contamination, had an effect on the performance of the Type I error. However, no DIF items were considered in this study.

Second, three matching methods, thin matching, deciles thick matching, and quintiles thick matching, are discussed and employed in this study. Donoghue and Allen (1993) studied seven forms of matching variable including “Thin matching variable,” “Equal interval matching variable,” “Percent of total sample matching variable,” “Percent of focal sample matching variable,” and three types of “Censored matching variable,” three types of “Minimum cell frequency matching variable,” and “No matching variable” (p. 136). The authors also compared the performance of each of these matching methods. The results of the current study apply only to the three matching methods used, but little research has been reported on these types of thick matching, which may provide some useful information.

Third, there are a variety of methods developed for detecting DIF including both traditional and modern techniques. However, only two of them, the MH and LR procedures, are examined in this study. Because they are regarded as two of the most popular techniques in DIF analysis (Narayanan & Swaminathan, 1996), the choice to focus on these two methods was reasonable.

Fourth, as indicated in the literature, a large number of factors affect the inflation of Type I error rates in DIF detection (e.g., DeMars, 2009; & Herrera & Gómez, 2008; Roussos & Stout, 1996). Nevertheless, only several of these factors are examined in this study. Besides the item discrimination parameter, mean ability difference, sample size and matching method are considered in the present study. Other factors also affect DIF
detection, such as test length, type of DIF (uniform or nonuniform), percentage of items with DIF, and the magnitude of DIF (Herrera & Gómez, 2008).

Finally, this research is based on Monte Carlo simulations, which may limit the generalizability of the results. For example, the distributions of the item discrimination parameters and difficulty parameters follow uniform distributions and the examinees’ response data will be generated based on particular IRT models. The test length, the sample size and the generation of item parameters will be selected so that current results are comparable to results from prior research. Although the Monte Carlo simulation is necessary in DIF studies and it is effective to track the correctness of the Type I error rates, the choices made in this study are based on previous research; however, all such choices are somewhat arbitrary and the simulated data may not adequately represent all data from actual examinees.

Definition of Terms

Differential item functioning (DIF). DIF, as mentioned above, is said to exist when the group members with identical latent traits have different probabilities of answering the studied item correctly (Angoff, 1993; Dorans & Holland, 1993, Narayanan & Swaminathan, 1996).

Item bias. “When a test item unfairly favors one group over another, it can be said to be biased” (Clauser & Mazor, 1998, p. 31). Item bias is a nuisance for test validity. The existence of item bias can result in the occurrence of DIF; however, the presence of DIF does not imply that an item is biased.
Reference and focal group. The reference group and focal group are traditionally referred to as a major and minor subgroup in DIF analysis, respectively. For example, males could be treated as a reference group and females could be a focal group when we conduct a DIF analysis across gender.

Impact. The ability mean difference between reference group and focal group in DIF studies is termed as impact. Commonly, the ability mean for reference group is set to 0. If the ability mean of focal group is not equal to 0, then impact exists. Most previous studies set impact to 1 standard deviation unit (e.g., Narayanan & Swaminathan, 1996; Güler & Penfield, 2009).

Classical test theory. Classical Test Theory (CTT) is a traditional method in measurement. It is test oriented. “CTT is based on the true score model. This model relates the individual’s observed score to his or her location on the latent variable” (De Ayala, 2009, p. 5). CTT indicates that any observed test score could be written as the sum of a true score and random error (Crocker & Algina, 1986).

Item response theory. Item Response Theory (IRT), based on an item-level model, permits a researcher to evaluate the psychometric characteristics of assessments or tests by assuming a latent construct underlying an observed score (Drasgow & Hulin, 1990; Embretson & Reise, 2000). The probability of responding to an item correctly is written as a function of the latent traits of examinees and the characteristics of items (Hambleton, Swaminathan & Rogers, 1991).

IRT parameters. There are three parameters in IRT models that need to be discussed. The item discrimination parameter is denoted by $a_i$. It exhibits the ability of an
item to separate examinees into different ability levels. That is, an item with a larger $a_t$ parameter has more capacity to distinguish higher ability examinees from lower ability examinees. The item difficulty parameter, denoted by $b_t$, indicates how difficult the item is. In general, the higher the parameter, the more difficult the item. For instance, Herrera and Gómez (2008) defined three levels of difficulty parameter in their study as low ($b_t \leq -1.5$), moderate ($-1.5 \leq b_t \leq 1.5$) and high ($b_t \geq 1.5$). They also stated two levels of discrimination parameter: low, $a_t < 0.5$, and moderate, $0.5 \leq a_t \leq 1$ (which implies that a parameter $a_t$ larger than 1 is considered high). The pseudo-chance or guessing parameter, denoted as $c_t$, stands for the probability that an examinee will get an item correct only by guessing. Using this parameter means that an examinee will have a certain chance of answering an item successfully even when the item’s difficulty is well beyond the examinee’s ability.

**Uniform DIF and nonuniform DIF.** There are two types of DIF: uniform DIF and nonuniform DIF (Mellenberg, 1982). “Uniform DIF exists when there is no interaction between ability level and the group membership on studied item performance” (Güler & Penfield, 2009, p. 315). That is, the studied item favors one group across the entire ability continuum. On the contrary, nonuniform DIF occurs when there is an interaction. In other words, if the studied item favors one group at one ability range but favors the other group at another ability range, nonuniform DIF exists. Based on the framework of Item Response Theory (IRT), if two comparable groups have equal discrimination parameters ($a$) and equal difficulty parameters ($b$), then no DIF exists. If an identical parameter $a$ and a distinct parameter $b$ are set for the two groups, it
represents uniform DIF. If there is an unequal parameter $a$ for two groups, it indicates the situation of nonuniform DIF no matter whether parameter $b$ is equal or not. Figure 1 and Figure 2, which were created by R software, illustrate uniform DIF and nonuniform DIF, respectively.

*Figure 1. Uniform DIF.*
Figure 2. Nonuniform DIF.

**Thin matching and thick matching.** Thin matching is when the total score is used as a matching variable. For instance, 50 items will form 51 scoring levels for thin matching. Based on the examinees’ total scores, they are assigned to those 51 scoring levels. The number of examinees in each level is not always equal for thin matching. Pooling total score levels to form homogeneous groups of the matching variable is regarded as thick matching (Donoghue & Allen, 1993). Deciles and Quintiles are two common types of thick matching. When ten score levels are created by pooling total score and these ten levels are used as the matching variable, deciles thick matching is conducted. Quintiles thick matching is when five scoring levels are formed by pooling total score and used as matching criterion. The number of examinees for each score group is approximately identical in these two thick matching methods. Generally speaking, as
Clauser and Mazor (1998) indicated, in thin matching more score levels are formed and fewer examinees are involved in each level. In contrast to thin matching, thick matching results in fewer score levels with more examinees contained in each level.

**P-value.** The $p$-value is an important indicator of statistical significance. It is convenient and effective for researchers to use $p$ values in investigating statistical inference issues (Hochberg & Benjamini, 1990).

**Type I error.** Stevens (1999) stated that Type I error rate is the probability of wrongly rejecting the true null hypothesis. In DIF analysis, “a Type I error is the incorrect identification of an item as displaying DIF when, in fact, it does not” (Jodoin & Gierl, 2001, p. 330). According to Aron and Aron (1997), the chance of making a Type I error is related to the significance level researchers set. The nominal significance level is .05.

**Organization of the Study**

This dissertation is organized into five Chapters. Chapter one contains the background and purposes of this study, problem statement, research questions and the significance, delimitations and limitations of this study and definition of terms. Chapter two focuses on review of literature related to this study. Chapter three covers the research design, Monte Carlo simulation, data generation, and data analysis procedures. Chapter four is devoted to results and results interpretation. Chapter five consists of discussion, conclusions, and recommendations. Furthermore, references and appendices are provided after the five chapters.
Chapter 2: Review of Literature

The significance of DIF analysis has been emphasized in many areas. In particular, “The detection of item bias, or differential item functioning (DIF), in achievement, licensure, and credentialing examinations has become an important issue in recent years” (Swaminathan & Rogers, 1990, p. 361). One important issue in DIF detection is the accuracy of the detecting rate, that is, the control of Type I error and statistical power. Paek (2010) claimed that it is necessary to conduct statistical testing in order that test developers can make correct decisions about removal of items in actual testing situations. As we know, when the null hypothesis is true (that is, there is no DIF), the accuracy of DIF detection mainly focuses on controlling for the inflation of Type I error. In contrast, the statistical power is investigated when DIF exists.

Review of Influencing Factors in DIF Detection

In recent decades, many previous studies have aimed to identify the main factors influencing detection accuracy (Güler & Penfield, 2009; Jodoin & Gierl, 2001; Mazor et al., 1992; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Whitmore & Schumacker, 1999). Herrera and Gómez (2008) summarized that “The most widely studied factors include the number of levels of attribute size in the focal and reference groups, the percentage of test items with DIF, the type and magnitude of DIF of the items studied, and the sizes of the groups compared” (p. 741). In fact, other factors such as test length and matching method also have an effect on the accuracy of detection rate.
Sample size is a significant component which is able to affect the DIF detection rate. Early simulation studies have shown that both the power and the inflation of Type I error is likely to increase as the sample size increases (e.g., DeMars, 2009; Roussos & Stout, 1996; Swaminathan & Rogers, 1990). The minimal number of examinees suggested by Wright and Stone (1979) was 200 total members when using the one-parameter Rasch model. Güler and Penfield (2009) noted that at least 200 to 250 examinees in each group are required to obtain enough power for the MH and LR procedure.

Group mean difference is a factor that needs attention in a DIF study. It is known that differences in the group ability distributions will have an effect on DIF detection (Mazor et al., 1992; Shealy & Stout, 1993). As Güler and Penfield (2009) stated, “The investigation of both equal and unequal ability distributions is of critical importance because of the known Type I error inflation of many DIF statistics when ability distributions differ” (p. 322). Group mean difference is also termed impact in previous studies. Roussos and Stout (1996) stated that the issue of inflated Type I error caused by impact should be addressed successfully in order to obtain appropriate statistics. Otherwise, the occurrence of impact may result in a remarkable false positive error rate even when no DIF exists.

The magnitude of the item discrimination parameter can have a significant effect on DIF detection performance. In previous studies, researchers manipulated item discrimination parameters to explore their effects on DIF detection. For example,
DeMars (2010) illustrated items with higher discrimination parameters generally resulted in more reliable scores.

Test length is a factor of interest because researchers believed that test length could affect the accuracy of total score. “The longer the test, the more reliable the total score” (Rogers & Swaminathan, 1993, p. 107). As Narayanan and Swaminathan (1996) stated, “Test length can have an impact on DIF detection rates because a longer test is likely to produce more reliable scores and, hence, more accurate ability estimates” (p. 263). In addition, Whitmore and Schumacker (1999) pointed out that test length should be considered because longer tests produced more reliable scores, when everything else was held constant. It is suggested by Wright and Stone (1979) that the minimum test length is 20 items in a Rasch model.

When the factors provided above are manipulated under different conditions, inflated Type I error may occur and vary. Roussos and Stout (1996) claimed that, “MH has potential Type I error problems when the number of examinees is small, the test is not long, the item response functions are non-Rasch, and/or the reference and focal group observed score distributions display impact” (p. 217).

**Type I Error in the MH Procedure**

A variety of statistical methods have been proposed by scholars for identifying DIF (e.g., Raju, Bode, & Larsen, 1989). “DIF procedures can be categorized into three broad classes: traditional classical test theory methods, chi-square methods, and latent trait theory methods” (Whitmore & Schumacker, 1999, p. 911). Among various techniques, the MH procedure, based on a chi-square statistic, is considered as one of the
most commonly used methods in applied analyses because of its mathematical simplicity, ease of use, and low cost (Herrera & Gómez, 2008). Considerable early research has been conducted to examine the performance of Type I error in the investigation of DIF using the MH statistic.

Holland and Thayer (1988) indicated that inflation of Type I error is not a potential problem in the one-parameter logistic (or Rasch) model for the MH procedure because the MH hypothesis is identical with the null hypothesis that there is no DIF when data follow a Rasch model. It was best explained by Roussos and Stout (1996) as “when a test can be modeled with one-parameter logistic (Rasch model) items, because test number-right score is a sufficient statistic for the test matching trait, the MH null hypothesis is indeed equivalent to a hypothesis of no DIF” (p. 217).

However, Raju et al. (1989) reminded researchers that there was possible risk of yielding inflated Type I error when the MH procedure was selected in DIF investigation because of the existence of numerous chi-square values. Zwick (1990) also claimed that inflated Type I errors might be expected using the MH procedure when group ability difference exists unless data follow a Rasch model or the test scores have very high reliability. In other words, in data where a Rasch model fits poorly, the MH statistic is likely to display inflated Type I error because of the existence of unequal group ability distributions. After Zwick, many researchers paid attention to exploring what kind of factors and how these factors affected the performance of the Type I error in the MH procedure. Although the Rasch model has a capacity to prevent inflated Type I error from occurring, in practice, many data for tests do follow a two-parameter (2PL) or three-
parameter (3PL) logistic IRT model rather than a Rasch model. Consequently, a good number of previous studies were mainly based on non-Rasch models.

Uttaro and Millsap (1994) explored how test length influenced the performance of the Type I error in a three-parameter IRT model. Both 20-item and 40-item tests were included in this study. The authors concluded that test length strongly impacts the Type I error rate. They contended that “For the 20-item test, the MH procedure gave inflated error rates but the inflation of Type I error rates disappeared entirely in the 40-item test” (p. 24). Nevertheless, only 200 replications were used within each cell of design.

An actual auto shop test with 25 items was examined by Roussos and Stout (1996) to investigate the effects of small sample size in the MH procedure. This real data followed a 3PL IRT model. The authors concluded that the MH statistic performed well in adhering to the nominal level of significance for small sample size; 100 examinees were in each group in this study. Roussos and Stout (1996) also examined the MH procedure under various IRT parameterizations. Their results indicated that there was no inflated Type I error when no latent group ability distribution difference was present regardless of sample size. However, inflated Type I error occurred when latent ability distributional differences existed under certain patterns, for instance, the condition of high discrimination and low difficulty.

Narayanan and Swaminathan (1996) simulated data for 40 items in order to investigate the feature of a relatively short test based on a three-parameter IRT model. The authors manipulated five factors: sample size, percent of DIF, group mean difference, DIF effect size, and item parameters. There were three findings about Type I
error rates in the MH procedure: first, they concluded that as sample size increases, the Type I error rates tended to go up as well. Moreover, higher Type I error rates were exhibited when group ability differences existed. For example, the Type I error rate was 4.1% for an equal ability distribution, indicating no inflation, whereas, for unequal ability distribution, it was 5.5% which was a slight inflation. Finally, no inflation was found in the results irrespective of the percent of items containing DIF. That is, the Type I error rates were always near nominal significance level whether 0%, 10% or 20% DIF items were included in tests.

Herrera and Gómez (2008) conducted a Monte Carlo experiment to display the extent to which sample size and item parameters affect the performance of Type I error rates in the MH procedure. Their data were simulated with a 3PL model. The examinees in the reference group were 500 and 1500, while the size of focal group ranged from 100 to 1500. One of their findings disclosed that a large sample size results in higher false rejection rates than small sample does. In fact, “The highest rate of false positives, 8%, was observed for the largest group sizes” (p. 752). Furthermore, three levels of difficulty parameter for the studied item were used: low \((b_i \leq -1.5)\), moderate \((-1.5 \leq b_i \leq 1.5)\) and high \((b_i \geq 1.5)\); two levels of discrimination parameter for the studied item were selected: low \((a_i < 0.5)\) and moderate \((0.5 \leq a_i \leq 1)\). The authors pointed out that “both reference group size and focal group sample size had a significant effect on the MH procedure’s rate of false positives with items of moderate difficulty and those which were of low difficulty and moderately discriminating” (p. 747). All the discrimination
parameters used by Herrera and Gómez (2008) were below 1, therefore, the examination of high discrimination parameters was not involved.

Another study was conducted by Güler and Penfield (2009) to estimate the Type I error rates for the MH procedure about sample size and group mean difference based on a 3PL IRT model. In this research, they covered small and large sample size and equal and unequal group ability distribution. However, the discrimination parameter of the studied item was fixed at 1. Under this condition, the authors found that the MH statistic maintained Type I error rates within the nominal level of significance level independent of sample size and equality of group ability distribution. This result might be due to the fixed discrimination parameter for the studied item.

DeMars (2009) carried out a series of Monte Carlo simulations to display the patterns of the Type I error rates for the MH procedure under different conditions with varied sample size, test length and group mean difference. She used 20-item, 40-item and 60-item tests. Data followed a 3PL IRT model. Three sample sizes were selected to be 250, 1,000 and 2,000 members in each group. Three test lengths across five levels of group mean differences under different sample sizes were examined and Type I error rates were graphed for several DIF procedures including the MH procedure. However, all item discrimination parameters were relatively large because DeMars believed that the inflation of Type I error was higher with larger discriminations. The results shown in this study illustrated that large sample size, short test, and high group mean ability differences tended to result in inflated Type I error. For example, inflated Type I error rates tends to increase as the sample size goes up when other factors are identical. The false rejection
rate becomes less in 60-item test comparing to 20-item tests if holding sample size and group mean difference.

DeMars (2010) also claimed that there would be Type I error inflation when impact existed. The author stated two potential explanations for the occurrence of inflation: the use of fallible, observed scores as the conditioning variable and the impact of guessing.

**Type I Error in the LR Procedure**

In 1987, Bennet, Rock and Kaplan proposed a logistic regression (LR) procedure and Swaminathan and Rogers (1990) first applied it to identify DIF. The LR procedure is particularly attractive since it is sensitive to the investigation of both uniform DIF and nonuniform DIF, as well as being simple to use. The strengths of the LR procedure have been demonstrated by many scholars, however, one disadvantage of the LR procedure is that it is prone to produce inflated Type I error rates (Jodoin & Gierl, 2001). Many previous studies have mentioned and discussed the performance of the Type I error in LR, and some of them have also indicated how the Type I error rate performs differently in the LR procedure compared to the MH procedure under a variety of conditions.

Swaminathan and Rogers (1990) mainly examined the power for detecting uniform and nonuniform DIF in both MH and LR procedures, however, the comparison of the Type I error rates for these two methods was also mentioned. They studied six conditions, which were simulated by crossing two levels of sample size (i.e., 250 and 500 members for each group) with three levels of test length corresponding to 40, 60 and 80 items. A 3PL IRT model was used to generate data in this study. Swaminathan and
Rogers concluded that the performance of the false rejection rate, that is, the Type I error rate, for the MH procedure was slightly better than that of the LR procedure. As the authors illustrated, a consistent result, 1% false positive rejection, was displayed when using the MH; while the false rejection rate for the LR varied from 1% to 6% under various conditions.

Narayanan and Swaminathan’s (1996) Monte Carlo study examined the performance of the power and Type I error under both uniform DIF and nonuniform DIF, and compared the MH and LR procedures as well. The design of the study as well as the findings for the Type I error in the MH procedure were reviewed in an early section. Consequently, only specific findings for the LR procedure are given here. The Narayanan and Swaminathan results indicated that sample size had a similar effect on the Type I error rate for the LR as it did for the MH. That is to say, inflation increased with larger sample size. However, the results of this study showed that the inflation of the LR procedure was higher than that of the MH procedure. Furthermore, the authors found that group ability differences tended to result in higher Type I error rates. For instance, using the LR procedure, the Type I error rate was 6.1% under the condition of equal ability distribution, which shows a slight inflation. When groups were of unequal ability, Type I error goes up 9.8%, a relatively high inflation. Compared to the MH, the LR exhibited higher inflation. The final finding was that inflation increased when the proportion of items with DIF increased. For example, 0, 4, and 8 DIF items were used in the study, and the results showed that the Type I error rates for the LR procedure were 7.5%, 8.6% and 8.1%, respectively, indicating inflation in all three conditions.
One study that mainly investigated the effect of sample size, group ability difference, and proportion of items containing DIF on the Type I error in the LR procedure was conducted by Jodoin and Gierl (2001). Forty items were generated based on a 3PL IRT model and three levels of sample size for both reference group and focal group (250, 500 and 1000) were selected. Reference group sample sizes were partially crossed with focal group sample sizes and six combinations were examined. The authors also had an equal mean ability and an unequal mean ability condition for each of the conditions listed above. The proportion of items with DIF was manipulated because the authors believed that “the percentage of DIF items can reduce the validity of the matching variable, it is also expected to affect Type I error and power rates” (p. 337). Three levels of DIF contamination were used: 4 DIF items (10%) and 8 DIF items (20%) favoring the reference group, and 8 DIF items with half favoring reference and the other half favoring focal group. Their results indicated that large sample size and unequal ability distribution between groups tended to result in larger inflated Type I error, regardless of the proportion of DIF items. However, the authors stated that the highest Type I error occurred when both reference and focal group had the largest members and group mean ability difference existed under all 8 DIF items favoring reference group.

Herrera and Gómez’s (2008) study measured the performance of the Type I error in both MH and LR procedures. Both the method of this study and the conclusions for the MH have been reviewed in a previous section. The findings in the LR will now be discussed. The authors pointed out that the LR procedure appeared to show somewhat higher Type I error rates than the MH procedure did under the identical conditions.
Additionally, the results indicated that sample size for focal group did have an appreciable influence on the rate of false rejection in the LR procedure for items with low discrimination that had either low or moderate difficulty. In contrast, one surprising finding suggested that the Type I error rate of the LR procedure was unaffected by reference group size.

Holmes Finch and French (2007) conducted a study with 30-item datasets in order to compare the performance of four DIF detection methods, which included the LR procedure. Four levels of total sample size (500, 750, 1000, and 1500) and three conditions of item contamination (0, 3, or 6 items) were covered in this study. Group ability differences and underlying IRT model (2PL and 3PL) were also examined. The results for Type I error in the LR obtained by Holmes Finch and French are not totally consistent with the previous studies. Holmes Finch and French pointed out that the false rejection rates were controlled within nominal significance level, that is, no inflation occurred under any manipulated conditions. In other words, no manipulated factors seemed to have a significant effect on the Type I error rate when using the LR procedure.

The research conducted by Güler and Penfield (2009) has been discussed previously. They found that the MH procedure performed well because no inflation of Type I error occurred. However, the results showed that inflated Type I error rates were obtained in the LR procedure in the presence of group ability difference. All the false rejection rates were maintained within nominal significance levels under the conditions of equal ability distributions.
The Monte Carlo studies conducted by DeMars (2009) aimed to display how factors such as sample size, test length, and group mean difference affect type I error rates in both MH and LR. The results for the MH reviewed above suggested that the Type I error rates for the MH procedure tended to inflate under the condition of short test, large sample size, and large difference between group means. The performance of the LR procedure was similar to the MH procedure, that is, as sample size and group mean difference expanded, or test length shrunk, the inflation of Type I error rates was expected to increase as well. However, in most situations, the pattern of results showed that the LR yielded slightly higher inflation than the MH did, which was consistent with previous studies.

**Thin Matching and Thick Matching**

Traditionally, researchers tended to select thin matching, that is, defining total score as a matching variable, when using the MH procedure to conduct DIF analysis. However, a large number of scoring levels in thin matching results in relatively fewer examinees from each group who can be assigned to individual levels. This is exacerbated when there is impact. Consequently, when sample size is small and/or group ability difference exists, there might be very few examinees located in some levels, especially those extreme levels, when we use thin matching. A method advocated by previous researchers is to collapse score levels or thicken the matching criteria.

Only a few of the early studies explored the performance of thick matching in DIF assessment. Wright (1986) evaluated the effect of 61 levels (i.e., thin matching) and 6 levels (thick matching) on the common odds ratio estimate indices in delta scale (Δ)
using the MH procedure with verbal SAT data. The results showed that 61 score groups yielded smaller average $\tilde{\Delta}$ than 6 score groups did. The author believed that six categories on matching variable were not enough to obtain a desirable estimation of $\tilde{\Delta}$ in this case. A major limitation of this research is the use of only two matching criteria. Raju et al. (1989) conducted comparisons of 2, 4, 6, 8 and 10 score groups using the MH procedure on a 40-item vocabulary test. They stated that four or more score groups were able to yield stable common odds ratio estimates. Additionally, the advantages and disadvantages of using a thick matching method in DIF analysis were explained by Donoghue and Allen (1993). They stated that using thick matching will ensure the stability of cell frequencies for each matching variable level.

A big concern of employing thick matching, however, is the possibility of confounding DIF with impact, especially when the number of categories is small. Donoghue and Allen (1993) explained that “As the number of levels of the matching variable increases, the MH procedure differs more and more from an analysis of impact” (p. 135). In this study, the authors also designed seven methods for pooling matching variables to identify the performance of the MH procedure. Donoghue and Allen (1993) found that thin matching tended to perform poorly in DIF detection when test length was short (e.g., 5 or 10 items). On the contrary, thin matching performed fairly well under the relatively long tests (e.g., 40 items), particularly with sufficient sample sizes (e.g., 1600 examinees in total). Among the seven thick matching strategies, three of them displayed the best performance. They are “equal interval matching” which means combining pairs of total scores to form levels, “censoring matching variable” that means the extreme
levels having at least one examinee and “minimum frequency matching” which means at least one examinee assigned to individual levels (p. 151). Clauser, Mazor and Hambleton (1994) conducted a simulation to investigate the behavior of the MH procedure through varying total sample size from 200 to 4000 and score group number from 2 to 81. They claimed that a large sample size and a small number of score levels resulted in a high Type I error inflation when group ability difference existed. The Type I error rates were maintained within a nominal significance level when the score group number was equal to or more than 20.

**Item Response Theory**

Item response theory (IRT) is constructed on an item-level framework. It is widely applied to evaluate psychometric properties and solve measurement and statistics issues (Drasgow & Hulin, 1990; Embretson & Reise, 2000; Hambleton et al., 1991). Compared to classical test theory (CTT), IRT provides researchers with more flexibility, as well as providing access to more accurate information to reduce measurement error. Currently, applications of IRT have penetrated into various areas, such as “test development, item banking, differential item functioning, adaptive testing, test equating, and test scaling” (Heh, 2007, p. 13)

In order to investigate bias or DIF successfully, one important thing is to determine an appropriate IRT model. Meredith and Millasp (1992) claimed that “When the chosen model is inadequate, use of the model may lead to incorrect decisions regarding bias, or the lack of bias” (p. 310). IRT models are becoming popular in DIF detection because they are constructed on an item-level rather than a test-level
framework. In all IRT models, the probability of responding to a studied item successfully should be able to be reflected or predicted by the latent trait of examinee and the properties of the item.

**IRT assumptions.** An IRT framework consists of three common assumptions. First, it is assumed that a single unobserved variable or latent trait (e.g., ability) is measured by a series of items in a test. This is referred to as unidimensionality assumption. Second, if an unobservable variable or latent trait has a constant impact on test performance, no interaction occurs between examinees’ responses on different items in a test. In other words, after controlling abilities, examinees’ responses on any pair of items should not be related to each other. This assumption is called local independence. Third, IRT models assume that the relationship between the observable variables (e.g., responses to items) and latent traits (e.g., examinees’ abilities) can be explained by an item response function (IRF). (Hambleton et al., 1991).

**IRT model.** It is possible to build a variety of models based on an IRT framework; however, currently three common logistic models are widely used in DIF analysis. They are referred to as one, two, and three-parameter IRT logistic models (Hambleton et al., 1991). The three-parameter logistic (3PL) model, the most comprehensive model, is commonly employed by researchers. It can be expressed mathematically as follows:

\[
P(U_i = 1|\theta) = c_i + \frac{(1-c_i)e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}} \quad i = 1,2 \ldots k
\]

Here \(P(U_i = 1|\theta)\) represents the probability that an examinee who has the ability \(\theta\) responds to the studied item \(i\) successfully; \(b_i\) is item difficulty parameter, indicating how
hard the item is; $a_i$ is the item discrimination parameter, indicating the ability of the item to discriminate between examinees; $c_i$ is the pseudo-guessing parameter, which stands for a probability of an examinee who provides a correct answer for the item only by guessing; $\theta$ is a person parameter, indicating individual ability; $D$ is a scaling factor, usually assigned to 1.7; $e$ is a transcendental number equaling 2.718 (Holmes Finch & French, 2007; Hambleton et al., 1991).

The discrimination parameter is an index to separate examinees into different ability levels. An item with a higher parameter $a_i$ has more capacity to distinguish higher ability examinees from lower ability examinees. Hambleton and his co-authors (1991) stated that the value of item discrimination parameters was usually restricted to (0, 2) even though theoretically parameter $a_i$ could be defined on the entire scale ($-\infty, +\infty$). A negative parameter $a_i$ should not be considered because it indicates a questionable item. Also, “it is unusual to obtain $a_i$ values larger than 2” (Hambleton et al., 1991, p. 15). Two levels of discrimination parameter were defined in the research of Herrera and Gómez (2008): low ($a_i < 0.5$) and moderate ($0.5 \leq a_i \leq 1$). It can be easily inferred that the authors believed that a parameter $a_i$ larger than 1 is a high item discrimination parameter. The difficulty parameter, as the name implies, is used to reflect the difficulty of an item. Generally, harder items have higher $b_i$ parameters. Hambleton et al. (1991) pointed out that the typical values for item difficulty parameters were set to [-2.0, 2.0]. An item with a parameter $b_i$ close to -2.0 indicates that this item is very easy for examinees. On the contrary, if an item has parameter $b_i$ near 2.0, it is very difficult to answer correctly. In the study of Herrera and Gómez (2008), three categories of
difficulty parameter were stated: low \( b_i \leq -1.5 \), moderate \((-1.5 \leq b_i \leq 1.5)\) and high \( (b_i \geq 1.5)\). Because of the existence of a guessing, an examinee will have a chance to respond to the item correctly even when the item is beyond the examinee’s ability. The value of guessing parameter in most of previous studies involving in three-parameter IRT model are set to 0.2 (e.g., Güler & Penfield, 2009; Narayanan & Swaminathan, 1996).

When parameter \( c_i \) equals zero, that is, no guessing is considered, the 3PL model becomes the two-parameter logistic (2PL) model. The equation for the 2PL model is

\[
P(U_i = 1|\theta) = \frac{e^{da_i (\theta - b_i)}}{1 + e^{da_i (\theta - b_i)}} \quad i = 1,2 \ldots k
\]  

If the discrimination parameter is held constant, the 2PL model will be simplified to the 1PL model (also called Rasch model). In the 1PL model, only one kind of item parameter, that is, parameter \( b_i \) exists. The item characteristic function for 1PL model can be given by the formula

\[
P(U_i = 1|\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad i = 1,2 \ldots k
\]

It is important to note that the difficulty parameter is referred to as the point on the latent trait scale where the probability of getting the studied item correct equals 0.5 in 1PL and 2PL IRT models. In other words, the higher the \( b_i \) parameter for an item, the greater ability needed for an examinee to get 50% possibility to answer the item successfully.

**Chapter Summary**

DIF detection has become a popular research area, especially as more emphasis is put on the fairness and validity of testing and assessments. In order to investigate DIF
accurately and effectively, a number of procedures have been developed in recent decades. The MH and LR, two prevalent techniques in DIF detection, are implemented in this study.

Type I error is a significant issue in DIF analysis. A substantial number of previous studies have focused on exploring the primary factors influencing the inflation of Type I error in the MH and LR procedures. Sample size, test length, group ability difference, the item parameters, the proportion of DIF items, the type of DIF and group width all may be related to inflated Type I error (e.g., DeMars, 2009; Herrera and Gómez, 2008; Mazor et al., 1992; Swaminathan & Rogers, 1990). A variety of Monte Carlo simulations have been conducted under different conditions of manipulating varied factors stated above to display how the Type I error rates perform. In this chapter, the literature concerning Type I error performance in both MH and LR procedures was reviewed.

Studies conducted by Roussos and Stout (1996) and Holland and Thayer (1988) indicated that the inflation of Type I error rates was not a big concern in the MH procedure when a Rasch model is used. Zwick (1990) stated that inflated Type I errors might occur using the MH procedure with unequal group abilities unless the data fits in a Rasch model or the measured variables are extremely reliable.

What is consistent with most of the early studies is the fact that large sample size and large group ability difference tend to result in inflated Type I error rates in both MH and LR procedures after taking other influencing factors into account (e.g., DeMars, 2009; Güler & Penfield, 2009). Additionally, scholars believed that the magnitude of the
item discrimination parameter may affect the Type I error using the MH and LR procedures; for instance, DeMars (2009) mentioned higher item discriminations might result in larger inflated Type I error rates. Unfortunately, most of the early studies were only related to either relatively high or relatively low discrimination parameters. Few studies focused on evaluating the effect of discrimination parameters on the performance of the Type I error rates. A Monte Carlo simulation conducted by Herrera and Gómez (2008) examined the effect of discrimination parameters on the Type I error rates through implementing two levels of discrimination parameters \((a_i < 0.5 \text{ and } 0.5 \leq a_i \leq 1)\); however, all of the discrimination parameters selected in this study were no more than 1. Consequently, there is a need to perform additional Monte Carlo research to investigate how the item discrimination parameters affect the Type I error rates under the manipulated sample size and group ability difference.

Finally, Chapter two also provided a brief overview on item response theory, IRT assumptions and three types of popular IRT models since the data generated in the present study will be based on IRT models. IRT is an item-level framework. Currently, more and more statistics and measurement issues are solved using IRT models because of its flexibility and richness of information (Crocker & Algina, 1986; Hambleton et al, 1991).
Chapter 3: Methodology

A primary purpose of this study was to explore how item discrimination parameters influence the performance of the Type I error rates in both MH and LR procedures under varied sample size, group mean difference and matching method based on data generated with IRT models. The first important thing for conducting this research was to generate examinees’ response data randomly based on items’ properties and examinees’ latent traits. Monte Carlo simulation is an appropriate choice for this study.

As Lix, Keselman, and Keselman (1996) stated, in a typical Monte Carlo simulation study, “pseudorandom sets of numbers are generated using a predefined computer algorithm and are sampled from populations with known characteristics” (p. 589). Additionally, the Type I error rates will vary due to sampling error over samples, that is, a large number of replications for generation are required to ensure the accuracy of the results. A Monte Carlo study was used in the current study.

Monte Carlo Simulation

Monte Carlo simulation is a popular and powerful tool for statistical scholars to solve some problems where analytical mathematical methods cannot be applied, especially under the conditions where the null hypothesis is not true, weak theory exists in statistics or the assumptions of statistical procedures are violated (Brooks, 1998; Mooney, 1997). A Monte Carlo simulation rests on the principle “that the behavior of a statistic in random samples can be assessed by the empirical process of actually drawing lots of random samples and observing this behavior” (Mooney, 1997, p. 3).
Some properties of statistical and social processes can be evaluated by Monte Carlo procedures. For instance, Kleijnen (1974) stated that Monte Carlo simulation can be used to study the robustness of a statistic. That is, Monte Carlo conditions are built to “violate underlying assumptions of a specific statistic or procedure to determine how sensitive it is to the given violations” (Brooks, 1998, p. 125). Mooney (1997) also pointed out that Type I error rates and power can be evaluated by conducting Monte Carlo simulation. In Monte Carlo research, an artificial population, which resembles a real population, is often created by computer algorithm (Heh, 2007). Basic procedures of Monte Carlo simulations have been described by Mooney (1997) in detail:

1. Specify the pseudo-population in symbolic terms in such a way that it can be used to generate samples. This usually means developing a computer algorithm to generate data in a specified manner.

2. Sample from the pseudo-population (a pseudo-sample) in ways reflective of the statistical situation of interest, for example, with the same sampling strategy, sample size, and so forth.

3. Calculate $\hat{\theta}$ in the pseudo-sample and store it in a vector, $\Theta$.

4. Repeat Steps 2 and 3 $t$ times, where $t$ is the number of trials.

5. Construct a relative frequency distribution of the resulting $\hat{\theta}$ values, which is the Monte Carlo estimate of the sampling distribution of $\hat{\theta}$ under the conditions specifically by the pseudo-population and the sampling procedures.

(p. 4)
Although theta ($\theta$) is often used to represent the ability (or latent trait) of the examinee in item response theory, Mooney used $\theta$ to represent any statistic of interest. Monte Carlo simulations are not too hard to understand and the application of Monte Carlo simulation methods can be straightforward. The study by An (2010), for example, examined the Type I error rates and statistical power in testing multiple hypotheses; this was investigated effectively by conducting a Monte Carlo simulation. In robustness and power studies, a large number of samples are generated under both the true null hypothesis and false null hypothesis, respectively, while the underlying assumptions are violated. Researchers can then calculate the false rejection (i.e., Type I error) rates of the true null hypothesis and true rejection (i.e., statistical power) rates of the false null hypothesis as a proportion of the number of iterations.

Monte Carlo designs often follow similar procedures as standard research designs and therefore can include “identification of the population, description of the sampling plan, data collection and data analysis” (Brooks, Barcikowski, and Robey, 1999, p. 3). The difficulties of simulations lie in how to develop computer code for data generation and in reporting the estimated sampling distribution (Mooney, 1997).

**Research Design**

A Monte Carlo study evaluating the Type I error performance in both MH and LR procedures was performed in this research. The number of items was set at 50, which based on the literature review may be considered a medium test length. Dichotomous item response data for both reference group and focal group were generated based on given item parameters (e.g., $\alpha_i$, $b_i$, $c_i$) and person parameters (e.g., $\theta$) following IRT
models. The first item was always chosen as the studied item in the present study. Since the items were in random order, choosing the first item was equivalent to selecting a random item.

Several factors that were believed to have significant effects on the Type I error rates in DIF analyses were manipulated in this study, such as sample size, group ability differences, item parameters, and score group width. There were three primary research questions to be investigated. The factors listed above were manipulated differently for each research question.

**Research question 1 (RQ1).** The first research question (RQ1) concerned the pattern of the Type I error rates for the MH and LR procedures under thin matching and both deciles and quintiles thick matching, with different sample sizes and varied group ability differences in a 2PL IRT model. Simulation I was conducted to address this question. Three levels of sample size were used. The choices for sample size were not entirely arbitrary. Güler and Penfield (2009) believed that at least 200 to 250 examinees in each group can ensure that the MH and LR procedures have adequate power. They defined 1000 examinees in each group as a relatively large sample size. In addition, 250, 500, and 1000 examinees in each group were selected by Jodoin and Gierl (2001) in their research to represent small, medium, and large sample sizes, respectively. In this study, \( N_R \) denotes the number of examinees in reference group and \( N_F \) represents the number of examinees in focal group. \( N_R=N_F=300 \) and \( N_R=N_F=1000 \) were chosen to represent a relatively small and a relatively large group size, respectively. \( N_R=N_F=650 \) was used as
well in the present study to include a medium group size. The aim was to investigate the
tendency of the Type I error performance for sample sizes varying from small to large.

Group mean difference, denoted as $\mu_d$, was also manipulated in this research
question. Group ability difference was also not entirely arbitrary, but was based on
previous studies. Many early researchers suggested that group ability differences tend to
result in inflated Type I error rates in the MH and LR procedures (e.g., Güler & Penfield,
However, group mean difference was set equal to 1 standard deviation in most of these
studies, which was thought of as a relatively large difference. DeMars (2009) specifically
investigated the pattern of the Type I error rates under the conditions that group mean
ability differences were manipulated from 0.0 to 1.0 in increments of 0.1SD using various
methods including the MH and LR procedures. Simulation I in this study followed the
same manipulation in group ability difference. That is, the mean of ability distribution for
reference group was set to 0.0 while the ability mean for focal group varied from 0.0 to -
1.0 with the intervals of 0.1SD. Standard deviation (SD) was always set to 1.0 for both
reference and focal group in all simulations of the current study.

The item discrimination parameter was the most important factor to be examined
in this study. Its effects on the Type I error rates in both MH and LR procedures has been
partially investigated by early researchers. For example, DeMars (2010) believed that
items with higher discrimination parameters commonly yield more reliable scores.
However, the pattern of the Type I error performance under the varied discrimination
parameters is still an unknown domain. Under Simulation I, item difficulty parameters
were randomly generated within the range of -2.0 to 2.0, following a uniform distribution. The item discrimination parameters were also randomly selected from a uniform distribution, ranging from 0.0 to 1.8, which included both strong and weak items.

Score group width was set at 51, 10 and 5 score levels, which corresponds to thin matching, deciles, and quintiles thick matching, respectively. When the total score, the sum of individual responses for each item, was used as a conditioning variable, thin matching was used. However, previous literature has indicated that thick matching might improve the DIF detection performance (e.g., Donoghue & Allen, 1993); therefore, deciles thick matching and quintiles thick matching were used as well. If examinees are grouped into ten score levels and scoring level is used as a matching criterion, deciles thick matching is conducted. Note that each score level has an approximately equal number of examinees under quintiles. Similarly, quintiles thick matching specifies that the conditioning variable is collapsed into five score levels and the number of examinees for each level is identical.

**Research question 2 (RQ2).** The second research question (RQ2) concerned the pattern of Type I error rates for the MH and LR procedures when the range of discrimination parameters for all items varied from very small to relatively large under large sample size (e.g., \(N_R=N_F=1000\)) and large group ability difference (e.g., \(\mu_d=1\)) using thin matching, deciles thick matching, and quintiles thick matching with data simulated according to 2PL IRT models. Simulation II addressed this question. A number of previous research studies mentioned that a larger sample size and a larger group ability difference may yield a larger inflated Type I error rates in the MH and LR procedures
(e.g., DeMars, 2009; Roussos & Stout, 1996). Consequently, in Simulation II, the maximum number of examinees in each group was selected and the largest group ability difference was used. The ability mean was set equal to 0.0 and -1.0 for reference group and focal group, respectively, so that $\mu_d = 1$.

The range of discrimination parameters for all items was set from 0.2 to 2.0. As explained previously, the discrimination parameter for an item is usually larger than 0 and no more than 2 even it can be defined on the entire scale (Hambleton et al., 1991). In this research question, the minimum range of all item discrimination parameters was set to 0.2, for example, the interval [0.5, 0.7] or [1.6, 1.8]. The maximum range of discrimination parameters for all items is 2.0, that is, the interval (0.0, 2.0). The range of difficulty parameters was kept the same as that of Simulation I. Under this research question, the Type I error rates were detected when the discrimination parameter for the studied item was not fixed. That is, the studied item had a random discrimination parameter from the assigned range for the purpose of calculating the Type I error rates.

**Research question 3 (RQ3).** The third research question (RQ3) concerned the pattern of the Type I error rates for the MH and LR procedures when the discrimination parameter for the studied item varied systematically from relatively low to relatively high and non-studied items both included and excluded items with low discrimination parameters under large sample size and large group ability difference using thin matching and deciles thick matching where data are simulated using both 2PL and 3PL IRT models.
Simulation III(a) and Simulation III(b) were conducted following 2PL IRT models. Sample size $N_R=N_I=1000$ and group ability difference $\mu_d = 1$ were used in these two simulations. For simulation III(a), the discrimination parameter for the studied item varied from 0.2 (low discrimination) to 2.0 (high discrimination). The discrimination parameters for the remaining 49 items varied from 0.2 to 1.8, which simulated both weak and strong items. In simulation III(b), the discrimination parameter for the studied item is still varied from 0.2 to 2.0, whereas the non-studied items have discrimination parameters from 1.2 to 2.0, which only includes comparatively strong items.

Simulation IV(a) and simulation IV(b), which also addressed the third research question, were carried out following a 3PL IRT model. All the conditions employed in these two simulations were exactly the same as those in simulation III(a) and simulation III(b), respectively, except that the guessing parameter for all items was set to 0.2.

**Monte Carlo Simulation in R**

The Monte Carlo simulations were conducted using the statistical programming language R, which contains various statistical packages, to simulate the required Type I error rates. It is worth mentioning that the source codes of R software can be obtained freely, which makes it popular and widely used (see [http://www.r-project.org](http://www.r-project.org)). A substantial number of built-in functions belonging to different R packages are easily used to process various statistical analyses. The discrimination parameters, difficulty parameters, and guessing parameters for each item were generated under different requirements. Accordingly, dichotomous item responses were created based on the simulated item parameters and given person parameters. The MH and LR procedures
were simultaneously conducted for the studied item and the false rejection rates are recorded in large matrices for different situations.

The needed number of iterations was based on previous literature. Robey and Barcikowski (1992) suggested that 5,422 iterations were required for a robust estimation of Type I errors. Mooney (1997) proposed that the more the better in choosing the number of iterations for Monte Carlo simulations. However, excessive iterations consume too much time. Hence, in this current study, 10,000 iterations were used. On the one hand, it can sufficiently ensure the stability and generalizability of the results. On the other hand, it avoids spending surplus time required to run more than 10,000 replications.

**Operational Definitions**

The following definitions provide researchers with a complete understanding about the methodology applied to the Monte Carlo study.

**P-value.** The $p$-value is an indicator of statistical significance. Researchers use $p$ values effectively and simply to examine the issues of statistical inference (Hochberg & Benjamini, 1990). The $p$ values of each design for the MH and LR procedures were available by developing specific R coding. These were then stored in large matrices.

**Type I error.** The Type I error occurs if a true null hypothesis is rejected falsely. In this study, all items were generated without DIF. A nominal significance level $\alpha = .05$ is used as the criterion. The null hypothesis is rejected when $p$-values obtained from the MH or LR procedures are less than .05. The number of rejections was summed and divided by the number of replications (e.g., 10,000) to yield the false rejection rate, that
is, the Type I error rate. The patterns of the Type I error rates were displayed under the manipulated factors.

**Thin matching.** If the total score is used as matching variable, thin matching is conducted. In the current study, 50 items was selected. Consequently, there are 51 score levels for thin matching (from 0 correct to 50 correct).

**Deciles thick matching.** When ten scoring levels are formed by pooling total score and these ten scoring levels are used as matching variable(s), deciles thick matching is conducted. Similarly, the number of examinees for each level is approximately identical. In the R program, the quantile function is used to form a new dataset for deciles thick matching.

**Quintiles thick matching.** When pooling total scores to form five groups and these five scoring levels are used as the matching criteria, quintiles thick matching is carried out. It is noted that the number of examinees for each score group is approximately equal. The same R quantile function was used to perform quintiles thick matching.

**MH procedure.** The MH procedure was used to detect DIF in this study. In the R program, thin matching for the MH was achieved by using the `difMH` function from the `difR` package, while deciles and quintiles thick matching was completed by employing the `mantelhaen.test` function in the Cronbach-Mesbah Curve (CMC) package.

**LR procedure.** The LR procedure is another method selected to examine DIF in the current research. The `glm` function, in the package called `stats`, was used to conduct the LR analysis. When raw score is chosen as a predictor, thin matching is completed, on
the contrary, when ten and five scoring levels are chosen as the data, deciles and quintiles thick matching is conducted, respectively.

**Data Generation**

The item parameter data including discrimination and difficulty parameters for 50 items were generated following a uniform distribution using R. The `runif` function in the *irtoys* package was used to achieve this goal. The uniform distribution, usually designated as U(0, 1), is “the building block of all Monte Carlo simulation work” (Brooks, 1998, p. 144). In a 2PL IRT model, guessing parameters for all items are identical to 0.0, while in 3PL IRT model, all items’ guessing parameters maintain 0.2 in this research, which was commonly used in several prior DIF studies (e.g., DeMars, 2009; Güler & Penfield, 2009). The simulated item parameters and population ability mean were used as inputs to the generation of dichotomous item response data for both reference group and focal group. The `sim` function, also belonging to the *irtoys* package, was employed in this generation. The population ability mean for reference group was always kept at 0.0, whereas the population ability mean for focal group varied from 0.0 to -1.0 with increments of -0.1SD.

**Data Collection Procedure**

The dichotomous item response data for each examinee were recorded in a matrix and, therefore, total score for each item was calculated. In the thin matching procedure, total score was treated as a conditioning variable. In deciles and quintiles thick matching, ten and five collapsed score groups were formed, respectively, according to the examinees’ total score. The scoring level for each examinee was determined in terms of
the definition of deciles and quintiles, which has been described in the previous section. In thick matching, collapsed scoring levels were used as the conditioning variable. Additionally, sample size, group mean ability and discrimination parameter were also manipulated under different conditions to form a substantial number of cells studied in each research question. For example, $11 \times 3 \times 3 \times 2$ cells (i.e., 198 cells) were required for RQ1 because there were 11 types of group ability difference, 3 types of sample size, 3 types of matching methods and 2 types of DIF detection procedures. Each cell was replicated 10,000 times and the corresponding samples were created. It took around 2 hours to complete 10,000 replications for each cell using a computer with i7-2600 processor and 3.40 GHZ CPU and 8.00 GB RAM. For each sample, the program performed DIF analyses using the MH and LR procedures under various manipulated factors. The $p$ values for all the conditions were extracted and stored in large matrices during the R program execution. The mean proportion of false rejections, that is, the Type I error rate for each condition, was calculated as the proportion of $p$ values less than the significance level.

Verification of the Data Collection Procedures

Based on the suggestions of Bratley, Fox and Schrage (1987), several steps for verifying the algorithms in a Monte Carlo study are necessary. First, manual verification for examining the logic of computer code is essential. It ensures that the same results can be obtained from both the computer analysis and hand calculation. Second, modular testing is needed to ensure each subroutine can yield a rational output when considering all possible inputs. Third, the results are compared with known solutions. Fourth,
sensitivity testing is used to ensure that the performance of computer code is sensible with varied parameters. Finally, stress testing is carried out for ensuring that there are no unexpected outcomes with strange values (Brooks, 1998). In the present study, the verification of the data collection procedures will be performed as indicated in the steps described above.

Manual verification was employed to explore the logic of computer code developed in this study. For example, a small sample size (e.g., 20 examinees/group) and small iterations (e.g., 20 trials) were used in generating item responses. Total score for each examinee was calculated by hand, which was used to compare with the results obtained from computer program. Additionally, a similar procedure was used in verifying the correctness of scoring levels of deciles and quintiles thick matching. The number of examinees in each scoring level yielding from R codes was exactly the same as the result computed by hand.

Modular testing was conducted in this study by examining several single simulations of data. For instance, item discrimination and difficulty parameters following a uniform distribution generated in R program were tested to ensure that they are indeed distributed uniformly.

Individual samples for the MH and LR procedures were analyzed by the R program and the results were recorded into a matrix. The dataset of each pattern for the last replication were pulled out through developing specific R codes and entered into SPSS (SPSS Inc., 2008). The same results exhibited by SPSS and R confirmed that the correct R codes and options were being used for the MH and LR procedures.
Sensitivity testing ensures that the performance of computer model constructed by researchers is sensible when parameters are varied (Brooks, 1998). In the 3PL IRT model for example, the guessing parameter was changed from 0.0 to 0.2 with increments of 0.02SD. The results produced from R testing codes confirmed the stability of the R program.

Stress testing is performed to ensure that no unexpected outcomes occur with strange or extreme values (Brooks, 1998). For example, varied iteration numbers, such as 0, 1, 10, 20, 100, and 1000, were used to complete stress testing. The results indicated that the R program performed well under all the conditions.

**Data Analysis Procedure**

The main purpose of the current study was to evaluate how the factors such as sample size, group ability difference, group width, especially item discrimination parameter, affect the Type I error rates in the MH and LR procedures. In this study, the Bradley’s (1978) stringent criterion interval [.045, .055], that is, $\alpha \pm .1\alpha$, was selected to evaluate the performance of the Type I error rates.

Three primary research questions were studied to investigate the effect of different combinations of these influencing factors on the performance of the Type I error rates. In order to explore these questions empirically, a Monte Carlo study was performed. In the following section, the research questions are restated and how data analysis was performed in the Monte Carlo study is introduced in detail.

RQ1: What is the pattern of the Type I error rates displayed in the MH and LR procedures under thin matching, deciles thick matching, and quintiles thick matching
with different sample sizes and varied group ability differences with data generated based on 2PL IRT models? Results of three levels of sample size and three types of matching method were analyzed under varied group ability differences with interval of 0.1. A total 198 cells for the Type I error rates were recorded and the tendency of the Type I error rates performance in each method were evaluated under various conditions.

RQ2: What is the pattern of the Type I error rates for the MH and LR procedures displayed when the range of discrimination parameters for all items varies from very small to relatively large under large sample size (e.g., 1,000) and large group ability difference (e.g., \( \mu_d = 1 \)) using thin matching, deciles thick matching, and quintiles thick matching with data simulated according to 2PL IRT models? The effect of the item discrimination parameter were evaluated on the Type I error rates when other influencing factors were manipulated. The false rejection rates were saved. The pattern of the Type I error rates provided an estimate of how well the MH and LR procedures performed in the DIF detection when item discrimination parameter varies.

RQ3: What is the pattern of the Type I error rates for the MH and LR procedures displayed when the discrimination parameter for the studied item varies systematically from low (e.g., 0.2) to high (e.g., 2.0) and the non-studied items include and exclude items with low discrimination parameters under large sample size (e.g., 1,000) and large group ability difference (e.g., \( \mu_d = 1 \)) using thin matching and deciles thick matching where data are simulated using both 2PL and 3PL IRT models? The result for each cell was recorded for 2PL and 3PL, respectively. The examination of non-studied items including or excluding weak items exhibited how the discrimination parameters affected
the DIF detection accuracy. Additionally, comparing the outcome from 2PL to that from 3PL indicated how guessing influenced the Type I error rates.

**Chapter Summary**

A Monte Carlo study was conducted to examine how the Type I error rates in the MH and LR procedures perform when a variety of factors are manipulated such as sample size, group ability difference, matching method and particular item discrimination parameter. Several corresponding Monte Carlo simulations were designed for this study. The patterns of the Type I error rates under various conditions were exhibited and the effects of influencing factors were evaluated.
Chapter 4: Results

The current study aimed to display the patterns of the Type I error rates of DIF detection using the MH and LR procedures, especially to explore how item discrimination parameters and group mean ability differences combine to affect the inflation of Type I error rates in these methods. In order to carry out this study, three research questions were addressed and Monte Carlo simulations were performed. The results are summarized in this chapter, organized in terms of the research questions.

Data Analysis

Analysis for research question one. What is the pattern of the Type I error rates displayed in the MH and LR procedures under thin matching, deciles thick matching, and quintiles thick matching with different sample sizes and varied group ability differences with data generated based on 2PL IRT models? In this research question, the Type I error rates were investigated under sample sizes of 300, 650, and 1,000 using three matching methods: thin, deciles, and quintiles.

The results for the MH procedure obtained from 10,000 replications for different sample sizes are given in Table 2, Table 3 and Table 4, respectively. A visual display of these results can be seen in Figure 3. This figure illustrates that when the group ability difference increases with increments of 0.1 SD, the inflated false rejection rates for the MH procedure tend to increase no matter what sample sizes and matching methods are used. That is, all the lines in Figure 3 are prone to go up from left to right.
### Table 2

*Type I Error Rates for the MH procedure with Sample Size of 300*

<table>
<thead>
<tr>
<th>Group Mean Difference</th>
<th>Thin</th>
<th>Deciles Thick</th>
<th>Quintiles Thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0403</td>
<td>0.0405</td>
<td>0.0423</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0396</td>
<td>0.0396</td>
<td>0.0385</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0379</td>
<td>0.0394</td>
<td>0.0415</td>
</tr>
<tr>
<td>0.3</td>
<td>0.0394</td>
<td>0.0396</td>
<td>0.0412</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0385</td>
<td>0.0411</td>
<td>0.0430</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0390</td>
<td>0.0403</td>
<td>0.0440</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0419</td>
<td>0.0454</td>
<td>0.0485</td>
</tr>
<tr>
<td>0.7</td>
<td>0.0448</td>
<td>0.0468</td>
<td>0.0534</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0481</td>
<td>0.0485</td>
<td>0.0570</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0485</td>
<td>0.0503</td>
<td>0.0636</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0509</td>
<td>0.0523</td>
<td>0.0660</td>
</tr>
</tbody>
</table>

### Table 3

*Type I Error Rates for the MH procedure with Sample of Size 650*

<table>
<thead>
<tr>
<th>Group Mean Difference</th>
<th>Thin</th>
<th>Deciles Thick</th>
<th>Quintiles Thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0424</td>
<td>0.0418</td>
<td>0.0421</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0434</td>
<td>0.0425</td>
<td>0.0421</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0441</td>
<td>0.0447</td>
<td>0.0468</td>
</tr>
<tr>
<td>0.3</td>
<td>0.0468</td>
<td>0.0479</td>
<td>0.0516</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0496</td>
<td>0.0503</td>
<td>0.0574</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0497</td>
<td>0.0521</td>
<td>0.0649</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0527</td>
<td>0.0549</td>
<td>0.0710</td>
</tr>
<tr>
<td>0.7</td>
<td>0.0556</td>
<td>0.0604</td>
<td>0.0836</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0620</td>
<td>0.0670</td>
<td>0.0964</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0686</td>
<td>0.0759</td>
<td>0.1087</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0729</td>
<td>0.0795</td>
<td>0.1195</td>
</tr>
</tbody>
</table>
Table 4

*Type I error Rates for the MH procedure with Sample Size of 1000*

<table>
<thead>
<tr>
<th>Group Mean Difference</th>
<th>Thin</th>
<th>Deciles Thick</th>
<th>Quintiles Thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0477</td>
<td>0.0464</td>
<td>0.0461</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0485</td>
<td>0.0475</td>
<td>0.0480</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0496</td>
<td>0.0509</td>
<td>0.0526</td>
</tr>
<tr>
<td>0.3</td>
<td>0.0529</td>
<td>0.0533</td>
<td>0.0584</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0534</td>
<td>0.0578</td>
<td>0.0693</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0597</td>
<td>0.0645</td>
<td>0.0823</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0635</td>
<td>0.0705</td>
<td>0.0938</td>
</tr>
<tr>
<td>0.7</td>
<td>0.0714</td>
<td>0.0777</td>
<td>0.1085</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0783</td>
<td>0.0863</td>
<td>0.1262</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0879</td>
<td>0.0987</td>
<td>0.1449</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0959</td>
<td>0.1055</td>
<td>0.1647</td>
</tr>
</tbody>
</table>

Notably, larger sample sizes exhibited larger inflation of Type I error rates when the other factors were manipulated identically. That is, the pattern of deciles thick matching in MH procedure, for example, shows that the inflation of Type I error rate increased from 0.0523 to 0.1055 as the sample size increased from small (N=300) to large (N=1,000) with the impact of 1.0 SD. In other words, the false rejection rate is within the Bradley’s (1978) stringent criterion interval [.045, .055], that is, $\alpha \pm .1\alpha$ under small sample size, while it is far outside the stringent criterion interval under the large sample size condition.

Examination of Table 2 to Table 4 indicates that among all of the matching methods, thin matching generally performs best and quintiles thick matching performs worst. That is, in general, thin matching yields the least inflated Type I error rates while quintiles thick matching shows the largest inflation of Type I error rates independent of
sample size or DIF detection methods. Correspondingly, it can been seen from the Figure 3 that the curve for sample size of 300 with thin matching is the flattest and the line for sample size of 1000 with quintiles thick matching is the steepest.

![Figure 3](image)

**Figure 3.** Type I error rates for the MH procedure.

The false rejection rates for the LR procedure are presented in Table 5, Table 6 and Table 7. In contrast to the MH results, the LR procedure causes more inflation of Type I error rates than the MH procedure does under the corresponding conditions. For example, in the condition of sample size of 300 with thin matching and impact of 0.9 SD, the false rejection rate is 0.0485 for the MH procedure, which lies within Bradley’s (1978) stringent criterion and 0.0602 for the LR procedure, which is beyond the stringent
criterion interval. A comparison of all similar results across tables shows this same general pattern.

Table 5

Type I Error Rates for the LR Procedure with Sample Size of 300

<table>
<thead>
<tr>
<th>Group Mean Difference</th>
<th>Thin</th>
<th>Deciles Thick</th>
<th>Quintiles Thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0516</td>
<td>0.0521</td>
<td>0.0532</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0492</td>
<td>0.0489</td>
<td>0.0484</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0496</td>
<td>0.0489</td>
<td>0.0495</td>
</tr>
<tr>
<td>0.3</td>
<td>0.0487</td>
<td>0.0506</td>
<td>0.0523</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0510</td>
<td>0.0532</td>
<td>0.0559</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0515</td>
<td>0.0532</td>
<td>0.0579</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0557</td>
<td>0.0581</td>
<td>0.0639</td>
</tr>
<tr>
<td>0.7</td>
<td>0.0558</td>
<td>0.0608</td>
<td>0.0687</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0583</td>
<td>0.0624</td>
<td>0.0746</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0602</td>
<td>0.0640</td>
<td>0.0809</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0619</td>
<td>0.0674</td>
<td>0.0833</td>
</tr>
</tbody>
</table>
Table 6

*Type I Error Rates for the LR Procedure with Sample Size of 650*

<table>
<thead>
<tr>
<th>Group Mean Difference</th>
<th>Thin</th>
<th>Deciles Thick</th>
<th>Quintiles Thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0499</td>
<td>0.0501</td>
<td>0.0504</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0501</td>
<td>0.0503</td>
<td>0.0509</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0510</td>
<td>0.0524</td>
<td>0.0547</td>
</tr>
<tr>
<td>0.3</td>
<td>0.0545</td>
<td>0.0565</td>
<td>0.0610</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0564</td>
<td>0.0595</td>
<td>0.0686</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0581</td>
<td>0.0637</td>
<td>0.0774</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0614</td>
<td>0.0702</td>
<td>0.0863</td>
</tr>
<tr>
<td>0.7</td>
<td>0.0654</td>
<td>0.0760</td>
<td>0.1004</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0736</td>
<td>0.0836</td>
<td>0.1132</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0794</td>
<td>0.0885</td>
<td>0.1274</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0856</td>
<td>0.0959</td>
<td>0.1384</td>
</tr>
</tbody>
</table>

Table 7

*Type I Error Rates for the LR Procedure with Sample Size of 1000*

<table>
<thead>
<tr>
<th>Group Mean Difference</th>
<th>Thin</th>
<th>Deciles Thick</th>
<th>Quintiles Thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0528</td>
<td>0.0521</td>
<td>0.0520</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0538</td>
<td>0.0536</td>
<td>0.0541</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0551</td>
<td>0.0579</td>
<td>0.0596</td>
</tr>
<tr>
<td>0.3</td>
<td>0.0562</td>
<td>0.0608</td>
<td>0.0643</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0609</td>
<td>0.0664</td>
<td>0.0786</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0686</td>
<td>0.0756</td>
<td>0.0945</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0738</td>
<td>0.0853</td>
<td>0.1090</td>
</tr>
<tr>
<td>0.7</td>
<td>0.0812</td>
<td>0.0931</td>
<td>0.1245</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0888</td>
<td>0.1036</td>
<td>0.1434</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0991</td>
<td>0.1176</td>
<td>0.1650</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1080</td>
<td>0.1243</td>
<td>0.1818</td>
</tr>
</tbody>
</table>
The pattern of the Type I error rates in the LR procedure has been displayed graphically in Figure 4. Similar to the MH procedure, all the lines tend to increase across the ability mean difference continuum. Also, the line for the sample size of 300 with thin matching was located highest while the line for the sample size of 1000 with quintiles thick matching was located lowest. That is, large sample size and large ability difference also resulted in higher inflated false rejection rates for the LR procedure. Another similar finding is that thin matching is generally better than deciles thick matching in the LR procedure. Deciles thick matching is in generally better than quintiles thick matching in the LR procedure as well. However, in almost every corresponding condition, the LR procedure exhibits slightly higher inflation than MH.

*Figure 4. Type I error rates for the LR procedure.*
In summary, the patterns of all of the findings from Simulation I are consistent with previous studies. First, the pattern of the Type I error rates for both MH and LR procedures indicates that larger sample size and larger group ability differences tend to yield higher inflation of Type I error rates, which is in agreement with much early research (e.g., DeMars, 2009; Roussos & Stout, 1996). Next, the conclusion that thin matching performs better than both deciles and quintiles thick matching may be because thin matching contains more abundant information. Also, thick matching may confound DIF and impact; it may increase statistical bias, and may not work as well with longer tests (Donoghue & Allen, 1993). Finally, compared to the MH procedure, the LR procedure almost always shows greater inflated Type I error rates, which agrees with DeMars’ (2009) study.

**Analysis for research question two.** What is the pattern of the Type I error rates for the MH and LR procedures displayed when the range of discrimination parameters for all items varies from very small to relatively large under large sample size (e.g., 1,000) and large group ability difference (e.g., $\mu_d=1$) using thin matching, deciles thick matching, and quintiles thick matching with data simulated according to 2PL IRT models?

The second simulation was conducted to answer this research question. It is noted that the outcome from the first research question verified that large sample size and large group impact are two issues related to inflated Type I error. Therefore, in the second research question, the potentially most troublesome conditions, that is, 1,000 sample size and group mean difference of 1.0 SD, were selected to investigate how the Type I error
rates performs when the ranges of all item discrimination parameters are generated from 0.2 to 2.0.

Table 8 presents the results of the MH procedure achieved from the second simulation over 10,000 replications for each condition. One important observation is that for the thin matching, the Type I error rates can be maintained fairly well, that is, near the nominal significance level (.05), when the range of all item discrimination parameters is very small, such as 0.2 in this research. In other words, all items are relatively homogeneous in discrimination. For example, the false rejection rates are 0.0441, 0.0452 and 0.0418 when the discrimination parameters for all items are generated in the intervals [0.5, 0.7], [0.8, 1.0] and [1.6, 1.8], respectively. On the contrary, as the interval of all item discrimination parameters increases, the Type I error rates became inflated. For example, in the interval [0.3, 2.0], the false rejection rate reached 0.0624, which is beyond Bradley’s (1978) stringent criterion when the range of item discrimination parameters was expanded to 1.7. The finding that the inflation can be controlled within nominal significance level when all discrimination parameters are generated in a tiny range (0.2) may be because it is similar to a Rasch model. It is in agreement with the study of Holland and Thayer (1988) that inflated Type I error is not a potential problem for the MH procedure when data are generated using the one-parameter (or Rasch) model.
Table 8

Type I Error Rates for the MH procedure with Varied Ranges of Discrimination Parameters for All Items with N=1,000 and Group Mean Difference=1.0

<table>
<thead>
<tr>
<th>Range of a</th>
<th>Interval of a</th>
<th>Thin</th>
<th>Deciles Thick</th>
<th>Quintiles Thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>a=0.5~0.7</td>
<td>0.0441</td>
<td>0.0481</td>
<td>0.0662</td>
</tr>
<tr>
<td>0.2</td>
<td>a=0.8~1.0</td>
<td>0.0452</td>
<td>0.0488</td>
<td>0.0789</td>
</tr>
<tr>
<td>0.2</td>
<td>a=1.6~1.8</td>
<td>0.0418</td>
<td>0.0441</td>
<td>0.1141</td>
</tr>
<tr>
<td>0.3</td>
<td>a=0.3~0.6</td>
<td>0.0515</td>
<td>0.0528</td>
<td>0.0698</td>
</tr>
<tr>
<td>0.3</td>
<td>a=0.8~1.1</td>
<td>0.0465</td>
<td>0.0509</td>
<td>0.0801</td>
</tr>
<tr>
<td>0.3</td>
<td>a=1.6~1.9</td>
<td>0.0420</td>
<td>0.0445</td>
<td>0.1183</td>
</tr>
<tr>
<td>0.4</td>
<td>a=0.3~0.7</td>
<td>0.0549</td>
<td>0.0570</td>
<td>0.0749</td>
</tr>
<tr>
<td>0.4</td>
<td>a=0.6~1.0</td>
<td>0.0478</td>
<td>0.0511</td>
<td>0.0799</td>
</tr>
<tr>
<td>0.4</td>
<td>a=1.4~1.8</td>
<td>0.0406</td>
<td>0.0466</td>
<td>0.1088</td>
</tr>
<tr>
<td>0.5</td>
<td>a=0.3~0.8</td>
<td>0.0601</td>
<td>0.0595</td>
<td>0.0794</td>
</tr>
<tr>
<td>0.5</td>
<td>a=0.6~1.1</td>
<td>0.0469</td>
<td>0.0527</td>
<td>0.0844</td>
</tr>
<tr>
<td>0.5</td>
<td>a=1.3~1.8</td>
<td>0.0425</td>
<td>0.0469</td>
<td>0.1084</td>
</tr>
<tr>
<td>0.6</td>
<td>a=0.2~0.8</td>
<td>0.0699</td>
<td>0.0721</td>
<td>0.0910</td>
</tr>
<tr>
<td>0.6</td>
<td>a=0.6~1.2</td>
<td>0.0489</td>
<td>0.0535</td>
<td>0.0862</td>
</tr>
<tr>
<td>0.6</td>
<td>a=1.2~1.8</td>
<td>0.0413</td>
<td>0.0473</td>
<td>0.1063</td>
</tr>
<tr>
<td>0.8</td>
<td>a=0.2~1.0</td>
<td>0.0706</td>
<td>0.0754</td>
<td>0.1023</td>
</tr>
</tbody>
</table>

Note. a=discrimination parameter.
Table 8 (continued)

<table>
<thead>
<tr>
<th>Range of $a$</th>
<th>Interval of $a$</th>
<th>Thin</th>
<th>Deciles Thick</th>
<th>Quintiles Thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>a=0.6~1.4</td>
<td>0.0480</td>
<td>0.0553</td>
<td>0.0932</td>
</tr>
<tr>
<td>0.8</td>
<td>a=1.0~1.8</td>
<td>0.0433</td>
<td>0.0489</td>
<td>0.1059</td>
</tr>
<tr>
<td>1.0</td>
<td>a=0.2~1.2</td>
<td>0.0697</td>
<td>0.0739</td>
<td>0.1087</td>
</tr>
<tr>
<td>1.0</td>
<td>a=0.6~1.6</td>
<td>0.0493</td>
<td>0.0568</td>
<td>0.0994</td>
</tr>
<tr>
<td>1.0</td>
<td>a=1.0~2.0</td>
<td>0.0442</td>
<td>0.0517</td>
<td>0.1118</td>
</tr>
<tr>
<td>1.2</td>
<td>a=0.2~1.4</td>
<td>0.0704</td>
<td>0.078</td>
<td>0.1197</td>
</tr>
<tr>
<td>1.2</td>
<td>a=0.4~1.6</td>
<td>0.0545</td>
<td>0.0611</td>
<td>0.1062</td>
</tr>
<tr>
<td>1.2</td>
<td>a=0.8~2.0</td>
<td>0.0458</td>
<td>0.0514</td>
<td>0.1137</td>
</tr>
<tr>
<td>1.5</td>
<td>a=0.2~1.7</td>
<td>0.0670</td>
<td>0.0771</td>
<td>0.1270</td>
</tr>
<tr>
<td>1.5</td>
<td>a=0.4~1.9</td>
<td>0.0531</td>
<td>0.0618</td>
<td>0.1176</td>
</tr>
<tr>
<td>1.5</td>
<td>a=0.5~2.0</td>
<td>0.0524</td>
<td>0.0597</td>
<td>0.1153</td>
</tr>
<tr>
<td>1.6</td>
<td>a=0.2~1.8</td>
<td>0.0674</td>
<td>0.0767</td>
<td>0.1292</td>
</tr>
<tr>
<td>1.6</td>
<td>a=0.3~1.9</td>
<td>0.0612</td>
<td>0.0675</td>
<td>0.1255</td>
</tr>
<tr>
<td>1.6</td>
<td>a=0.4~2.0</td>
<td>0.0566</td>
<td>0.0625</td>
<td>0.1183</td>
</tr>
<tr>
<td>1.7</td>
<td>a=0.2~1.9</td>
<td>0.0669</td>
<td>0.0773</td>
<td>0.1339</td>
</tr>
<tr>
<td>1.7</td>
<td>a=0.3~2.0</td>
<td>0.0624</td>
<td>0.0688</td>
<td>0.1239</td>
</tr>
<tr>
<td>1.8</td>
<td>a=0.2~2.0</td>
<td>0.0663</td>
<td>0.0773</td>
<td>0.1379</td>
</tr>
<tr>
<td>2.0</td>
<td>a=0.0~2.0</td>
<td>0.0937</td>
<td>0.1049</td>
<td>0.1664</td>
</tr>
</tbody>
</table>
Additionally, in the thin matching, strong item discrimination parameters tend to make the Type I error rates adhere to the nominal significance level. Comparing the interval [0.2, 1.2] with [1.0, 2.0] shows that the false rejection rates decrease from 0.0697 to 0.0442, indicating that even when the range of all item discrimination parameters (1.0) is wide, no inflation occurs if all items are relatively strong. The deciles thick matching method exhibited a similar tendency, except that it shows slightly more inflation than thin matching. The information from Table 8 shows that the performance of quintiles thick matching was always worse, that is, all the false rejection rates were beyond Bradley’s stringent criterion (1978).

Finally, the MH odds ratio was also investigated in simulation II. Figure 5 illustrates the pattern of odds ratio with no impact. As described previously, the given item will favor the reference group, which was the group with a higher mean, if the odds ratio is greater than 1 but the item will favor the focal group, which was the group with a lower mean, if the odds ratio is less than 1 (Dorans, 1989). It can be seen that the points in Figure 5 are distributed almost evenly on the two sides of the line that represents an MH odds ratio equal to 1. That is, it illustrates that neither reference group nor focal group was favored by DIF when no impact exists. The pattern of MH odds ratio with impact of 1.0 SD is given in Figure 6. This figure shows that for discrimination indices above 1.0 when there is impact, more items with MH odds ratios are greater than 1.0, which therefore favor the group with the higher mean. For items with discrimination indices below 1.0 in the presence of large impact, most of the items show that the MH odds ratio is less than 1.0, thereby favoring the group with the lower mean. Consequently, the
pattern in Figure 6 indicates that higher discrimination item appears to favor group with higher ability (reference group) and lower discrimination item tends to favor the group with lower ability (focal group). The finding is in agreement with the study of Zwick (1990).

*Figure 5.* The MH odds ratio with impact of 0.0.
Table 9 displays the false rejection rates for the LR procedure under the condition of sample size 1000 and impact of 1.0 SD. It can be seen that almost all the Type I error rates are higher than .05; however, the LR procedure exhibits a similar tendency as the MH procedure.

First, narrower range of all item discrimination parameters appears to yield smaller inflated false rejection rates under thin matching and deciles thick matching. For example, under the thin matching, when the discriminating parameters for all items are simulated in the interval of [0.5, 0.7] and [0.3, 0.6], the Type I error rates are .0525 and .0578, respectively, which are close to the nominal significance level. Whereas, for the interval of [0.2, 1.7], the false rejection rate increases to .0775, indicating a relatively large inflation.
Next, for the thin matching, all items with strong discriminating parameters are prone to maintain the Type I error rates at or near the nominal significance level no matter that the generation interval for all item discrimination parameters is small or large. Specifically, as shown in Table 9, the Type I error rates equal to .0493, .0487, .0478, .0482 and .0512 for the interval of [1.6, 1.8], [1.4, 1.8], [1.3, 1.8], [1.2, 1.8] and [1.0, 1.8], respectively. That is, even the range of interval goes up from 0.2 to 0.8, the Type I error rates still can be controlled within Bradley’s (1978) stringent criterion interval [.045, .055], $\alpha \pm .1\alpha$, if all the items are relatively strong. Furthermore, the false rejection rate is .0518 for the interval of [1.0, 2.0], while the false rejection rate is .0783, indicating an inflation, for the interval of [0.2, 1.2]. Notice that the above two conditions have the same range for all item discrimination parameters (1.0). Consequently, it confirms the conclusion described above that the Type I error rates are maintained fairly well when no weak items are included no matter the range of all items discrimination parameters.

Finally, in contrast to thin matching, thick matching methods perform worse, especially for quintiles thick matching. Nevertheless, the deciles thick matching method displays a similar tendency as thin matching does, except that it always shows more inflation. As seen in Table 9, quintiles thick matching always exhibits highly inflation regardless of the interval of all item discrimination parameters.
Table 9

Type I Error Rates for the LR procedure with Varied Ranges of Discrimination Parameter for All Items with $N=1,000$ and Group Mean Difference=1.0

<table>
<thead>
<tr>
<th>Range of $a$</th>
<th>Interval of $a$</th>
<th>Thin</th>
<th>Deciles Thick</th>
<th>Quintiles Thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>$a=0.5\sim0.7$</td>
<td>0.0525</td>
<td>0.0556</td>
<td>0.0787</td>
</tr>
<tr>
<td>0.2</td>
<td>$a=0.8\sim1.0$</td>
<td>0.0531</td>
<td>0.0607</td>
<td>0.0909</td>
</tr>
<tr>
<td>0.2</td>
<td>$a=1.6\sim1.8$</td>
<td>0.0493</td>
<td>0.0579</td>
<td>0.1336</td>
</tr>
<tr>
<td>0.3</td>
<td>$a=0.3\sim0.6$</td>
<td>0.0578</td>
<td>0.0621</td>
<td>0.0793</td>
</tr>
<tr>
<td>0.3</td>
<td>$a=0.8\sim1.1$</td>
<td>0.0533</td>
<td>0.0607</td>
<td>0.0925</td>
</tr>
<tr>
<td>0.3</td>
<td>$a=1.6\sim1.9$</td>
<td>0.0518</td>
<td>0.0592</td>
<td>0.1401</td>
</tr>
<tr>
<td>0.4</td>
<td>$a=0.3\sim0.7$</td>
<td>0.0612</td>
<td>0.0670</td>
<td>0.0861</td>
</tr>
<tr>
<td>0.4</td>
<td>$a=0.6\sim1.0$</td>
<td>0.0543</td>
<td>0.0617</td>
<td>0.0922</td>
</tr>
<tr>
<td>0.4</td>
<td>$a=1.4\sim1.8$</td>
<td>0.0487</td>
<td>0.0571</td>
<td>0.1289</td>
</tr>
<tr>
<td>0.5</td>
<td>$a=0.3\sim0.8$</td>
<td>0.0638</td>
<td>0.0704</td>
<td>0.0915</td>
</tr>
<tr>
<td>0.5</td>
<td>$a=0.6\sim1.1$</td>
<td>0.0545</td>
<td>0.0627</td>
<td>0.0976</td>
</tr>
<tr>
<td>0.5</td>
<td>$a=1.3\sim1.8$</td>
<td>0.0478</td>
<td>0.0593</td>
<td>0.1267</td>
</tr>
<tr>
<td>0.6</td>
<td>$a=0.2\sim0.8$</td>
<td>0.0787</td>
<td>0.0821</td>
<td>0.1033</td>
</tr>
<tr>
<td>0.6</td>
<td>$a=0.6\sim1.2$</td>
<td>0.0560</td>
<td>0.0647</td>
<td>0.0975</td>
</tr>
<tr>
<td>0.6</td>
<td>$a=1.2\sim1.8$</td>
<td>0.0482</td>
<td>0.0604</td>
<td>0.1247</td>
</tr>
<tr>
<td>0.8</td>
<td>$a=0.2\sim1.0$</td>
<td>0.0797</td>
<td>0.0880</td>
<td>0.1167</td>
</tr>
</tbody>
</table>

*Note.* $a=$discrimination parameter.
<table>
<thead>
<tr>
<th>Range of (a)</th>
<th>Interval of (a)</th>
<th>Thin</th>
<th>Deciles Thick</th>
<th>Quintiles Thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>(a=0.2^\sim1.4)</td>
<td>0.0554</td>
<td>0.0655</td>
<td>0.1083</td>
</tr>
<tr>
<td>0.8</td>
<td>(a=1.0^\sim1.8)</td>
<td>0.0512</td>
<td>0.0617</td>
<td>0.1248</td>
</tr>
<tr>
<td>1.0</td>
<td>(a=0.2^\sim1.2)</td>
<td>0.0783</td>
<td>0.0896</td>
<td>0.1225</td>
</tr>
<tr>
<td>1.0</td>
<td>(a=0.6^\sim1.6)</td>
<td>0.0575</td>
<td>0.0671</td>
<td>0.1142</td>
</tr>
<tr>
<td>1.0</td>
<td>(a=1.0^\sim2.0)</td>
<td>0.0518</td>
<td>0.0623</td>
<td>0.1314</td>
</tr>
<tr>
<td>1.2</td>
<td>(a=0.2^\sim1.4)</td>
<td>0.0778</td>
<td>0.0927</td>
<td>0.1322</td>
</tr>
<tr>
<td>1.2</td>
<td>(a=0.4^\sim1.6)</td>
<td>0.0632</td>
<td>0.0741</td>
<td>0.1244</td>
</tr>
<tr>
<td>1.2</td>
<td>(a=0.8^\sim2.0)</td>
<td>0.0546</td>
<td>0.0642</td>
<td>0.1304</td>
</tr>
<tr>
<td>1.5</td>
<td>(a=0.2^\sim1.7)</td>
<td>0.0775</td>
<td>0.0917</td>
<td>0.1453</td>
</tr>
<tr>
<td>1.5</td>
<td>(a=0.4^\sim1.9)</td>
<td>0.0614</td>
<td>0.0769</td>
<td>0.1342</td>
</tr>
<tr>
<td>1.5</td>
<td>(a=0.5^\sim2.0)</td>
<td>0.0614</td>
<td>0.0746</td>
<td>0.1354</td>
</tr>
<tr>
<td>1.6</td>
<td>(a=0.2^\sim1.8)</td>
<td>0.0769</td>
<td>0.0927</td>
<td>0.1480</td>
</tr>
<tr>
<td>1.6</td>
<td>(a=0.3^\sim1.9)</td>
<td>0.0688</td>
<td>0.0852</td>
<td>0.1416</td>
</tr>
<tr>
<td>1.6</td>
<td>(a=0.4^\sim2.0)</td>
<td>0.0630</td>
<td>0.0767</td>
<td>0.1372</td>
</tr>
<tr>
<td>1.7</td>
<td>(a=0.2^\sim1.9)</td>
<td>0.0777</td>
<td>0.0942</td>
<td>0.1513</td>
</tr>
<tr>
<td>1.7</td>
<td>(a=0.3^\sim2.0)</td>
<td>0.0694</td>
<td>0.0845</td>
<td>0.1436</td>
</tr>
<tr>
<td>1.8</td>
<td>(a=0.2^\sim2.0)</td>
<td>0.0766</td>
<td>0.0952</td>
<td>0.1550</td>
</tr>
<tr>
<td>2.0</td>
<td>(a=0.0^\sim2.0)</td>
<td>0.1037</td>
<td>0.1234</td>
<td>0.1867</td>
</tr>
</tbody>
</table>
Analysis for research question three. What is the pattern of the Type I error rates for the MH and LR procedures displayed when the discrimination parameter for the studied item varies systematically from low (e.g., 0.2) to high (e.g., 2.0) and the non-studied items include and exclude items with low discrimination parameters under large sample size (e.g., 1,000) and large group ability difference (e.g., $\mu_d=1$) using thin matching and deciles thick matching where data are simulated using both 2PL and 3PL IRT models?

Through conducting simulation III (a) and (b), the results of research question 3 for both the 2PL and 3PL IRT models are provided in the following tables. In these two simulations, the first item was always selected as the studied item and the value of discriminating parameter for this given item varied from low (0.2) to high (2.0). In simulation III (a), the non-studied items, that is, the remaining 49 items, included weak items, which are items with low discrimination parameters. The generation interval for non-studied items under this simulation was [0.2 to 2.0]. In simulation III (b), the non-studied items excluded weak items and the generation interval was set to [1.2, 2.0].

Table 10 indicates the false rejection rates for both MH and LR procedures when non-studied items contain both low and high discrimination parameters in the 2PL IRT model. It is displayed graphically in Figure 7. For the MH procedure, the Type I error rates were controlled within the stringent Bradley (1978) criterion interval under the conditions where the discrimination parameter of studied item varied from 0.6 to 1.2 when using thin matching and from 0.5 to 1.0 when using deciles thick matching. The LR procedure exhibited slightly greater inflation than MH in the corresponding conditions.
However, a similar inflation tendency occurred in the LR procedure. The false rejection rates for the LR procedure can be controlled within Bradley’s (1978) stringent criterion interval when the discrimination parameter of studied item is generated from 0.7 to 1.0 under thin matching. For thick matching, the Type I error rates can be controlled within the stringent Bradley criterion interval only under the situation of 0.6 and 0.7, however, the error rates are tolerable from 0.5 to 0.9.

Table 10

<table>
<thead>
<tr>
<th>Parameter $a$ for studied item</th>
<th>MH_thin</th>
<th>MH_deciles</th>
<th>LR_thin</th>
<th>LR_deciles</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.1273</td>
<td>0.1105</td>
<td>0.1406</td>
<td>0.1125</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0825</td>
<td>0.0692</td>
<td>0.0936</td>
<td>0.0707</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0645</td>
<td>0.0536</td>
<td>0.0764</td>
<td>0.0574</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0534</td>
<td>0.0468</td>
<td>0.0620</td>
<td>0.0532</td>
</tr>
<tr>
<td>0.7</td>
<td>0.0478</td>
<td>0.0441</td>
<td>0.0541</td>
<td>0.0526</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0464</td>
<td>0.0476</td>
<td>0.0537</td>
<td>0.0563</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0443</td>
<td>0.0467</td>
<td>0.0524</td>
<td>0.0594</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0462</td>
<td>0.0527</td>
<td>0.0541</td>
<td>0.0683</td>
</tr>
<tr>
<td>1.2</td>
<td>0.0554</td>
<td>0.0734</td>
<td>0.0657</td>
<td>0.0914</td>
</tr>
<tr>
<td>1.4</td>
<td>0.0691</td>
<td>0.0917</td>
<td>0.0808</td>
<td>0.1159</td>
</tr>
<tr>
<td>1.6</td>
<td>0.0864</td>
<td>0.1168</td>
<td>0.0996</td>
<td>0.1482</td>
</tr>
<tr>
<td>1.8</td>
<td>0.1073</td>
<td>0.1491</td>
<td>0.1215</td>
<td>0.1828</td>
</tr>
<tr>
<td>2.0</td>
<td>0.1283</td>
<td>0.1781</td>
<td>0.1462</td>
<td>0.2161</td>
</tr>
</tbody>
</table>

The overall results for both MH and LR procedures are graphed in Figure 7. The patterns for the four conditions exhibit similarities. The figure shows that the false rejection rates decrease as the value of the discrimination parameter for the studied item
increases, however, after the value is beyond a threshold value, the Type I error rates increase again. For instance, the threshold values are 1.0 and 0.8, respectively, under the LR procedure using thin and deciles thick matching.

Additionally, when the value of the discrimination parameter for the given item is extremely small (e.g., 0.2) or extremely large (e.g., 2.0), regardless of DIF detection approaches or matching methods, the Type I error rates were highly inflated. For example, under the MH procedure using thin matching, the Type I error rates were .1273 and .1283 when the discrimination parameters for the studied item are 0.2 and 2.0, respectively. These values are far beyond Bradley’s (1978) stringent criterion interval.

Figure 7. Type I error rates when non-studied items include low discrimination parameters in the 2PL IRT model.
The Type I error rates for both MH and LR procedures when no weak items are included in non-studied items in 2PL IRT model are provided in Table 11. The results for the MH procedure show that the Type I error rates are restricted to Bradley’s stringent criterion interval of [.045, .055], or \( \alpha \pm .1\alpha \), when the discrimination parameter of studied item varies from 1.0 to 2.0 on the condition of thin matching. For deciles thick matching, the false rejection rates are controlled within the Bradley stringent criterion interval when the value of discrimination parameter for the studied item varies from 0.7 to 1.6.

The results for the LR procedure indicated that higher inflated Type I error rates result in contrast to the MH procedure. Nevertheless, the inflation is tolerable when the discrimination parameter of studied item is set from 1.2 to 2.0 using thin matching and from 0.7 to 1.2 using deciles thick matching.

Figure 8 presents the patterns of the false rejection rates for both MH and LR procedures in terms of matching method. It shows that when non-studied items are all strong items, if the discriminating parameter for the studied item is fixed as an extremely small value (e.g., 0.2), a high inflation of Type I error rates will occur independent of DIF investigation techniques or matching methods. For example, under the LR procedure using thin matching, the Type I error rate is .1288, which is far beyond Bradley’s (1978) stringent criterion interval when the discrimination parameter for the given item is 0.2. Review of Figure 8 suggests that the patterns of the four conditions have a similar tendency, that is, as the value of the discrimination parameter for the studied item increases, the false rejection rates decrease dramatically. Nevertheless, Type I error rates
increase slightly at a certain point. For instance, under the MH procedure with thin and deciles thick matching, Type I error rates begin to increase at 1.8 and 1.2, respectively.

Table 11

*Type I Error Rates When Non-studied Items Exclude Low Discrimination Parameters in the 2PL IRT Model*

<table>
<thead>
<tr>
<th>Parameter a for studied item</th>
<th>MH_thin</th>
<th>MH_deciles</th>
<th>LR_thin</th>
<th>LR_deciles</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.1210</td>
<td>0.1030</td>
<td>0.1288</td>
<td>0.1074</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0969</td>
<td>0.0786</td>
<td>0.1035</td>
<td>0.0836</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0830</td>
<td>0.0664</td>
<td>0.0897</td>
<td>0.0686</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0721</td>
<td>0.0592</td>
<td>0.0792</td>
<td>0.0604</td>
</tr>
<tr>
<td>0.7</td>
<td>0.0661</td>
<td>0.0524</td>
<td>0.0729</td>
<td>0.0558</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0607</td>
<td>0.0518</td>
<td>0.0693</td>
<td>0.0565</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0556</td>
<td>0.0459</td>
<td>0.0637</td>
<td>0.0516</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0493</td>
<td>0.0443</td>
<td>0.0589</td>
<td>0.0508</td>
</tr>
<tr>
<td>1.2</td>
<td>0.0454</td>
<td>0.0447</td>
<td>0.0555</td>
<td>0.0537</td>
</tr>
<tr>
<td>1.4</td>
<td>0.0454</td>
<td>0.0476</td>
<td>0.0542</td>
<td>0.0591</td>
</tr>
<tr>
<td>1.6</td>
<td>0.0437</td>
<td>0.0491</td>
<td>0.0518</td>
<td>0.0628</td>
</tr>
<tr>
<td>1.8</td>
<td>0.0456</td>
<td>0.0580</td>
<td>0.0554</td>
<td>0.0721</td>
</tr>
<tr>
<td>2.0</td>
<td>0.0472</td>
<td>0.0620</td>
<td>0.0578</td>
<td>0.0783</td>
</tr>
</tbody>
</table>
Comparison between Table 10 and Table 11 shows that the performance of the Type I error rates is different when the non-studied items include or exclude weak items in 2PL IRT model. For instance, under MH thin matching, when the discrimination parameter for the studied item is 0.6, the false rejection rate is .0534 when weak items are included in non-studied items, while the false rejection rate is inflated to .0721 when all non-studied items are relatively strong. Furthermore, if the value of the discrimination parameter for the studied item is 1.6, the Type I error rate is .0864, beyond Bradley’s (1978) stringent criterion, when both weak and strong items are contained in non-studied items. The Type I error rate is .0437 and within the stringent criterion interval, however, when no weak items exist among non-studied items.
The false rejection rates for both MH and LR procedures under the condition that non-studied items are involved in both weak and strong items in 3PL IRT model were provided in Table 12. It can be seen that the Type I error rates are mostly beyond Bradley’s (1978) stringent criterion interval regardless of the DIF detection techniques and matching methods. Compared to the results of analogous condition in 2PL IRT model, shown in Table 10, it indicates that guessing has an effect on the performance of the Type I error rates. Additionally, review of Table 12 also shows that unlike in 2PL IRT model, deciles thick matching generally performs better than thin matching in 3PL IRT model.

Table 12

<table>
<thead>
<tr>
<th>Parameter a for studied item</th>
<th>MH_thin</th>
<th>MH_deciles</th>
<th>LR_thin</th>
<th>LR_deciles</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.1287</td>
<td>0.1168</td>
<td>0.1454</td>
<td>0.1201</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0666</td>
<td>0.0580</td>
<td>0.0770</td>
<td>0.0618</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0510</td>
<td>0.0485</td>
<td>0.0598</td>
<td>0.0529</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0491</td>
<td>0.0499</td>
<td>0.0552</td>
<td>0.0587</td>
</tr>
<tr>
<td>0.7</td>
<td>0.0526</td>
<td>0.0579</td>
<td>0.0606</td>
<td>0.0699</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0625</td>
<td>0.0708</td>
<td>0.0696</td>
<td>0.0873</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0797</td>
<td>0.0934</td>
<td>0.0893</td>
<td>0.1111</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0971</td>
<td>0.1173</td>
<td>0.1093</td>
<td>0.1392</td>
</tr>
<tr>
<td>1.2</td>
<td>0.1457</td>
<td>0.1708</td>
<td>0.1610</td>
<td>0.1972</td>
</tr>
<tr>
<td>1.4</td>
<td>0.1957</td>
<td>0.2249</td>
<td>0.2159</td>
<td>0.2550</td>
</tr>
<tr>
<td>1.6</td>
<td>0.2438</td>
<td>0.2782</td>
<td>0.2707</td>
<td>0.3139</td>
</tr>
<tr>
<td>1.8</td>
<td>0.2893</td>
<td>0.3258</td>
<td>0.3174</td>
<td>0.3604</td>
</tr>
<tr>
<td>2.0</td>
<td>0.3260</td>
<td>0.3628</td>
<td>0.3567</td>
<td>0.3928</td>
</tr>
</tbody>
</table>
The patterns of the false rejection rates are displayed in Figure 9. All lines show a similar tendency. It illustrates that as the value of discrimination parameter for the studied item increase, the Type I error rates gradually decrease, however, after the value breaks through a single point, the Type I error rates will increase sharply. Some of them are even above 0.3, which is far beyond Bradley’s (1978) stringent criterion interval. Moreover, for the extreme cases, such as the discriminating parameter for the studied item is 0.2 or 2.0, the Type I error rates are highly inflated independent of DIF detection approach or matching methods. For instance, the highest Type I error rate shown in this table reaches .3928 under the condition of deciles thick matching in the LR procedure when the discrimination parameter for the studied item equals to 2.0.

Figure 9. Type I error rates when non-studied items include low discrimination parameters in the 3PL IRT model.
Table 13 shows the false rejection rates for both MH and LR procedures under the condition that non-studied items only include strong items in 3PL IRT model. They are also graphically presented in Figure 10.

Similarly, almost all the Type I error rates are beyond Bradley’s (1978) stringent criterion interval independent of the DIF detection techniques and matching methods. It confirms that guessing appears to be a nuisance determinant on the Type I error rates. Additionally, the performance of deciles thick matching is also generally better than thin matching.

Table 13

<table>
<thead>
<tr>
<th>Parameter $a$ for studied item</th>
<th>MH thin</th>
<th>MH deciles</th>
<th>LR thin</th>
<th>LR deciles</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.1225</td>
<td>0.1112</td>
<td>0.1433</td>
<td>0.1160</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0917</td>
<td>0.0813</td>
<td>0.1078</td>
<td>0.0840</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0781</td>
<td>0.0674</td>
<td>0.0924</td>
<td>0.0724</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0717</td>
<td>0.0620</td>
<td>0.0844</td>
<td>0.0657</td>
</tr>
<tr>
<td>0.7</td>
<td>0.0648</td>
<td>0.0562</td>
<td>0.0776</td>
<td>0.0597</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0611</td>
<td>0.0542</td>
<td>0.0746</td>
<td>0.0576</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0583</td>
<td>0.0536</td>
<td>0.0697</td>
<td>0.0579</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0589</td>
<td>0.0526</td>
<td>0.0690</td>
<td>0.0580</td>
</tr>
<tr>
<td>1.2</td>
<td>0.0604</td>
<td>0.0574</td>
<td>0.0713</td>
<td>0.0656</td>
</tr>
<tr>
<td>1.4</td>
<td>0.0640</td>
<td>0.0643</td>
<td>0.0770</td>
<td>0.0729</td>
</tr>
<tr>
<td>1.6</td>
<td>0.0674</td>
<td>0.0727</td>
<td>0.0827</td>
<td>0.0827</td>
</tr>
<tr>
<td>1.8</td>
<td>0.0771</td>
<td>0.0842</td>
<td>0.0952</td>
<td>0.0954</td>
</tr>
<tr>
<td>2.0</td>
<td>0.0851</td>
<td>0.0947</td>
<td>0.1089</td>
<td>0.1074</td>
</tr>
</tbody>
</table>
The patterns of the results in the four conditions conducted in this simulation exhibit a similar tendency. That is, as the value of the discrimination parameter for the studied item increases, the false rejection rates gradually decrease, nevertheless, after it is beyond a threshold point, the Type I error rates will increase smoothly. Moreover, when the value of discriminating parameter for the given item is extremely small (e.g., 0.2) or extremely large, regardless of DIF detection approach or matching methods, the inflation of Type I error rates are relatively high. For example, under the MH thin matching, the Type I error rates are .1225 and .0851, respectively, when the discrimination parameters for the studied item are 0.2 and 2.0. Both of them are far beyond Bradley’s (1978) stringent criterion interval.

Figure 10. Type I error rates when non-studied items exclude low discrimination parameters in the 3PL IRT model.
Comparing Table 12 with Table 13 suggests that the performance of the Type I error rates is different when the non-studied items include or exclude weak items in the 3PL IRT model. For example, under the condition of MH thin matching, when the discrimination parameter for the studied item is 0.6, the false rejection rate is .0491, which is controlled fairly well when weak items were included in non-studied items. The false rejection rate is .0717 and beyond Bradley’s (1978) stringent criterion interval, however, when no weak items existed in non-studied items. In addition, if the value of the discrimination parameter for the studied item is 1.6, the Type I error rate was .2438, far beyond the stringent criterion interval under the condition where non-studied items included both weak and strong items. The Type I error rate is .0674, which indicated mild inflation, when all non-studied items were relatively strong.
Chapter 5: Discussion, Conclusions and Recommendations

The results obtained from the current Monte Carlo study will be discussed in this chapter organized according to the three research questions proposed previously. Moreover, several conclusions will be reached and a few recommendations will be provided to practitioners.

Discussion of the Results

The aim of the current research was to examine how item discrimination parameters affect Type I error rates while manipulating other factors such as sample size, group mean ability, and matching method in both MH and LR procedures. The first research question explored the pattern of the Type I error rates under varied sample size and group mean ability when item discrimination parameters were generated from the entire reasonable interval from 0.0 to 1.8. This research question provided the researcher an entire picture of how Type I error rates performed when the factors mentioned above were manipulated.

Research question 1 (RQ1). What is the pattern of the Type I error rates displayed in the MH and LR procedures under thin matching, deciles thick matching, and quintiles thick matching with different sample sizes and varied group ability differences with data generated based on 2PL IRT models?

The results of the first research question showed that the Type I error rates for both MH and LR procedures tended to increase as the group mean differences increased, independent of sample size and matching methods. These results indicated that differences in the group ability distributions had an effect on the occurrence of false
rejection rates, which is consistent with many previous studies (e.g., Güler & Penfield, 2009; Mazor et al., 1992; Shealy & Stout, 1993).

Additionally, an important observation of this research question was that larger sample sizes tended to yield larger Type I error rates if other factors were manipulated no matter what type of DIF approaches were applied. This finding is also in agreement with some earlier studies (e.g., DeMars, 2009; Roussos & Stout, 1996; Swaminathan & Rogers, 1990).

Furthermore, Donoghue and Allen (1993) explored the performance of seven matching strategies including thin matching and quintiles matching, called “percent of total sample matching variable” in that study (p. 136). It indicated that thin matching performed better than quintiles matching when test length was relatively long (e.g., 20 or 40 items). In the present study, thin matching performed better than thick matching methods including both deciles and quintiles methods, that is, thin matching yielded the least inflated Type I error rates regardless of sample size and DIF detection method. Additionally, deciles thick matching performed better than quintiles thick matching. In other words, the performance of quintiles thick matching was the worst among the three matching methods. The finding that thin matching performs better than thick matching may be explained by the reason that thin matching covers more information. In addition, it may also be because thick matching is likely to confound DIF and impact and increase statistical bias. The performance of thick matching is not satisfactory with longer tests (Donoghue & Allen, 1993).
Finally, the results obtained from RQ1 indicated that the LR procedure generally showed more inflation than the MH procedure. That is to say, independent of sample size and matching methods, compared to the MH procedure, slightly larger false rejection rates occurred in the LR procedure in this study. This confirms the results of DeMars (2009).

Based on the results of the first research question that larger sample size and larger group ability difference were prone to yield larger Type I error rates, the second research question was constructed to study only large sample size and large group ability difference. The second research question sought to investigate how item discrimination parameters affect the performance of the false rejection rates when other factors such as test length, item difficulty parameter, sample size, and group mean difference were manipulated. The second research question is described as follows.

**Research question 2 (RQ2).** What is the pattern of the Type I error rates for the MH and LR procedures displayed when the range of discrimination parameters for all items varies from very small to relatively large under large sample size (e.g., 1,000) and large group ability difference (e.g., $\mu_d=1$) using thin matching, deciles thick matching, and quintiles thick matching with data simulated according to 2PL IRT models?

Examination of the results for the MH procedure indicated several findings. One of the most interesting observations from this research question is that under thin matching, Type I error rates can be controlled within a small range around the nominal significance level (.05) when the range of all item discrimination parameters is very narrow, no matter whether the items are relatively weak or strong. For example, the false
rejection rate is 0.0441, indicating no inflation, when all item discrimination parameters are generated in the intervals [0.5, 0.7]. It can be seen that under this circumstance that all items are relatively weak. When discrimination parameters for all items are generated in the interval [1.6, 1.8], indicating all items are relatively strong, the Type I error rate is 0.0418, also no inflation. Holland and Thayer (1988) claimed that the inflation of Type I error is not a big concern in the one-parameter (or Rasch) IRT model for the MH procedure because the MH hypothesis is in accordance with the null hypothesis that there is no DIF when data follow a Rasch model. Roussos and Stout (1996) made a further explanation and verification on this viewpoint. The finding in this study that the inflation can be maintained fairly well when all item discrimination parameters are simulated in a small interval may be because it resembles a Rasch model when discrimination parameters for all items are relatively homogeneous. This finding provides empirical evidence to support the study of Zwick (1990).

Another interesting finding is that strong discrimination parameters for all items are expected to yield lower false rejection rates under both thin matching and deciles thick matching methods, however, deciles thick matching shows only a slightly higher inflation than thin matching does. No previous literature has reported a similar finding about deciles thick matching to what was found in the current study. Also, results represented that all the Type I error rates in quintiles thick matching are beyond the stringent criterion interval [.045, .055], $\alpha \pm 1\alpha$ (Bradley, 1978). These findings reveal that thin matching performs better in regard to Type I error than either thick matching used here and that quintiles thick matching performs least well.
Additionally, the MH odds ratio was also examined in the second simulation. It can be seen that when there is no mean ability difference between reference group and focal group, neither group is favored by DIF. However, when group mean difference exists, higher discrimination items tend to favor the group with higher ability, and lower discrimination items are prone to favor the focal group, which is the group with poorer ability. This finding is consistent with Zwick’s (1990) study.

Review of the results for the LR procedure showed similar findings as those of the MH procedure, except that the LR procedure almost always suffered more inflation. First, under thin matching, the Type I error rates adhered to Bradley’s (1978) stringent criterion interval, $\alpha \pm .1\alpha$, when all item discrimination parameters were generated in a small range (0.2). It also can be explained by the same reason as in the MH procedure that data fit a Rasch model when all item discrimination parameters are relatively homogenous. Second, similar to thin matching in the MH procedure, when all items were simulated with relatively strong discrimination parameters, the Type I error rates tended to be controlled within the stringent criterion interval, independent of the range of generation interval for all item discrimination parameters.

Third, thick matching methods generally displayed inflated Type I error rates no matter the interval of discrimination parameters for all items, particularly under quintiles thick matching. Nevertheless, deciles thick matching method showed a similar tendency as that of thin matching, with the exception of higher inflation under deciles thick matching.
Through investigating the second research question, the influence of discrimination parameters for all items on the pattern of the false rejection rates in MH and LR procedures was identified. Until now, no previous literature investigated how the discrimination index values of studied items and non-studied items is relevant to the Type I error rates. The third research question was therefore created to study this problem. Because the results obtained from the second research question showed that quintiles thick matching performed poorly, only thin matching and deciles thick matching were considered in the third research question. Additionally, both 2PL and 3PL IRT models were examined in the third research question because the 3PL IRT model is commonly applied in a variety of domains.

**Research question 3 (RQ3).** What is the pattern of the Type I error rates for the MH and LR procedures displayed when the discrimination parameter for the studied item varies systematically from low (e.g., 0.2) to high (e.g., 2.0) and the non-studied items include and exclude items with low discrimination parameters under large sample size (e.g., 1,000) and large group ability difference (e.g., μd=1) using thin matching and deciles thick matching where data are simulated using both 2PL and 3PL IRT models?

In the 2PL IRT model, when the value of discrimination parameter for the studied item varied from a weak item to a strong item and non-studied items included weak items, four patterns, for thin and deciles under the MH procedure and for thin and deciles under the LR procedure, were demonstrated in Figure 7. These patterns suggest that the four conditions exhibit the similar tendency; however, the intervals of the discrimination parameter for the studied item which are able to avoid inflated Type I error are slightly
different under the four conditions. For instance, under thin matching and deciles thick matching in the MH procedure, the false rejection rates can be maintained well when the studied item discrimination is located in the interval \([0.6, 1.2]\) and \([0.5, 1.0]\), respectively. It is clearly noted that the interval with no inflation for thin matching is slightly wider than for deciles thick matching, which is reasonable because more information is contained in thin matching. Figure 7 also illustrates that the LR procedure generally suffers more inflation and that the interval of no inflation is narrower than that of the MH procedure under corresponding conditions independent of matching method, which is not surprising. It is consistent with the study of DeMars (2009). In addition, an interesting finding is that regardless of DIF detection procedures or matching methods, the highest inflated false rejection rates will occur when the studied item has an extremely large (e.g., 2.0) discrimination index.

The patterns of the Type I error rates in the 2PL IRT model when non-studied items excluded weak items were displayed in Figure 8, which are rather different from those in Figure 7. Only extremely small discrimination indices for the studied item tends to yield inflated Type I error rates. The intervals of no inflation for the four conditions all shift toward high discrimination indices. For example, the interval is \([1.0, 2.0]\) for thin matching and \([0.7, 1.6]\) for deciles thick matching in the MH procedure. It can be seen that in contrast to the condition where non-studied items include weak items, the false rejection rates can be controlled fairly well when the studied item has a relatively higher discrimination value. Discrimination parameters for non-studied items were generated from 1.2 to 2.0 in this case. These results seem reasonable because the Type I error rates
tend to perform well when the value of discrimination parameter for the studied item is close to the distribution of discrimination indices for non-studied items. Nevertheless, there are several similar performances as that in Figure 7 needed to be mentioned. First of all, the four conditions in this case also exhibit a similar tendency. Second, thin matching has a wider interval with no inflated Type I error rates than that of thick matching independent of DIF investigation method. Third, the interval with no inflation for the LR procedure is narrower than that for the MH procedure regardless of matching methods.

In the 3PL IRT model, the most interesting finding is that almost all the Type I error rates inflate no matter what type of DIF detection methods and matching methods are used, and whether non-studied items include or exclude weak items. This finding is in agreement with the study of DeMars (2010) that the guessing has an effect on DIF detection rate when large group mean differences exist. Moreover, the patterns in 3PL IRT model appear to show a similar tendency as in the 2PL IRT model under analogous conditions no matter whether the non-studied items include or exclude weak items. For instance, extremely large discrimination values for the given item are expected to yield the highest inflated false rejection rates when both weak and strong items are involved in non-studied items. Extremely small discrimination parameter values for the given item are prone to result in the highest inflation of Type I error rates when no weak items exist in non-studied items. The similar patterns further confirm the rationale of the results obtained from prior simulations. However, because the existence of guessing, all the curves in the 3PL IRT model increase more sharply than those in the 2PL IRT model.
Conclusions

According to earlier studies, a number of factors are able to affect the performance of the Type I error rates in the MH and LR procedures. Many influencing factors such as test length, sample size, group mean difference, and the percentage of items with DIF have been examined by previous researchers (e.g., Güler & Penfield, 2009; Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). However, no studies had been conducted to investigate how the discrimination index effects Type I error rates when other factors are manipulated. The current Monte Carlo study sought to investigate the relationship between the item discrimination parameters and false rejection rates in DIF detection. The main conclusions are summarized according to the results and findings.

First, larger sample size and larger group mean differences are expected to yield inflated Type I error rate regardless of DIF detection techniques and matching methods based on the patterns displayed in Figure 3 and Figure 4. Specifically, when all the influencing factors except sample size are manipulated, in both MH and LR procedures, the Type I error rates become higher when sample size becomes larger. Meanwhile, if controlling all the influencing factors with the exception of group mean differences, the false rejection rates for both MH and LR procedures will increase as group mean differences increase.

Second, thin matching is generally preferable for both MH and LR procedures based on the first and the second research question. Examination of all the simulations in this study indicates that thin matching shows generally lower inflated false rejection rates
than thick matching methods does. Quintiles thick matching is the worst and almost always suffers the largest inflated Type I error rates. The performance of deciles thick matching are acceptable in many circumstances; however, the deciles thick matching generally led to higher false rejection rates than thin matching no matter how sample sizes, group mean differences, and item discrimination indices were manipulated.

Third, for all simulations constructed in the current study, the LR and MH procedures exhibit similar patterns of the Type I error rates. Type I error inflation, however, is generally somewhat higher in the LR procedure than in the MH procedure.

Fourth, in both MH and LR procedures, a small range (e.g., 0.2 to 0.4) of discrimination parameters for all items is a far less serious issue in causing inflated false rejection rates for thin matching in light of Table 8 and Table 9, even with large sample size and large group mean differences. In other words, the Type I error rates are controlled fairly well when discrimination indices for items are relatively homogenous no matter whether the items are relatively weak or relatively strong. For deciles thick matching, the MH procedure still performs well while the LR procedure suffers slightly higher inflation in these conditions. Additionally, in both MH and LR procedures, all items with relatively high discrimination parameters are prone to yield less inflation independent of the range of discrimination parameters for all items when using thin matching and deciles thick matching. Even though the expansion of the interval of item discrimination parameters led to greater inflated false rejection rates, the Type I error rates can be maintained within or close to the stringent Bradley (1978) criterion interval [0.045, 0.055], if all items are relatively strong, especially for thin matching.
Fifth, in a 2PL IRT model, when non-studied items include both weak and strong items, the Type I error rates are able to be controlled well when the discrimination parameter for the studied item is located in a certain interval. The interval varies among the combinations of DIF detection techniques and matching methods. For example, the intervals are $[0.6, 1.2]$, $[0.5, 1.0]$, $[0.7, 1.0]$ and $[0.6, 0.7]$ under thin matching in MH, deciles thick matching in MH, thin matching in LR and deciles thick matching in LR, respectively. Additionally, for the condition that the non-studied items exclude weak items, that is, the condition that all the non-studied items have relatively high discrimination parameters, the false rejection rates can be controlled fairly well when the studied item has a relatively larger discrimination value, adhering to the distribution of discrimination indices for non-studied items. For example, the interval of discrimination parameters for the studied item that avoids inflated Type I error rates varies from $[0.6, 1.2]$ to $[1.0, 2.0]$ under thin matching in MH procedure depending on non-studied items including or excluding weak items, respectively.

Finally, in contrast to the 2PL IRT model, the false rejection rates in the 3PL IRT model are intended to inflate due to the existence of guessing through comparing Table 10 with Table12, and Table 11 with Table 13. That is to say, guessing is likely to have an influence on the inflation of Type I error rates. Nevertheless, the patterns of false rejection rates are similar to those in the 2PL IRT model under corresponding conditions.

Overall, according to the present Monte Carlo study, in addition to sample size and group mean difference, item discrimination indices do have an effect on Type I error rates. Generally speaking, relative homogeneity of discrimination parameters for all items
is helpful in reducing the occurrence of inflated false rejection rates. Assuming that item
discrimination parameters are heterogeneous, Type I error inflation is not a serious issue
if all items are relatively strong. When manipulating these influencing factors, thin
matching generally performs best and quintiles thick matching performs worst in both
MH and LR procedures. Moreover, the value of the discrimination parameter for the
studied item and the non-studied item discrimination indices do have an influence on
Type I error inflation. Finally, guessing is a concern that may yield inflated false rejection
rates.

**Recommendations**

A number of previous researchers have made great efforts to investigate the effect
of various influencing factors on the performance of DIF detection rates in both MH and
LR procedures. The current study was conducted primarily to examine the influence of
discrimination indices on Type I error rates. Several observations and findings from the
study are able to provide certain practical implications to practitioners. Meanwhile,
recommendations will be made for future study due to the existence of limitations and
delimitations in the present research.

**Recommendations for practitioners.** Type I error inflation commonly occurs in
DIF analysis when using the MH and LR procedures and the inflated false rejection rates
do have damaging effects (Roussos & Stout, 1996). The findings from this study are able
to provide researchers some recommendations from practical perspectives.

One observation is that larger sample size is prone to yield larger false rejection
rates with group impact. In order to control the inflation of Type I error, it seems that
researchers should limit the sample size; however, power will decrease as sample size decreases. Therefore, DIF scholars need to strike a balance when they decide the number of examinees.

Larger group mean differences are expected to cause larger Type I error rates. Generally, no inflation exists under equal group mean distributions; however, the false rejection rates tend to inflate when group impact expands, especially with large sample size. Of course, in some practical research, ability differences between reference group and focal group do exist. This observation reminds DIF researchers that when they select a large group of examinees, it is better for them to control group impact within a proper range before conducting a DIF analysis.

In this research, in general, the performance of thin matching is better than thick matching while quintiles thick matching generally performs worst and led to the highest inflated Type I error, indicating that thin matching is the best choice in DIF analysis. Nevertheless, in some certain circumstances, thick matching is needed, particularly for the MH procedure under the condition of short test and small number of examinees (Donoghue & Allen, 1993). Based on the results from the present study, deciles thick matching appears to be a good choice because its performance is generally acceptable. Quintiles thick matching might be ignored since excessively high Type I error inflation occurs. According to the study of Donoghue and Allen (1993), three matching methods displayed the best performance: “equal interval matching”, “censoring matching variable” and “minimum frequency matching” (p. 151). Consequently, before conducting DIF research, scholars should select an appropriate matching method.
Examination of the results from the 2PL IRT model indicates that when groups are equivalent on ability, there is no Type I error inflation concern. When groups are not equivalent in abilities, homogeneity of discrimination parameters for all items helps control the Type I error rates well. Holland and Thayer (1988) pointed out that the Type I error inflation is not a potential problem in the one-parameter logistic (or Rasch) model for the MH procedure. There is no doubt that data that fit a Rasch model is the best choice to prevent the inflated false rejection rates; however, the 2PL IRT model is commonly needed and used in a variety of fields. The findings remind researchers that if a 2PL model is used and weak items are included in an assessment, an excessively wide interval of item discrimination indices should be avoided in order to better control the false rejection rates. The results also show that all items with relatively high discrimination parameters are prone to yield less inflation even when the interval of all item discrimination parameters is relatively wide. Therefore, another way to reduce Type I error inflation is for researchers to only design or select strong items, that is, items with relatively high discrimination parameters. Under this situation, the permitted interval of item discrimination indices is much wider.

The present study illustrates how the discrimination parameter for the studied item and discrimination indices for non-studied items affect the performance of the Type I error rates in both the 2PL and 3PL IRT models. It can be seen that when the discrimination parameter for the studied item is located within certain intervals, the Type I error inflation can be controlled or reduced no matter whether non-studied items include or exclude weak items. Roussos and Stout (1996) stated that inflated Type I error rates
would lead test developers to remove some good items. The results from this study remind researchers that sometimes the Type I error inflation may occur only because the discrimination index for the studied item is not appropriate rather than it is a bad item.

It can be seen that the MH procedure generally works slightly better than the LR procedure in this study, which is consistent with some prior studies (e.g., DeMars, 2009). The MH procedure has advantages of easy access, simple computation, and less limitation on sample sizes selection (Narayanan & Swaminathan, 1996). If an investigator is only concerned with uniform DIF detection, the MH procedure would be a better choice to avoid type I error inflation. Though the LR procedure generally suffers more inflation, it is sensitive to both uniform and nonuniform DIF. Therefore, the LR procedure should be considered when nonuniform DIF must be investigated.

**Recommendations for future research.** Admittedly, this Monte Carlo study was not able to explore all issues related to DIF and further research is needed to be conducted in several aspects. First, the primary purpose of this study was to investigate the performance of Type I error rates on DIF detection; consequently, no DIF items were generated in this research. However, some literature mentioned that the percentage of items with DIF affected the false rejection rates (e.g., Jodoin & Gierl, 2001). Therefore, in future studies, different levels of DIF contamination should be considered.

Second, there are a variety of matching methods proposed by previous scholars. Raju et al. (1989) conducted 2, 4, 6, 8 and 10 score groups as matching variables on a 40-item vocabulary test. Donoghue and Allen (1993) designed seven forms of matching variable including “Thin matching variable”, “Equal interval matching variable”,
“Percent of total sample matching variable”, “Percent of focal sample matching variable”, and three types of “Censored matching variable”, three types of “Minimum cell frequency matching variable” and “No matching variable” (p. 136). Nevertheless, only three matching methods, thin matching, deciles thick matching and quintiles thick matching, are examined in this study. In the current study, quintiles thick matching performed rather poorly. More matching methods should be explored in the future research because thick matching is necessary to be used in some conditions, especially for the MH procedure with a short test and few examinees (Donoghue & Allen, 1993).

Third, there are a number of methods available for detecting DIF. The LR procedure is widely used in recent DIF analysis because it is sensitive to both uniform DIF and nonuniform DIF. The results of this study indicated that the LR procedure suffered more inflation than the MH procedure, which is in agreement with prior literature. In the current study, no interaction of ability and group membership was considered, that is, interaction may not be a reason for yielding higher Type I error inflation for LR procedure based on the present study. Nevertheless, no explanation can be provided and further research should be conducted to examine primary reasons for this issue. Additonally, DeMars (2009) conducted a Monte Carlo study to investigate the performance of the Type I error rates not only using MH, LR, SIBTEST, but using linear regression-corrected MH and LR procedures. Only two approaches, the MH and LR procedures, were examined in this study. Further study can explore the effect of item discrimination parameters in other DIF detection methods.
Fourth, test length is a significant factor influencing DIF detection rates. Narayanan and Swaminathan (1996) stated that “Test length can have an impact on DIF detection rates because a longer test is likely to produce more reliable scores and, hence, more accurate ability estimates” (p. 263). Additionally, Whitmore and Schumacker (1999) mentioned that test length must be considered since longer tests were intended to yield more reliable scores. Therefore, short, medium and long tests should be involved in future research.

Fifth, the current study indicated that Type I error inflation is not a big concern if all items are relatively strong, even if item discrimination parameters are heterogeneous. However, in practice, sometimes it is not possible to avoid weak items involved in a test, therefore, future research should investigate effective methods for minimizing Type I error inflation. For example, researchers might set aside weaker items temporarily only for DIF analyses rather than deleting these items completely from the test. That is, only items with higher discrimination values would be included in the DIF analyses (a different form of item purification). Additionally, if scholars have less confidence in items, it is necessary to explore how to strike a balance among influencing factors such as sample size, test length, group mean differences, and matching methods in order to control inflated Type I error. For instance, if researchers use thin matching in the MH procedure with nonequivalent groups when they are not confident of the quality of all items, longer tests might be considered. Such adaptations must be studied further.

Sixth, Herrera and Gómez (2008) examined how the combination of unbalanced group size and item discrimination indices affected the performance of the Type I error
rates. However, only low and moderate discrimination parameters were applied in their studies. Even though wider intervals of discrimination parameters were considered in the current research, only equal sample size for the reference and focal groups was used. Consequently, similar studies with unequal group sizes should be conducted.

Seventh, a variety of non-linear patterns are shown for the second and the third research questions (e.g., Figures 7 and 8) and several useful findings and conclusions are obtained; however, the author is unable to provide an interpretation on the reasons for these patterns. Further study is necessary to investigate the issue.

Eighth, DeMars (2010) pointed out that guessing was a nuisance for the accuracy of detecting DIF, especially with large group mean differences. This study only confirms that guessing does have an influence on inflated Type I error. Further study should focus more on how the guessing affects the false rejection rates so that effective approaches can be developed to control the Type I error inflation in 3PL IRT model.

Finally, a good number of early DIF research used Monte Carlo studies to explore the performance of the Type I error in MH and LR procedures; however, some of them used replications less than 1,000 for each conditions. For example, only 100 trials were applied in the study of Narayanan and Swaminathan (1996) and Jodoin and Gierl (2001), and 200 trials were used by Güler and Penfield (2009). As stated early, sufficient replications as 10,000 trials for each condition in the present study can ensure the stability and generalizability of the results. Consequently, researchers should select appropriate replications when they conduct Monte Carlo studies.
References


### Appendix A: R Code of RQ2 for Thin Matching and Deciles Thick Matching

```r
### remove memory and set seed for generation
rm(list=ls())
set.seed(28760103)

### install packages
library(irttoy)
library(difR)
library(vcd)
library(CMC)

# initialize various variables
aI<-0
bI<-0
abI<-0
num_r<-0
reliab<-0
MH_typeI<-0

mh_pvalueI<-0
log_pvalueI<-0
log_pvalue_rawI<-0
wf_pvalueI<-0

sum_MH_pvalue<-0
power_MH_pvalue<-0
sum_mh_pvalueI<-0
power_mhI<-0
sum_wf_pvalueI<-0
power_wfI<-0
sum_log_pvalueI<-0
power_logI<-0
sum_log_pvalue_rawI<-0
power_log_rawI<-0

### set IRT parameters values
a1=1.6
a2=1.8
b1=-2.0
b2=2.0
c1=0.0
c2=0.0

### set replication, sample size and group mean difference
rep<-10000
samplen<-1000
```
iterations<-samplen * 2  # iteration equals to double sample size
mean_difference<-0

###generate data
for(t in 1:rep)
{
  ###for Type I error rates
  DifparameterI<-cbind(runif(50,a1,a2),runif(50,b1,b2),runif(50,c1,c2))  ## 50 items for no DIF
  ##use IRTparameters to generate data responses
  DifresponseI<-sim(ip=DifparameterI,x=rnorm(samplen,mean=0,sd=1))
  DifresponseII<-sim(ip=DifparameterI,x=rnorm(samplen,mean=-1,sd=1))
  DifresponseII<-rbind(DifresponseI,DifresponseII)

  ###calculate total scores for examinees
  rawscoreI<-rowSums(DifresponseI)
  rawscoreII<-rowSums(DifresponseII)
  rawscoreI<-as.matrix(rawscoreI)
  rawscoreII<-as.matrix(rawscoreII)
  rawscoreIII<-c(rawscoreI,rawscoreII)

  ###build up dataset
  aI<-DifresponseI[,1]
  bI<-DifresponseII[,1]
  groupI<-rep(0:1,each=samplen)                  ##equals to sample size
  groupI
  abI<-c(aI,bI)
  abI
  class(abI)
  dataIII<-cbind(groupI,DifresponseIII)
  m_groupI<-as.matrix(groupI)
  m_abI<-as.matrix(abI)
  dataI<-cbind(m_groupI,m_abI)
  dataI
  class(dataI)
  reliab<-alpha.cronbach(DifresponseIII)

  ###use deciles thick matching in MH method (rejection rate)
  ###build up scoregroup
  quantile(rawscoreIII, probs=c(.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
  quant <- quantile(rawscoreIII, probs=c(.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
  quant <- as.matrix(quant)
  quant
  scoregroupI<-matrix(0,iterations,1)
for (p in 1:iterations) 
{
  if (rawscoreIII[p,1]>=0 & rawscoreIII[p,1]<=quant[1])          {scoregroupI[p,1]<-1}
  else     {scoregroupI[p,1]<-99}
}
##scoregroupI
datasetI<-cbind(dataI,scoregroupI)
datasetI

###forming dataset for quintiles thick matching MH methods
###initialize cell variables
q1<-0
q2<-0
q3<-0
q4<-0
q5<-0
q6<-0
q7<-0
q8<-0
q9<-0
q10<-0
q11<-0
q12<-0
q13<-0
q14<-0
q15<-0
q16<-0
q17<-0
q18<-0
q19<-0
q20<-0
q21<-0
q22<-0
q23<-0
q24<-0
q25<-0
q26<-0
q27<-0
q28<-0
q29<-0
q30<-0
q31<-0
q32<-0
q33<-0
q34<-0
q35<-0
q36<-0
q37<-0
q38<-0
q39<-0
q40<-0

for (r in 1:iterations)
{
  if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==1) {q1<-q1+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==1) {q2<-q2+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==0&datasetI[r,3]==1) {q3<-q3+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==2) {q4<-q4+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==2) {q5<-q5+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==0&datasetI[r,3]==2) {q6<-q6+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==3) {q7<-q7+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==3) {q8<-q8+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==4) {q9<-q9+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==4) {q10<-q10+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==5) {q11<-q11+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==5) {q12<-q12+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==6) {q13<-q13+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==6) {q14<-q14+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==7) {q15<-q15+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==7) {q16<-q16+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==8) {q17<-q17+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==8) {q18<-q18+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==9) {q19<-q19+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==9) {q20<-q20+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==10) {q21<-q21+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==10) {q22<-q22+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==11) {q23<-q23+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==11) {q24<-q24+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==12) {q25<-q25+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==12) {q26<-q26+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==13) {q27<-q27+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==13) {q28<-q28+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==14) {q29<-q29+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==14) {q30<-q30+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==15) {q31<-q31+1}
  else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==15) {q32<-q32+1}
  else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==16) {q33<-q33+1}
```r
else if (datasetI[r,1]==1 & datasetI[r,2]==1 & datasetI[r,3]==9) {q34<-q34+1}
else if (datasetI[r,1]==0 & datasetI[r,2]==0 & datasetI[r,3]==9) {q35<-q35+1}
else if (datasetI[r,1]==1 & datasetI[r,2]==0 & datasetI[r,3]==9) {q36<-q36+1}
else if (datasetI[r,1]==0 & datasetI[r,2]==1 & datasetI[r,3]==10) {q37<-q37+1}
else if (datasetI[r,1]==1 & datasetI[r,2]==1 & datasetI[r,3]==10) {q38<-q38+1}
else if (datasetI[r,1]==0 & datasetI[r,2]==0 & datasetI[r,3]==10) {q39<-q39+1}
else if (datasetI[r,1]==1 & datasetI[r,2]==0 & datasetI[r,3]==10) {q40<-q40+1}
}
if (q1+q2+q3+q4==0 | q5+q6+q7+q8==0 | q9+q10+q11+q12==0
  | q13+q14+q15+q16==0 | q17+q18+q19+q20==0 | q21+q22+q23+q24==0
  | q25+q26+q27+q28==0 | q29+q30+q31+q32==0 | q33+q34+q35+q36==0
  | q37+q38+q39+q40==0)
  {mh_pvalueI[t]<-NA
   num_r<-num_r+1}
else
  {
    datasetnewI<-array(c(q1,q2,q3,q4,
      q5,q6,q7,q8,
      q9,q10,q11,q12,
      q13,q14,q15,q16,
      q17,q18,q19,q20,
      q21,q22,q23,q24,
      q25,q26,q27,q28,
      q29,q30,q31,q32,
      q33,q34,q35,q36,
      q37,q38,q39,q40),
    dim = c(2, 2, 10),
    dimnames = list(
      group = c("reference", "focal"),
      Response = c("succeed", "failure"),
      group.Level = c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10"))
    datasetnewI

    ### pull out p-value of MH
    mantelhaen.test(datasetnewI)
    mantelhaen.test(datasetnewI)$p.value
    mh_pvalueI[t]<-mantelhaen.test(datasetnewI)$p.value
    wf_pvalueI[t]<-woolf_test(datasetnewI)$p.value

    ### extract the odds ratio
    mantelhaen.test(datasetnewI)$estimate
    woolf_test(datasetnewI)

    ### use thin matching and deciles thick matching in LR procedure
    logdata1I<-data.frame(abI,groupI,scoregroupI)
    logdata2I<-data.frame(abI,groupI,rawscoreIII)
    logdatatrial1I<-glm(abI~groupI+scoregroupI,family=binomial(link=logit),data=logdata1I)
    logdatatrial2I<-glm(abI~groupI+rawscoreIII,family=binomial(link=logit),data=logdata2I)
```

summary(logdatatrial1I)                                ##using scoregroup in LR
summary(logdatatrial1I)$coefficients
summary(logdatatrial1I)$coefficients[2,4]
summary(logdatatrial2I)                               ##using rawscore in LR
summary(logdatatrial2I)$coefficients
summary(logdatatrial2I)$coefficients[2,4]

###pull out p-value of LR
log_pvalueI[t]<-summary(logdatatrial1I)$coefficients[2,4]
log_pvalue_rawI[t]<-summary(logdatatrial2I)$coefficients[2,4]
summary(logdatatrial1I)
###count rejection number
{
  if (mh_pvalueI[t]<0.05)        {sum_mh_pvalueI<-sum_mh_pvalueI+1
    mh_ratioI[t]<-mantelhaen.test(datasetnewI)$estimate
    para_aI[t]<-DifparameterI[1,1]
    para_bI[t]<-DifparameterI[1,2]}
if (mh_pvalueI[t]>0.05)        {para_aI[t]<-NA
  para_bI[t]<-NA
  mh_ratioI[t]<-NA}

if (log_pvalueI[t]<0.05)             {sum_log_pvalueI<-sum_log_pvalueI+1}
if (log_pvalue_rawI[t]<0.05)    {sum_log_pvalue_rawI<-sum_log_pvalue_rawI+1}
if (wf_pvalueI[t]<0.05)              {sum_wf_pvalueI<-sum_wf_pvalueI+1}
}
### calculate the rejection rate (the Type I error rate) in thin matching and quintiles thick matching
power_MH_pvalue<-sum_MH_pvalue/rep
power_mhI<-sum_mh_pvalueI/(rep-num_r)
power_wfl<-sum_wf_pvalueI/(rep-num_r)
power_logI<-sum_log_pvalueI/(rep-num_r)
power_log_rawI<-sum_log_pvalue_rawI/(rep-num_r)

###show the results
power_MH_pvalue
power_mhI
power_log_rawI
power_logI

###save results
typeI<-data.frame(a1,a2,mean_difference,samplen,power_MH_pvalue,power_mhI,power_wfl,
power_log_rawl,power_logl)

write.table (type1,file="result_DIF/1000_type_I_error_10thick and thin_1.3_1.8.csv",quote=FALSE,sep=","\"na\\" ",row.names=FALSE,col.names=TRUE)
Appendix B: R Code of RQ2 for Thin Matching and Quintiles Thick Matching

###remove memory and set seed for generation
rm(list=ls())
set.seed(28760103)

###install packages
library(irtoys)
library(difR)
library(vcd)
library(CMC)

#initialize various variables
aI<-0
bI<-0
abl<-0
num_r<-0
reliab<-0
MH_typeI<-0

mh_pvalueI<-0
log_pvalueI<-0
log_pvalue_rawI<-0
wf_pvalueI<-0

sum_MH_pvalue<-0
power_MH_pvalue<-0
sum_mh_pvalueI<-0
power_mhI<-0
sum_wf_pvalueI<-0
power_wfI<-0
sum_log_pvalueI<-0
power_logI<-0
sum_log_pvalue_rawI<-0
power_log_rawI<-0

###set IRT parameters values
a1=1.6
a2=1.8
b1=-2.0
b2=2.0
c1=0.0
c2=0.0

###set replication, sample size and group mean difference
rep<-10000
samplen <- 1000
iterations<-samplen * 2                # iteration equals to double sample size
mean_difference<-0

###generate data
for(t in 1:rep)
{
    ###for Type I error rates

    DifparameterI<-cbind(runif(50,a1,a2),runif(50,b1,b2),runif(50,c1,c2))  ## 50 items for no DIF
    ##use IRTparameters to generate data responses
    DifresponseI<-sim(ip=DifparameterI,x=rnorm(samplen,mean=0,sd=1))
    DifresponseII<-sim(ip=DifparameterI,x=rnorm(samplen,mean=-1,sd=1))
    DifresponseIII<-rbind(DifresponseI,DifresponseII)

    ###calculate total scores for examinees
    rawscoreI<-rowSums(DifresponseI)
    rawscoreII<-rowSums(DifresponseII)
    rawscoreI<-as.matrix(rawscoreI)
    rawscoreII<-as.matrix(rawscoreII)
    rawscoreIII<-c(rawscoreI,rawscoreII)
    rawscoreIII<-as.matrix(rawscoreIII)

    ###build up dataset
    aI<-DifresponseI[,1]
    bI<-DifresponseII[,1]
    groupI<-rep(0:1,each=samplen)                  ##equals to sample size
    abI<-c(aI,bI)
    abI
    class(abI)
    dataIII<-cbind(groupI,DifresponseIII)
    m_groupI<-as.matrix(groupI)
    m_abI<-as.matrix(abI)
    dataI<-cbind(m_groupI,m_abI)
    dataI
    class(dataI)
    reliab<-alpha.cronbach(DifresponseIII)

    ##use thin matching in MH method
    MH_typeI<-
difMH(Data=dataIII[,2:51],group=dataIII[,1],focal.name=1,correct=TRUE,save.output = TRUE, output = c("MHresults","default"))
    #MH_typeI
MH\_ST<-as.matrix(MH\_typeI$MH)
#MH\_ST
as.numeric(MH\_ST[1,])
#class(as.numeric(MH\_ST[1,]))
MH\_stat[t]<-as.numeric(MH\_ST[1,])

###pull out p-value of MH
MH\_pvalue[t]<-pchisq(MH\_stat[t],df=1, lower.tail = F)

###calculate rejection number
if (MH\_pvalue[t]<0.05) {sum\_MH\_pvalue<-sum\_MH\_pvalue+1}

###use quintiles thick matching in MH method (rejection rate)
###build up scoregroup
quantile(rawscoreIII, probs=c(.2,.4,.6,.8,1))
quant <- quantile(rawscoreIII, probs=c(.2,.4,.6,.8,1))
quant <- as.matrix(quant)
quant
scoregroupI<-matrix(0,iterations,1)
for (p in 1:iterations)
{
  if (rawscoreIII[p,1]>=0&rawscoreIII[p,1]<=quant[1]) {scoregroupI[p,1]<-1}
  else if (rawscoreIII[p,1]>quant[1]&rawscoreIII[p,1]<=quant[2]) {scoregroupI[p,1]<-2}
  else if (rawscoreIII[p,1]>quant[2]&rawscoreIII[p,1]<=quant[3]) {scoregroupI[p,1]<-3}
  else if (rawscoreIII[p,1]>quant[3]&rawscoreIII[p,1]<=quant[4]) {scoregroupI[p,1]<-4}
  else if (rawscoreIII[p,1]>quant[4]&rawscoreIII[p,1]<=quant[5]) {scoregroupI[p,1]<-5}
  else {scoregroupI[p,1]<-99}
}
datasetI<-cbind(dataI,scoregroupI)
datasetI

###forming dataset for quintiles thick matching MH methods
###initialize cell variables
q1<-0
q2<-0
q3<-0
q4<-0
q5<-0
q6<-0
q7<-0
q8<-0
q9<-0
q10<-0
q11<-0
q12<-0
q13<-0
q14<-0
q15<-0
q16<-0
q17<-0
q18<-0
q19<-0
q20<-0

###count examinee number in each cell
for (r in 1:iterations)
{
    if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==1)  {q1<-q1+1}
    else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==1)  {q2<-q2+1}
    else if (datasetI[r,1]==1&datasetI[r,2]==0&datasetI[r,3]==1) {q4<-q4+1}
    else if (datasetI[r,1]==0&datasetI[r,2]==0&datasetI[r,3]==1) {q3<-q3+1}
    else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==2)  {q6<-q6+1}
    else if (datasetI[r,1]==1&datasetI[r,2]==0&datasetI[r,3]==2)  {q8<-q8+1}
    else if (datasetI[r,1]==0&datasetI[r,2]==0&datasetI[r,3]==2)  {q7<-q7+1}
    else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==3)  {q10<-q10+1}
    else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==3)  {q9<-q9+1}
    else if (datasetI[r,1]==1&datasetI[r,2]==0&datasetI[r,3]==3)  {q11<-q11+1}
    else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==4)  {q16<-q16+1}
    else if (datasetI[r,1]==1&datasetI[r,2]==0&datasetI[r,3]==4)  {q18<-q18+1}
    else if (datasetI[r,1]==0&datasetI[r,2]==1&datasetI[r,3]==5)  {q17<-q17+1}
    else if (datasetI[r,1]==1&datasetI[r,2]==0&datasetI[r,3]==5)  {q19<-q19+1}
    else if (datasetI[r,1]==1&datasetI[r,2]==1&datasetI[r,3]==5)  {q20<-q20+1}
}
if (q1+q2+q3+q4==0&q5+q6+q7+q8==0&q9+q10+q11+q12==0
    |q13+q14+q15+q16==0|q17+q18+q19+q20==0)
    {num_r<-num_r+1}
else
{
    datasetnew1<-array(c(q1,q2,q3,q4,
                          q5,q6,q7,q8,
                          q9,q10,q11,q12,
                          q13,q14,q15,q16,
                          q17,q18,q19,q20),

                          dim = c(2, 2, 5),
                          dimnames = list(
                              group = c("reference", "focal"),
                              Response = c("succeed", "failure"),
                              group.Level = c("1", "2", "3", "4", "5")))}
datasetnewI

###pull out p-value of MH
mantelhaen.test(datasetnewI)
mantelhaen.test(datasetnewI)$p.value
mh_pvalueI[t]<-mantelhaen.test(datasetnewI)$p.value
wf_pvalueI[t]<-woolf_test(datasetnewI)$p.value

###extract the odds ratio
mantelhaen.test(datasetnewI)$estimate
woolf_test(datasetnewI)

###use thin matching and quintiles thick matching in LR procedure
logdata1I<-data.frame(abI,groupI,scoregroupI)
logdata2I<-data.frame(abI,groupI,rawscoreIII)
logdatatrial1I<-glm(abI~groupI+scoregroupI,family=binomial(link=logit),data=logdata1I)
logdatatrial2I<-glm(abI~groupI+rawscoreIII,family=binomial(link=logit),data=logdata2I)

summary(logdatatrial1I)                                ##using scoregroup in LR
summary(logdatatrial1I)$coefficients
summary(logdatatrial1I)$coefficients[2,4]
summary(logdatatrial2I)                                ##using rawscore in LR
summary(logdatatrial2I)$coefficients
summary(logdatatrial2I)$coefficients[2,4]

###pull out p-value of LR
log_pvalueI[t]<-summary(logdatatrial1I)$coefficients[2,4]
log_pvalue_rawI[t]<-summary(logdatatrial2I)$coefficients[2,4]
summary(logdatatrial1I)$coefficients[2,4]
summary(logdatatrial1I)

###count rejection number
{
  if (mh_pvalueI[t]<0.05)            {sum_mh_pvalueI<-sum_mh_pvalueI+1
                                      mh_ratio[t]<-mantelhaen.test(datasetnewI)$estimate
                                      para_a[t]<-DifparameterI[1,1]
                                      para_b[t]<-DifparameterI[1,2]}
  if (mh_pvalueI[t]>0.05)            {para_a[t]<-NA
                                      para_b[t]<-NA
                                      mh_ratio[t]<-NA}
  if (log_pvalueI[t]<0.05)           {sum_log_pvalueI<-sum_log_pvalueI+1}
  if (log_pvalue_rawI[t]<0.05)       {sum_log_pvalue_rawI<-sum_log_pvalue_rawI+1}
  if (wf_pvalueI[t]<0.05)            {sum_wf_pvalueI<-sum_wf_pvalueI+1}
}
}## ending replication
###calculate the rejection rate (the Type I error rate) in thin matching and quintiles thick matching

\[
\text{power}_{\text{MH}}\text{pvalue} <- \frac{\sum \text{MH}_\text{pvalue}}{\text{rep}} \\
\text{power}_{\text{mhl}} <- \frac{\sum \text{mh}_\text{pvalueI}}{(\text{rep}-\text{num}_\text{r})} \\
\text{power}_{\text{wfl}} <- \frac{\sum \text{wf}_\text{pvalueI}}{(\text{rep}-\text{num}_\text{r})} \\
\text{power}_{\text{logI}} <- \frac{\sum \text{log}_\text{pvalueI}}{(\text{rep}-\text{num}_\text{r})} \\
\text{power}_{\text{log\_rawI}} <- \frac{\sum \text{log\_pvalue\_rawI}}{(\text{rep}-\text{num}_\text{r})}
\]

###show the results

\[
\text{power}_{\text{MH}}\text{pvalue} \\
\text{power}_{\text{mhl}} \\
\text{power}_{\text{wfl}} \\
\text{power}_{\text{logI}} \\
\text{power}_{\text{log\_rawI}}
\]

###save results

\[
\text{typeI} <- \text{data.frame}(a1,a2,\text{mean\_difference},\text{samplen},\text{power}_{\text{MH}}\text{pvalue},\text{power}_{\text{mhl}},\text{power}_{\text{wfl}}, \\
\text{power}_{\text{log\_rawI}},\text{power}_{\text{logI}})
\]

\[
\text{write.table} (\text{typeI},\text{file}="\text{result\_DIF/1000\_type\_I\_error\_5thick and thin\_1.3\_1.8.csv"},\text{quote}=\text{FALSE},\text{sep}="",\text{na}="",\text{row\_names}=\text{FALSE},\text{col\_names}=\text{TRUE})
\]