LORD'S WALD TEST FOR DETECTING DIF IN MULTIDIMENSIONAL IRT MODELS:

A COMPARISON OF TWO ESTIMATION APPROACHES

by

SOO YOUN LEE

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Education

Written under the direction of

Youngsuk Suh

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2015

ABSTRACT OF THE DISSERTATION

Lord's Wald Test for Detecting DIF in Multidimensional IRT Models:

A Comparison of Two Estimation Approaches

by SOO YOUN LEE

Dissertation Director:

Youngsuk Suh

Lord's Wald test for differential item functioning (DIF) has not been extensively studied particularly in the context of multidimensional IRT (MIRT) framework. Lord's Wald test was implemented using two estimation approaches in the MIRT framework: Marginal maximum likelihood (MML) estimation based on expectation maximization (EM) algorithm and the Bayesian Markov chain Monte Carlo (MCMC) estimation based on Metropolis-Hastings algorithm. This study investigated the recovery of item parameters, the Type I error, and the power of Lord's Wald tests obtained from the two estimation approaches under various simulation conditions, including DIF type differences, DIF magnitude differences, test length differences, and different combinations of sample sizes. Item responses were generated under multidimensional two-parameter logistic and three-parameter logistic models. Specific concerns for designing DIF detection conditions in MIRT framework were outlined based on the literature review on unidimensional and multidimensional DIF methods. The relative

performances of the two estimation methods compared and summarized under the simulation conditions considered in this study. Furthermore, English usage data were used to illustrate the use of Lord's Wald test with the two estimation approaches. Finally, the summary and implications of the results, the limitations of the present study, and directions for further studies were discussed.

ACKNOWLEDGEMENTS

I could not have finished this dissertation without the contributions of people who I am grateful. First and foremost, I would like to thank Dr. Youngsuk Suh for her consistent guidance and support for this study, and for my graduate school, GSE in general. Her endless expertise, insight, and patience to provide immediate and through feedback were invaluable. I greatly appreciate her lavish support, including mentorship, travel funding, and searching for a job. I am forever thankful to have had such mentoring and the opportunity to learn from one of the great minds in psychometrics.

Thank you to my committee members: Dr. Jimmy de la Torre, Dr. Chia-Yi Chiu, and Dr. Okan Bulut. I am indebted to them for their time and support, the thorough perspective each brought to our discussions, and their thoughtful comments.

Thank you to the beloved of my life, my friends: Katie, Muteb, Jinnie. Thank you to my friends in the program: Charlie, Nate, Wenchao, Eugene, Lokman, Ragip, Immanuel, and Sumbo. Especially, Mehmet, he encouraged me all the time, gave me brotherly care, and shared the pain through graduate school from the beginning. They kept me sane, indulged my complaints, and were there for me every step of the way. I not only would not have made it through graduate school without them, but I would not be who I am today. I am truly lucky to have such support in my life.

Finally, deep appreciation and love go to my family. My dad, Hyun-Jin Lee and my mom, In-Ok Cho, they have always believed in me. And they were with me at the best and worst moments of my Ph.D. journey. I could not have made it to this point without their generous support, love, and encouragement. My brother, Youngjae Lee, he

has always made sure of my financial well-being and endless support of love. I could never have hoped or expected to be so blessed. Thank you from the bottom of my heart.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In the educational and psychological measurement literature, the term *differential item functioning* (DIF) was created to define concerns about item bias within the context of test bias. DIF is present when persons from one group have different probabilities of answering an item correctly compared to persons from other groups after conditioning on the same ability level (Lord, 1977). Detecting DIF is an essential step in enhancing the validity of tests and can be a crucial step in establishing the fairness and validity of high-stakes tests that determine achievement, certification, and licensure. Over the years, DIF detection has been widely used in the context of the item response theory (IRT) model approach (e.g., Clauser & Mazor, 1998; Finch & French, 2007; Kim & Cohen, 1995; Oshima & Morris, 2008; Woods, Cai, & Wang, 2013) and compared to observed-score approaches (e.g., Finch, 2005; Finch & French, 2007). Observed-score approaches require fewer assumptions and are relatively easier of implement than IRT approaches. However, the results of the observed-score approaches might be sample-specific and therefore insufficient for ensuring measurement invariance (Budgell, Raju, & Quartetti, 1995; Hulin, Drasgow, & Parsons, 1983).

IRT approaches have been typically used within the unidimensional IRT (UIRT) framework. However, most educational and psychological tests are usually designed to measure subskills. These tests, such as the uniform certified public accountant (CPA) examination with four subtest areas (American Institute of CPAs®, 2015), the Graduate Record Examinations® (GRE®) general test (Educational Testing Service®, 2015), and the Standardized test (SAT®) Reasoning test (The College Board, 2015), consist of several

subsets of items that measure multiple domains. Consequently, the assumption of unidimensionality may not be true in practice (Ackerman, Gierl, & Walker, 2003; Snow & Oshima, 2009). In other words, when a test contains more than one target subscale or different groups of items measure distinctly different latent skills, unidimensional IRT models may not be applied correctly and can lead to erroneous results. Therefore, DIF in multidimensional tests must be examined with multidimensional DIF analyses (Snow & Oshima, 2009; Wang, Wilson, & Adams, 1997).

1.1 Differential Item Functioning

Since DIF analyses were first carried out in response to public concern that cognitive ability tests in 1960s discriminated against minority examinees (Angoff, 1993), over the past five decades, extensive psychometric research involving DIF has been conducted. The goal of DIF analyses is to detect irrelevant items on a test and then edit or eliminate them from the final test set (Angoff, 1993). In general, DIF is examined by comparing item responses for two groups of examinees, usually labeled the reference group and the focal group. In most applications, these groups represent types of examinees based on demographic characteristics such as gender or race (Finch & French, 2007). There are two characteristically different forms of DIF: uniform DIF and nonuniform DIF (Mellenberg, 1982). Uniform DIF occurs when item characteristic curves (ICCs) for two groups differ only in terms of the difficulty parameter. The relative advantage for one group is uniform in relation to the other group across the score scale. Nonuniform DIF exists when the ICCs for the two groups differ in the discrimination parameters and/or pseudo-guessing parameter with or without a difference in the

difficulty parameter (Clauser & Mazor, 1998). Therefore, the two ICCs can cross each other. In this study, both types of DIF are considered.

1.2 DIF Detection Methods

Various methods for detecting DIF have been introduced in the literature. These methods include the Mantel-Haenszel procedure (MH; Holland & Thayer, 1988; Mantel, 1963; Mantel & Haenszel, 1959; Penfield, 2001), logistic regression (Swaminathan & Rogers, 1990), proportion difference measures (Dorans & Kulick, 1983, 1986; Dorans & Schmitt, 1991), the simultaneous item bias test (SIBTEST; Bolt & Stout, 1996; Shealy & Stout, 1993), Lord's Wald test (Lord, 1980), area methods (Raju, 1988, 1990; Raju, van der Linden, & Fleer, 1995), the IRT likelihood ratio test (IRT-LR; Thissen, Steinberg, & Wainer, 1993), differential functioning of items and tests (DFIT; Raju et al., 1992, 1995), and simple area indices (Hambleton & Rogers, 1989; Rudner, 1977; Rudner, Getson, & Knight, 1980).

Among these DIF detection approaches, the IRT-LR test (also known as the χ^2 difference test), which always involves the comparison of two models, a compact model and an augmented model (Judd & McClelland, 1989), is more flexible than other DIF methods (Teresi, Kleinman, & Ocepek-Welikson, 2000; Thissen et al., 1993; Wainer, 1995). However, the IRT-LR DIF detection approach has several disadvantages: the painstaking model refittings and intensive computational time. As addressed by Millsap and Everson (1993), a set of unbiased anchor items must be accessible before the IRT-LR test is implemented. If the test is conducted with biased anchor items without prescreening, the results can be misleading. Furthermore, a dearth of reliable software for

performing the required computation is a disadvantage of the IRT-LR test since computing the likelihood requires multiple-group concurrent estimates.

Although Lord's Wald test for detecting DIF is asymptotically comparable to the IRT-LR test and computationally less intensive, the Wald test has not been considered a viable methodological algorithm for detecting DIF in almost a decade (Woods et al., 2013) because of two major shortcomings: severe Type I error inflation (Donoghue & Isham, 1998; Kim, Cohen, & Kim, 1994; Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987) and inaccuracy in estimating the covariance matrix (Donoghue & Isham, 1998; Kim et al., 1994; McLaughlin & Drasgow, 1987). However, Lord's Wald test was recently improved (Cai, 2012; Cai, Thissen, & du Toit, 2011; Langer, 2008) to overcome these shortcomings.

1.3 Motivation of the Study

To date, several studies have been reported on DIF detection using Lord's χ^2 test (Lord, 1977, 1980), also referred as the Wald test or Lord's Wald test in IRT approaches (Cohen & Kim, 1993; Kim & Cohen, 1995; Kim et al., 1994; Kim, Cohen, & Park, 1995; Langer, 2008; Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987; Woods et al., 2013; Yao & Li, 2010). In the past decade, particularly in the context of the multidimensional IRT (MIRT) framework, the performance of Lord's Wald (1943) test has not been extensively studied. Only a few studies have been published on the MIRT DIF detection method (Bolt & Johnson, 2009; Yao & Li, 2010). Bolt and Johnson (2009) used a multidimensional nominal response model (NRM; Bock, 1972). The aim of the study was not to explore dichotomously scored data for DIF detection but, utilizing ordered rating scale data from other factors that might be a potential cause of DIF, to investigate

response-style effects. Yao and Li (2010), however, used a Markov chain Monte Carlo (MCMC) algorithm to estimate item parameters for studying DIF in the MIRT context. Lord's Wald test was compared to Raju's volume test (a variant of Raju's area measures [Raju, 1988]). Yao and Li (2010) found that Raju's volume test had lower Type I error rates than Lord's Wald test, whereas Lord's Wald test had lower Type II error rates, and thus higher power than Raju's volume test. Yao and Li did not discuss how to correct the uncontrolled Type I error in the Wald test.

A newly developed (yet to be evaluated within the context of MIRT) Lord's Wald test, the improved Lord's Wald test (Cai, 2012; Cai et al., 2011), was evaluated for DIF detection in this study. Woods et al. (2013) evaluated the improved version of Lord's Wald test not under the MIRT framework but under the unidimensional IRT framework. The primary purpose of the DIF analysis proposed by Woods et al. (2013) was to investigate the strengths and weaknesses of the Wald-1 test (one-stage; Cai et al., 2011), which requires anchor items be indicated by the researcher, in relation to the Wald-2 test (two-stage; Langer, 2008), which does not require anchor items in the unidimensional IRT framework. At the same time, the improved Lord's Wald test (Wald-1 and Wald-2) was compared to the IRT-LR test, which requires concurrent calibration of the item parameters.

If tests are multidimensional, the multidimensional extension of the improved Lord's Wald test would be expected to be much better suited for studying DIF diagnostically than the improved Lord's Wald test under the unidimensional IRT framework. In addition, this study will shed light on the direct comparison of the marginal maximum likelihood (MML) and MCMC estimation methods regarding the

performance of Lord's Wald test in the MIRT framework. Notwithstanding the popularity of using MCMC to estimate parameters, the comparison of MCMC estimation using the Metropolis-Hastings algorithm (Bayesian approach) and MML estimation using an expectation maximization (EM or the supplementary EM [SEM; Cai, 2008; Meng & Rubin, 1991]) algorithm has never been implemented in terms of DIF detection especially in the MIRT framework. Previous results using MML estimation and marginal Bayes estimation (MBE; Mislevy, 1986) methods for two-parameter logistics (2PL) showed that both methods provided more accuracy and less inflation of Type I error rates than joint maximum likelihood (JML; as implemented in LOGIST [Wood, Wingersky, & Lord, 1976]) estimation for Lord's Wald test (Cohen & Kim, 1993; Lim & Drasgow, 1990). Kim et al. (1994) indicated that results for the applicability of Lord's Wald test for the three-parameter logistic (3PL) model were lacking in MML estimation and MBE (Millsap & Everson, 1993). To address this concern and different from previous studies (e.g., Cohen & Kim, 1993; Kim & Cohen, 1995; Kim et al., 1994; Langer, 2008; Lim & Drasgow, 1990; Woods et al., 2013; Yao & Li, 2010) of DIF detection methods, the unique contribution of the present study is the evaluation of the improved Lord's Wald test based on MML estimation for DIF detection and the comparison to Lord's Wald test based on the MCMC algorithm of the Bayesian approach under the MIRT framework. The advantages and disadvantages of using each estimation method was investigated. Exploring a more effective estimation method for detecting DIF under the various DIF conditions of multidimensional IRT-based Lord's Wald test is meaningful.

1.4 Objective of the Study

The current procedures for detecting DIF using Lord's Wald test fall short in several ways. The two most salient are that the latent trait or ability of interest is limited to the unidimensional context, and there is no direct comparison of item parameter estimation methods in the extant literature. This study delineates how comparable and reasonable the MML and MCMC approaches are under various DIF conditions in the context of the MIRT framework.

The primary purpose of this study is to evaluate the improved Lord's Wald test using a MML estimation approach for detecting DIF in an underlying multidimensional framework and to compare it to Lord's Wald test using the Bayesian MCMC estimation approach. In contrast to other previous DIF detection studies (e.g., Langer, 2008; Woods et al., 2013; Yao & Li, 2010) that focus on a single estimation method, this study used two estimation approaches for detecting DIF in an MIRT model.

For multidimensional estimation, a compensatory multidimensional two-parameter logistic (M-2PL) model and a multidimensional three-parameter logistic (M-3PL) model (both models are explained in the multidimensional IRT models section, Section 2.4) were used to estimate the item parameters for the Wald test. Four simulation factors for evaluating DIF were investigated, including (a) DIF types, (b) DIF magnitudes, (c) test lengths, and (d) sample sizes for the reference and focal groups. Despite the promising findings from previous MIRT DIF studies, research comparing two estimation methods for detecting DIF in a multidimensional framework remains somewhat limited in terms of various factors. That is also why the current study considered four specifically chosen factors to make direct comparisons with the previous studies. For instance,

numerous studies (e.g., Narayanan & Swaminathan, 1996; Oshima, Raju, & Flowers, 1997) have shown that larger samples and bigger magnitudes of DIF are easier to identify; however, in this study, the results were checked with various simulation conditions, which are implemented under the MIRT framework. Finally, many studies have investigated the effects of sample size on DIF detection. Researchers have consistently shown higher power in detecting DIF in the IRT approach with large samples (e.g., Rogers & Swaminathan, 1993; Swaminathan & Gifford, 1986; Swaminathan, Hambleton, Sireci, Xing, & Rizavi, 2003). A Monte Carlo simulation was conducted to examine three evaluation criteria: the recovery of model parameters, Type I error, and the power of two DIF detection estimation methods with the manipulated factors.

Relevant research questions on the performance of two estimation approaches for DIF detection in terms of the three evaluation criteria are as follows:

(a) How differently were the recoveries under the Bayesian MCMC and MML estimation methods affected by the manipulated factors?

(b) How differently did the Bayesian MCMC and MML estimation methods perform in terms of Type I error rates for Lord's Wald test under different simulation conditions?

(c) How differently did the Bayesian MCMC and MML estimation methods perform in terms of the power rates of Lord's Wald test under different simulation conditions?

(d) How differently did the M-2PL and M-3PL models perform in terms of the recovery of item parameters, Type I error rates, and power rates?

The present study investigated how each factor influenced the DIF detection results of the two estimation approaches under M-2PL and M-3PL.

CHAPTER 2

LITERATURE REVIEW

This chapter is organized as follows: The first section provides an overview of the theoretical background of unidimensional DIF models. The second section reviews the literature on Lord's Wald test in the unidimensional DIF application. The third section presents a thorough review of multidimensional DIF methods, with a comparison of model specifications, strengths, weaknesses, and potential problems in their use. The fourth section presents different multidimensional IRT models. The fifth section reviews Lord's Wald test in multidimensional IRT models. Finally, the last section explores three estimation methods, JML, MML, and MCMC, in terms of DIF studies.

2.1 Unidimensional DIF Methods

Methods for detecting DIF can be categorized within two major approaches: (1) observed-score approaches and (2) latent variable/item response theory (IRT) approaches (Millsap & Everson, 1993). The approaches share an assumption: The DIF items have been matched on the same dimension as the matching variable. That is, DIF implies that after controlling for the ability levels on the same dimension of interest, the effects of group difference on item responses still exist. The two approaches differ fundamentally in that the observed-score approaches use observed scores (i.e., total scores) as the matching variable, whereas the IRT approaches use ability parameter estimates as a function of observed data. This distinction determines how DIF is defined and measured.

There are two types of procedures if one needs to categorize DIF detection methods in terms of whether they use a mathematical model: (a) parametric and (b)

nonparametric procedures. Parametric procedures use a functional form (i.e., a mathematical model) for the specified relationship between the item score and the matching variable, whereas nonparametric procedures do not use a functional form. The dichotomous classification scheme is similar to the distinction between observed-score and IRT approaches, but it is not always clear-cut (Potenza & Dorans, 1995; Scheuneman & Bleistein, 1989; Wainer, 1993). The main advantage of parametric procedures is that parameter estimates resulting from parametric procedures have practical information and can provide insight into how an item or test could be revised. One problem in parametric procedures is the introduction of colinearity in standard errors (SEs), as clearly discussed in previous literature (Lord, 1980; Potenza & Dorans, 1995; Ramsay, 1991; Thissen & Wainer, 1982). That is, very high SEs of the parameter estimates can arise when an item has large sampling covariance among item parameter estimates. For instance, Thissen and Wainer (1982) showed that the error of estimation of the guessing parameter in the 3PL model was strongly related to the error of estimation of the difficulty parameter, and thus, the guessing parameter and the difficulty parameters were poorly estimated.

The major advantage of nonparametric procedures is that the procedures do not require item parameter estimation. Therefore, detecting DIF by comparing item or test scores can be done directly from examinee responses, thus avoiding difficulties during the item parameter estimation process (Raju & Ellis, 2002). Nonparametric procedures avoid parametric procedure-related problems such as matching variable and colinearity (Lopez, 2012). However, in contrast to parametric procedures, nonparametric procedures provide little information that may result in the potential causes of DIF, whereas parametric procedures provide a wealth of information (Lopez, 2012). Consequently,

parametric DIF detection procedures are often preferred; however, the information handled by parametric and nonparametric DIF detection procedures differs and remains clear (Drasgow & Hulin, 1990; Hulin et al., 1983; Lopez, 2012; Stark, Chernyshenko, & Drasgow, 2006). Two distinctions of matching variable and two classifications of DIF procedures can be crossed and applied in the context of dichotomous DIF methods as shown in Table 2.1 (Potenza & Dorans, 1995).

Table 2.1

Classification of Dichotomous DIF Procedures according to Parametric or Nonparametric

Type of procedure and matching variable	Parametric	Nonparametric
Observed Score	Logistic regression	MH STD
IRT	IRT-LR Raju's Area Measures Lord's Wald Latent class logistic regression	SIBTEST

2.1.1 Observed-score Approach. Observed-score approaches provide techniques as alternatives to IRT approaches. Several observed-score approaches for detecting DIF on dichotomously scored items have been introduced in the literature (Holland & Wainer, 1993; Scheuneman & Bleistein, 1989; Zumbo, Liu, Wu, Shear, Astivia, & Ark, 2015). Three widely used observed-score approaches are the MH procedure (Holland & Thayer, 1988; Mantel & Haenszel, 1959), standardization of difference measures (STD; Dorans & Kulick, 1983, 1986; Dorans & Schmitt, 1991), and logistic regression (Swaminathan &

Rogers, 1990). These three procedures are observed-score approaches because a common definition of null hypothesis DIF on the item level is shared and an observed score measure of the test construct of interest is used as a matching variable (Potenza & Dorans, 1995).

A major drawback of the observed-score approach is in the matching variable (Holland & Thayer, 1988; Swaminathan & Rogers, 1990; Zwick, 1990). In general, total scores are used as the matching variable to ensure comparability between the reference and focal groups. Millsap and Everson (1993) suggested that when many items are biased, the observed-score approaches tend to result in bias in the total score used for matching examinees. Van der Flier, Mellenbergh, Ader, and Wijn (1984) proposed an iterative approach that can improve DIF detection results. Removing biased items from the total score should alleviate the problem of bias in the matching variable. Another issue of the matching variable is that Type I error may be inflated when the observed score is not a sufficient/adequate measure of the underlying ability (Meredith & Millsap, 1992; Zwick, 1990).

Mantel-Haenszel (MH). Among the nonparametric DIF detection methods that have been widely introduced and investigated, the MH method is probably the most commonly used observed-score approach. The MH method uses a series of 2 x 2 contingency tables. Holland and Thayer (1988) used it for DIF detection and investigated group differences on a dichotomously scored test. The MH method is an extension of the χ^2 test and compares the performance of individuals between two groups (the reference and focal groups) after conditioning on a matching test score, which serves as an alternative for the latent trait measured by the instrument. The null hypothesis of the DIF

definition for the MH method is based on the assumption of a common odds ratio; the odds are compared for responding to an item correctly in the focal and reference groups. The null hypothesis is that the odds are independent of group membership for different score levels on the observed score (i.e., total score).

The MH method has two strong disadvantages in practice, including the matching variable problem (Holland & Thayer, 1988; Swaminathan & Rogers, 1993; Zwick, 1990) and sensitivity in detecting uniform or nonuniform bias (Holland & Thayer, 1988; Millsap & Everson, 1993). The first problem concerns the matching variable (i.e., total score) as a substitute for the latent trait. Several theoretical studies (Meredith & Millsap, 1992; Millsap & Meredith, 1992; Zwick, 1990) have shown that when complex IRT models were used to generate the item responses, the MH method tends to indicate DIF when there is no DIF present. This becomes more serious on short tests (e.g., fewer than 20 items) than on longer tests (Millsap & Everson, 1993). Additionally, IRT methods, particularly those that use MML estimation, have been shown to outperform the MH method for detecting DIF on short tests. IRT methods that use MML estimation can be applied to tests with as few as 10 to 20 items, whereas the MH method may not be appropriate for very short tests (Bock, 1993). Several studies have confirmed these results (Donoghue, Holland, & Thayer, 1993; Millsap & Everson, 1993); for example, when the total number of items is fewer than 20 on a test, the MH method has been shown to produce inflated Type I error rates.

Potenza and Dorans (1995) commented that the MH method is occasionally viewed as a parametric procedure because it measures the amount of DIF under the constant odds-ratio model as a particular type of violation of null DIF across all score

levels. Thus, the MH method is frequently referred to as a *uniform* DIF model. Rogers and Swaminathan (1993) compared the performance of logistic regression and MH methods for detecting DIF. The authors described that the MH method was not as powerful as the logistic regression method in detecting nonuniform DIF. The problem is more significant in practical settings. At times, applying different DIF techniques identifies different items as displaying DIF. As a result, the MH method is not typically designed for detecting nonuniform DIF (Clauser & Mazor, 1998).

The MH method has a major advantage in terms of the required sample size (Mazor, Clauser, & Hambleton, 1991; Spray, 1989). According to Clauser and Mazor (1998), this method has been shown to be effective when the sample was reasonably small (e.g., 200 for each group). The MH method is highly efficient in terms of statistical power and computational requirements (Clauser & Mazor, 1998). Thus, this method has been utilized extensively as an evaluation DIF detection method (Clauser & Mazor, 1998; Millsap & Everson, 1993; Thissen, 2001).

Standardization (STD). The STD for DIF detection method was introduced by Dorans and Kulick (1983, 1986). The MH and STD procedures are similar: The total score is used as the matching variable, and the 2 x 2 base data contingency table are utilized in MH and STD procedures to correctly interpret the proportions (Dorans, 1989; Dorans & Holland, 1992; Potenza & Dorans, 1995). According to Dorans (1989), however, the two procedures may differ meaningfully in terms of how they operate on the 2 x 2 contingency table to compare the performance of two groups of examinees. For example, the STD focuses on differences in proportion correct at each score level *k*,

whereas the MH uses the odds ratios to compare the base (reference) group to the focal group.

In contrast to the MH method, the STD method focuses on the difference percentage (proportion) correct at each score level *s,* D_s , which is calculated with the following:

$$D_s = P_{Fs} - P_{Rs}, \tag{1}$$

$$P_{Fs} = R_{Fs} / N_{Fs}; P_{Rs} = R_{Rs} / N_{Rs}, \tag{2}$$

where P_{Fs} is the proportions correct of the studied item for the focal group, and P_{Rs} is the percent correct of the studied item for the reference group at score level *s*, respectively. Referred to as the standardized *p* differences, the STD *p*-DIF is the average overall index of DIF obtained by Dorans and Holland (1993) and expressed as follows:

$$\text{STD } p\text{-DIF} = \sum_{s=1}^{S} K_s (P_{Fs} - P_{Rs}) / \sum_{s=1}^{S} K_s, \tag{3}$$

where $K_s / \sum K_s$ is the weighting factor at each score level *s*. The set of weights used for standardization depends on the purposes of the investigation. Some possible options are the following:

$K_s = N_{Ts}$, the number of people at *s* in the total group;

$K_s = N_{Rs}$, the number of people at *s* in the reference group;

$K_s = N_{Fs}$, the number of people at *s* in the focal group; or

$K_s =$ the relative number of people in some standard reference group.

Typically, $K_s = N_{Fs}$ has been used because it provides the greatest weight to differences in P_{Fs} and P_{Rs} at the score levels most frequently obtained by the focal group

of the study. The following are interpretation guidelines for STD *p*-DIF to evaluate the DIF effect size:

Type A items. $0 \leq |$ STD *p*-DIF$| \leq .05$: Items with negligible or nonsignificant DIF.

Type B items. $.05 \leq |$ STD *p*-DIF$| \leq .10$: Items with moderate DIF or significant DIF.

Type C items. $|$ STD *p*-DIF$| > .10$: Items with large DIF or significant DIF.

Dorans and Kulick (1986) observed that absolute values above .10 are unusual and should be inspected carefully.

Logistic Regression. Swaminathan and Rogers (1990) introduced parametric method for DIF detection based on logistic regression in which a parametric procedure uses an observed-score matching variable, as shown in Table 2.1. The logistic regression method for detecting DIF has become widespread since Swaminathan and Rogers (1990) applied it to dichotomous scored items (Monahan, McHorney, Stump, & Perkins, 2007). Several advantages of the logistic regression method are that it can be adapted for use with a wide variety of models such as extension to multiple examinee groups and polytomous item scores (Millsap & Everson, 1993). According to Swaminathan and Rogers's (1990) results, when logistic regression was compared to the MH method for detecting uniform and nonuniform DIF under various sample sizes and test lengths, logistic regression was equally powerful under most conditions of uniform DIF and was superior to the MH method under nonuniform DIF conditions. Additionally, the estimates of the regression coefficients produced by the logistic regression model may be very useful for locating DIF in the plot (Miller, Spray, & Wilson, 1992). A disadvantage is that the distributions of the test statistics are usually obscure and, thus, require

computationally intensive resampling methods to obtain critical values for hypothesis testing. In a simulation study, Swaminathan and Rogers (1990) found that the logistic regression test was three or four times more time-intensive than the MH statistic.

The logistic regression method can be estimated using the logit of endorsing the item and can be displayed as:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x + \beta_2 g + \beta_3 xg \,, \tag{4}$$

where P is the conditional proportion of examinees who endorse an item. β_0 is the intercept parameter of the model and β_1 is a parameter for the total score, x. β_2 is the group indicator as a dummy variable, g, for group (1 = reference, 0 = focal), and β_3 is a parameter for the interaction term, xg, between the total score and group membership. When β_2 is nonzero, the item has uniform DIF whereas, when β_3 is zero, the item has nonuniform DIF.

2.1.2 Item Response Theory (IRT) Approach. Although applications of observed-score approaches require few assumptions and are relatively easy to implement, the results may be sample-specific and, thus, inadequate for ensuring measurement invariance (Budgell, Raju, & Quartetti, 1995; Hulin et al., 1983). IRT approaches propose a latent trait or ability, usually denoted θ, that underlies the item responses and shares the use of the estimate of the latent trait as a matching variable rather than the observed score. These approaches often provide an advantage over observed-score approaches due to the ability to differentiate group mean differences on the latent trait (i.e., impact) from actual DIF. The limitation of IRT approaches was highlighted by Clauser and Mazor (1998), who specified that the data must meet the strong assumption of unidimensionality

of the models. Furthermore, a well-known drawback of IRT approaches indicated by Swaminathan and Rogers (1990) is that IRT approaches are very sensitive to sample size and model fit. According to Clauser and Mazor (1998), IRT approaches require large samples for accurately estimating the model parameters, especially when 2PL or 3PL models are used.

In practical applications that use the IRT approach, one of the most common models is the 3PL model, often used for responses to multiple-choice items in educational research, which assumes that the probability that an examinee with ability value θ will respond correctly to item *j* is

$$P(\mathrm{U}_{ij}=1\left|\theta_i, a_j, b_j, g_j\right.) = g_j + \frac{1-g_j}{1+e^{-1.7a_j(\theta_i-b_j)}}, \quad (5)$$

where $P(\mathrm{U}_{ij}=1\left|\theta_i, a_j, b_j, g_j\right.)$ is the probability of an examinee *i* with ability θ_i responding correctly, a_j is the item discrimination of the item, b_j is the difficulty of the item, and g_j is the lower asymptote, or pseudo-guessing parameter of the item. This discrimination and difficulty form of the 3PL model are used for computational ease; however, the literature often uses a intercept parameter, *d* rather than the difficulty parameter (*b*), and the relationship between *b* and *d* is *b = d/a*, if a negative intercept is modeled (i.e., $a*\theta-d$). When the pseudo-guessing value is set to zero, the 3PL model becomes the 2PL model. The item and ability parameters are estimated from the examinees' responses to a set of items. Among the many DIF detection methods in the context of unidimensional IRT models, four methods have been extensively studied and widely applied in the literature for evaluating DIF: Raju's (1988) area measures, the differential functioning of items and tests (DFIT; Raju et al., 1992, 1995), the likelihood

ratio test (IRT-LR, Thissen, Steinberg, & Gerrard, 1986; Thissen et al., 1988, 1993), and Lord's Wald test (Lord, 1980). Last, a new method, latent class logistic regression from Zumbo's third-generation DIF, will be discussed (Zumbo et al., 2015).

Raju's Area Measures. A wide variety of parametric DIF detection methods are available; however, among them, two main methods have been introduced in the context of IRT approaches. One is Lord's Wald test (which will be discussed later in this section), and the other is *Raju's area measures*. The area measures focus on comparing IRFs from the focal and reference groups of examinees by measuring areas between them over a selected interval on the θ scale (Raju, 1988, 1990; Raju et al., 1995). Several studies have indicated that two methods produce similar results, especially when sufficiently large samples are involved and the test length is appropriately long (Cohen & Kim, 1993; Kim & Cohen, 1995; Raju, Drasgow, & Slinde, 1993; Shepard et al., 1981; Shepard et al., 1984, 1985). Let $P_R(\theta)$ and $P_F(\theta)$ define the IRFs on the DIF item for the reference and focal groups, and the area measure is calculated as

$$A = f_s\left[P_R(\theta) - P_F(\theta)\right], \tag{6}$$

with θ located in the interval $S = (\theta_L, \theta_U)$ where *L* and *U* indicate the lower and upper bounds, respectively. There are various choices for selecting the function *f* and the interval boundaries: (1) absolute, unsigned, or signed differences, (2) bounded or unbounded in the interval *S*, (3) continuous integration or discrete approximation is utilized in *f*, and (4) the differences in *f* are equally weighted or differentially weighted. According to Raju (1988) and Camilli and Shepard (1994), the signed and unsigned areas can be defined as closed-form formulas:

$$\text{Signed Area (SA)} = SA_{kl} = \int_{-\infty}^{\infty} [P_R(\theta) - P_F(\theta)] d\theta, \quad (7)$$

$$\text{Unsigned Area (UA)} = UA_{kl} = \int_{-\infty}^{\infty} \left| P_R(\theta) - P_F(\theta) \right| d\theta. \quad (8)$$

In the SA and UA cases, the smaller the area, the lower the DIF values (Camilli & Shepard, 1994).

Several disadvantages of area measures exist. One problem is that their values depend on the endpoints of the selected interval: This choice is arbitrary to some extent. Choosing between the lower and upper of the selected interval (bounded and unbounded) area measures remains imprecise (Millsap & Everson, 1993). Thus, a disadvantage of the unbounded measures is that they are infinite when there are group differences in the guessing parameter in the 3PL model. Another disadvantage is that distributions of the test statistics in the area measure method are commonly unidentified, and thus, computationally intensive resampling is required to attain critical values for hypothesis testing (Lopez, 2012).

Differential Functioning of Items and Tests (DFIT). Raju et al. (1995) introduced a new method for detecting DFIT. This method is also an IRT-based approach and has the advantage of providing richer information, by using the differential test functioning (DTF) index in addition to the item DIF indices. Two types of DIF indices were developed in the DFIT framework: noncompensatory DIF (NCDIF) and compensatory DIF (CDIF) (Oshima & Morris, 2008). As a DIF index in the context of DFIT framework, NCDIF can be defined as the average squared distance between the item characteristic functions for the focal and reference groups. In the use of the dichotomous IRT model, the difference in item probabilities for item j (d_j) is explained

as the difference in the probability of a correct response on item *j* for a given ability level (θ) for examinee *s* between the focal and reference groups,

$$d_j(\theta_s) = P_{jF}(\theta_s) - P_{jR}(\theta_s), \tag{9}$$

$$NCDIF_j = E_F\left[d_j(\theta_s)^2\right], \tag{10}$$

where E_F represents the expectation postulated on the θ distribution from the focal group. In the notation for NCDIF, taking the square of the difference is critical, and thus, differences in opposite directions will not cancel DIF. NCDIF, therefore, detects uniform and nonuniform DIF. Raju et al. (1995) developed CDIF, which relates item- and test-level differential functioning in a very simple relationship. DTF can be calculated as two scores for each examinee based on ability level (θ): T_{sF} for the F group and T_{sR} for the R group. DTF can be calculated as:

$$DTF = E_F(\mathrm{T}_{sR} - \mathrm{T}_{sF})^2 = E_F\left[(\sum_{j=1}^{n} d_{js})^2\right] \tag{11}$$

CDIF can be expressed by considering the covariance, $Cov(d_j, D)$, and the mean, μ. $Cov(d_j, D)$ is the covariance between d_j and the difference between the two true scores (*D*). CDIF is defined as

$$CDIF_j = E_F(d_j D) = Cov(d_j, D) + \mu_{d_j}\mu_D, \tag{12}$$

According to Oshima and Morris (2008), a major advantage of the DFIT analysis is that it tests at the DIF and DTF levels. NCDIF is similar to DTF, except that the two total characteristic function curves are matched.

IRT Likelihood Ratio Test (IRT-LR). Thissen et al. (1988) mentioned that the IRT-LR test is preferable for theoretical reasons to other DIF detection under IRT

approaches such as Lord's Wald test and Raju's area measures in terms of computational brevity. These other methods require computing the variance and covariance matrices of item parameter estimates precisely from the second derivatives on the likelihood, which often impedes progress. In comparison, the IRT-LR test is not as computationally demanding (Kim & Cohen, 1995). Three main advantages of using the IRT-LR test can be summarized as follows: First, it does not require a linking process to transform item parameter estimates on common matrices across focal and reference groups because of the simultaneous item parameter estimation in each group when concurrent calibration is available. Second, the IRT-LR test, in comparison with Lord's Wald test, requires only the log likelihood values for two models being compared, which, in this case, is easier to compute (Thissen et al., 1988). Third, the IRT-LR test can be extended to polytomous data to evaluate DIF detection and is an effective method for detecting uniform and nonuniform DIF and DTF (Lopez, 2012).

The IRT-LR test can be obtained by comparing χ^2 values between two nested models: a compact model (*C*) and an augmented model (*A*). In this procedure, the null hypothesis, $H_0 : \mathrm{C}$ can be rejected, in favor of the alternative, using the likelihood ratio test. The test statistic is

$$G^2(df) = -2\ln\left[\frac{L(A)}{L(C)}\right]. \tag{13}$$

Where, $L(A)$ is the maximum likelihood for model *A* and $L(C)$ is the maximum likelihood for model *C*. The IRT-LR test statistic follows a large-sample χ^2 distribution with *df* equal to the difference in the number of parameters between the two models (Thissen et al., 1993).

Lord's Wald test. Lord's Wald test (1977, 1980) is generally used in the unidimensional IRT framework for detecting DIF by comparing vectors of IRT parameters between the focal group and the reference group. Initially, Lord proposed the evaluation of DIF for the location (difficulty) parameters only:

$$Z_j = \frac{\hat{b}_{F_j} - \hat{b}_{R_j}}{\sqrt{Var(\hat{b}_{F_j}) + Var(\hat{b}_{R_j})}}, \tag{14}$$

where $\hat{b}_{F_j}$ and $\hat{b}_{R_j}$ are the maximum likelihood estimates of the parameter b_j for each focal group and reference group, and $Var\left(\hat{b}_{F_j}\right)$ and $Var\left(\hat{b}_{R_j}\right)$ are the corresponding estimates of the sampling variance of $\hat{b}_{F_j}$ and $\hat{b}_{R_j}$, respectively. For example, to test a single parameter, *b*, the difference between the estimated *b*s across groups is compared to its SE;

$$SE(b_F - b_R) = \sqrt{Var(b_F) + Var(b_R)}. \tag{15}$$

Z^2 is a chi-square distributed with *df* = 1 for large samples. Lord (1980) extended this test to a generalized test of the joint difference between the vectors of discrimination and difficulty parameters across focal and reference groups. The test statistics, χ^2 is

$$\chi_j^{\,2} = \mathbf{v}_j' \sum^{-1} \mathbf{v}_j, \tag{16}$$

for the 2PL model, where $\mathbf{v}_j'$ is $[\hat{a}_{F_j} - \hat{a}_{R_j}, \hat{b}_{F_j} - \hat{b}_{R_j}]$, and let $\sum^{-1}$ define the corresponding variance-covariance matrix. For large samples, the $\chi_i^{\,2}$ test statistic follows a chi-square distribution with *df* = 2. In general, the *df* are the number of parameters per item *j*, which is being tested for DIF. The null hypothesis of no DIF to be tested for the 2PL is:

$$H_0 : a_{F_j} = a_{R_j}, b_{F_j} = b_{R_j} \ . \quad (17)$$

A common finding of the comparison between Lord's Wald test and other IRT approaches in many studies that use the 2PL or 3PL model (Langer, 2008; McCauley & Mendoza, 1985; Shepard, Camilli, & Averill, 1981; Shepard, Camilli, & Williams, 1984, 1985; Thissen et al., 1988; Woods et al., 2013; Yao & Li, 2010) is that the performance of the Wald statistic is closely related to that of unsigned area indexes proposed by Raju (1988) and Camilli and Shepard (1994). Lord (1980) advised that using the 3PL model and MML estimation methods with prior distributions on the guessing parameter, *g*, may allow wider use of the chi-square test.

Lord's Wald test has been criticized because the null hypothesis may be rejected even when there is a small difference in the area between two IRFs (Millsap & Everson, 1993). However, Lord's Wald test and Raju's area measures (unbounded area measures) offer the benefit of mathematical tractability, which use SEs and a formal hypothesis test (Millsap & Everson, 1993). Millsap and Everson (1993) explained that the best approach may be achieved by adding the calculation of a bounded area measure to the Wald statistic when the chi-square test is significant.

Using the improved version of the Wald test (Cai, 2012; Cai et al., 2011; Woods et al., 2013), the covariance matrix can be estimated accurately, and when the constant latent scale has been held, the concurrent calibration of item parameter estimation for the focal group is allowed. The improved version of Lord's Wald test can calibrate all parameters concurrently, whereas Lord's original test used separate calibration for item parameters (Woods et al., 2013). The improved version of the Wald test that requires anchor items by user is referred to as *the Wald-1 test*, and the improved version of the

Wald test that does not require anchor items is referred to as *the Wald-2 test* to avoid labeling confusion. In this study, only the Wald-1 test was used to detect DIF using the MML estimation method.

Latent Class Logistic Regression. Recently, a novel methodology of item response, latent class logistic regression, was introduced (Zumbo et al., 2015). Several essential differences should be emphasized in contrasting the MH and/or logistic regression, previous IRT-based models, and multidimensional models under DIF frameworks (Zumbo, 2007a). Zumbo et al. (2015) noted that compared to the first framework (MH and/or logistic regression) and the second framework (IRT-based models), the latent class logistic regression model, as highlighted in *third-generation DIF methodology*, focuses on latent variable mixture models, whereas the other two frameworks focused on manifest grouping variables (e.g., gender, ethnicity, or language of the test) are used for flagging potentially problematic items (Zumbo, 2007a). In the latent class logistic regression, the question of whether important grouping variables as potential causes of DIF were not easily directly observed is addressed (Zumbo et al., 2015). As an extension of widely used logistic regression DIF methods, but unlike the traditional logistic regression model that has parameters that account only for relationships between observed variables, the latent class logistic regression model can include one or more discrete latent variables. As an extension of the IRT DIF methods, but unlike the conventional IRT methods that were indicated by the continuous latent variable, the latent class logistic regression model allows latent classes to be indicated by discrete latent variables (Zumbo et al., 2015). Thus, binary, ordinal, and nominal logistic

regression and therefore any of these item response types, or combination in testing can be carried out with the latent class logistic regression model (Zumbo et al., 2015).

2.2 Previous Research on Lord's Wald Test in Unidimensional IRT Models

DIF can be an indicator of irrelevant variance that can affect test scores, and thus, gathering evidence of test score validity is an essential step (Finch & French, 2007). Various studies have evaluated and improved the accuracy of detecting DIF methods using several DIF methods in the context of the unidimensional framework. However, the present study is concerned only with Lord's Wald test using the MML and MCMC estimation methods.

McLaughlin and Drasgow (1987) evaluated the performance of the Wald test using two sample sizes (N = 250 and 1,000) under 3PL and 2PL. For both models with the two sample sizes, the Type I error rates when the person ability was known were below the expected nominal alpha levels of .0005, .001, .005, .01, .05, and .10. According to McLaughlin and Drasgow (1987), the Type I error rates tended to be inflated when both item and person parameters were estimated at the same time.

Lim and Drasgow (1990) evaluated the effectiveness of two estimation methods (i.e., MML and Bayes modal estimation) compared to the JLM estimation method using the Wald test with unidimensional and multidimensional data under the 2PL. Unidimensional data were generated with 2PL, and the multidimensional data were generated using Schmid and Leiman's (1957) hierarchical factor analysis model (Lim & Drasgow, 1990). Additional description and technical details of the Schmid-Leiman model for generating multidimensional data were given in the Drasgow and Parsons (1983) and Lim and Drasgow (1990) studies. The Lord's Wald test statistic was used a

chi-square distribution with 2 *df* for unidimensional and multidimensional datasets. Results from the study indicated that the MML and Bayes modal estimation methods provided parameter estimates that were more accurate and less inflated Type I error rates than JLM. In a large sample (N = 750), the MML and Bayes modal estimation methods had very similar SEs, except for items with large discrimination and extreme difficulty, in which case the SEs of MML were larger than those of Bayes modal estimation method. One interesting finding was that these patterns were observed regardless of the dimensionality of the data. In general, the MML and Bayes modal estimation methods produced similar results: higher power rates of flagging DIF with larger samples. Again, however, the dimensionality of the data had only a negligible influence on the power rates.

Cohen and Kim (1993) examined the effectiveness of two Raju's area measures (the *Z* test for exact signed area and the *Z* test for exact unsigned area), the IRT-LR test, and Lord's Wald test for different test length, sample size, proportion of DIF items, and item parameter estimation conditions using 2PL. Item and person ability parameters were estimated using MML estimation and MBE. They reported that the Type I error rates in DIF detection tended to be lower for Lord's Wald test and Raju's area measures than those for the IRT-LR test.

Kim et al. (1994) proposed the use of Lord's Wald test to evaluate 2PL, 3PL, and 3PL with a fixed guessing parameter using MML and MBE methods via a simulation study and a real data analysis. As Kim et al. (1994) noted, the Type I error rates for 3PL consistently exceeded the expected nominal alpha values of 0.05 and 0.10. For the 3PL model with a fixed guessing parameter and for the 2PL model, the Type I error rates were

consistently below the expected nominal alpha value. Results supported that Type I error rates for the 3PL with a fixed guessing parameter and 2PL models were within the expected nominal alpha values with a larger sample (N = 1,000). Although these results were more apparent in MML than in MBE, the differences between MML estimation and MBE were negligible.

Kim and Cohen (1995) analyzed three commonly used IRT procedures for detecting DIF using 2PL under unidimensional IRT: Lord's Wald test, Raju's two area measures (signed and unsigned), and the IRT-LR test in terms of practical considerations such as linking metrics and scale purification. An iterative procedure was described for the IRT-LR test G_j^2 to purify the anchor items. The data were also examined using MH χ_j^2 and Pearson's K_j^2. Results of the comparisons suggested agreement among the iterative procedures and noniterative procedures. The agreement with iterative procedures such as Lord's Wald test, Raju's area measures, and the IRT-LR test on flagging DIF items was slightly higher for the final iteration results than for the first iteration. For noniterative procedures such as MH χ_j^2 and Pearson's K_j^2, the agreement was generally higher for the first iteration results. Using Lord's Wald and Raju's area measures, the Type I error rates for both DIF methods tended to be within the expected nominal alpha value 0.05.

Kim et al. (1995) presented a DIF detection method for the multiple-group condition. The multiple pairwise comparisons statistic for DIF detection used in this study is closely related to Lord's Wald test and can be used to estimate item parameter in two groups. The two important aspects of this study are as follows: (a) a method for DIF detection in multiple groups was presented; (b) the method from (a) is the generalized

Lord's Wald test statistic introduced in this paper. Typically, a DIF study is conducted in two groups, but a practical situation may arise among several groups. The extension of the error variance-covariance matrix used in the generalized Lord's Wald test to the M-2PL and M-3PL models (which will be illustrated in Lord's Wald test in the multidimensional IRT models section later) was introduced. Three groups consisting of the reference group and two focal groups were used. Two hundred examinees for each group (for the reference and the two focal groups) were selected from real data and used with the 2PL, MML estimation. Kim et al. (1995) reported that differences between Lord's Wald test and the pairwise statistic occurred because of differences in the linking coefficients.

Recently, Lord's Wald test (Wald-2) was enhanced and reexamined by Langer (2008). The Wald-2 test performed well in terms of power and Type I error rates using the 3PL model. Item parameters were estimated using MML estimation. All three parameters (*a*, *b*, and *g*) have been investigated. As a result, the simulation condition of a large sample (e.g., 1,000) per group for a long test length (e.g., 40 items) showed the Type I errors were significantly less than or close to the nominal value (i.e., 0.05). Langer (2008) found that these results support the hypothesis that the prior on the *g* parameter induces close to zero Type I error rates for the guessing parameter in DIF detection, leading the Wald test to be conservative. The results also showed that after the prior on the *g* is removed, the alpha rates appeared to improve and close to the nominal value as the sample size increased. Moreover, the results confirmed that shifting the *b* parameter in detecting DIF has a greater effect in terms of power rates than reducing or increasing the *a* parameter in DIF detection (Langer, 2008).

Woods et al. (2013) examined the performance of the Wald test in DIF detection for multiple groups (e.g., two or three groups) under the unidimensional IRT framework using the 2PL model and Samejima's (1997) graded response model. Equal sample sizes (1,000 and 500) for all groups and unequal sample sizes (1,500/500 and 750/250) for two groups and (1,500/500/500 and 750/250/250) three groups were used with the 24 five-category ordinal item test. Woods et al.'s (2013) results showed that the Type I error rates for Wald-1 were lower than those of Wald-2. Wald-1 is recommended because of the controlled Type I error rates and greater power rates regardless of the sample sizes. When the Wald-1 test was compared to the IRT-LR test, the Wald-1 was preferred because the Type I error rates of Wald-1 were within the expected nominal alpha values (0.05) for all groups. The power of the IRT-LR test was similar to that of Wald-1; however, Wald-1 provided higher power in the unequal sample sizes. The Wald-1 test showed equal to or greater power in detecting DIF compared to the IRT-LR test.

2.3 Multidimensional DIF Methods

Although development in IRT has promoted predominantly unidimensional IRT DIF studies as a method for efficiently detecting DIF, the unidimensional IRT DIF studies provide insufficient information for utilizing DIF detection in the context of multidimensional IRT. Several MIRT DIF detection methods evolved from unidimensional IRT DIF detection methods. DIF detection methods in the context of the multidimensional framework can be either parametric or nonparametric approaches in the same manner as unidimensional DIF detection methods. Stout, Li, Nandakumar, and Bolt developed multidimensional SIBTEST (MULTISIB; 1997) as a nonparametric approach, which was an extension of the unidimensional SIBTEST methodology to two-

dimensional data. Although MULTISIB has been suggested as a very effective methodology for identifying DIF items, it is limited because two primary dimensions seem to be the maximum that can be applied to the data.

Only a handful of studies have detected DIF with the multidimensional IRT framework. Oshima et al. (1997) extended the DFIT method from Raju et al. (1995) to the multidimensional DIF analysis using dichotomous data. Their DIF analysis was designed for multidimensional data structures, and the multidimensional DFIT could identify DIF correctly only after a linking process (Snow & Oshima, 2009). That is, multidimensional linking was required to adjust the location differences as well as the variance and covariance differences in ability dimensions for the groups being compared (Oshima et al., 1997; Suh & Cho, 2014). During this linking process, errors related to the linking process were likely to emerge in detecting DIF (Shepard et al., 1984). Using M-2PL, Oshima and her colleagues (1997) simulated two-dimensional data with known DTF and DIF. After the appropriate linking, the results showed DIF was identified correctly in various conditions, including when the distributions of θ differed for the reference and focal groups.

The IRT-LR test (also known as a chi-square difference test) benefited from avoiding a linking process because the item parameter estimation can be calibrated simultaneously. The IRT-LR test of DIF and global DIF (GDIF) detection at the test level in the context of MIRT, as an extension of unidimensional IRT, was investigated extensively by Suh and Cho (2014). The simulation study examined the performance of the LR tests obtained from limited information estimation methods (i.e., robust weighted least square estimators; RWLS) implemented in Mplus 6 (Muthén & Muthén, 2010). To

estimate the MIRT parameter, two RWLS estimators were used: (1) weighted least square with adjusted means and variance [WLSMV] and (2) weighted least square with adjusted means [WLSM]. Two sample sizes (N = 500 and 1,000) were used with the 40-item test under three DIF conditions, (1) non-DIF, (2) uniform DIF, and (3) nonuniform DIF, and under three GDIF conditions, (1) non-GDIF, (2) unidirectional GDIF, and (3) balanced-directional GDIF conditions. These DIF and GDIF conditions were used to investigate the Type I error rates and rejection rates of the chi-square difference tests. The results from Suh and Cho's (2014) study showed that for the chi-square tests for detecting GDIF, WLSM tended to produce inflated Type I error rates for small-sample conditions, while WLSMV appeared to yield lower error rates than the expected value on average. In addition, WLSM produced higher rejection rates than WLSMV. For the χ^2 tests for detecting DIF, WLSMV tended to yield somewhat higher rejection rates than WLSM. The error rates for both estimators were close to the expected value on average. The reasonable conclusion was found in the results that the power rates increased as the sample size increased, consistent with previous research (e.g., Kim & Yoon, 2001; Suh & Bolt, 2011).

Yao and Li (2010) proposed a DIF detection procedure to identify only items that have adverse DIF[1] (Douglas, Roussos, & Stout, 1996) using real and simulated data in the MIRT framework. In that regard, Lord's Wald test was compared with Raju's volume test in the context of a two-dimensional IRT DIF study. Using their procedure, benign

[1] Adverse DIF (Bolt & Stout, 1996; Douglas et al., 1996) occurs when the nuisance dimension (which is clearly extraneous to the construct intended to measure) diminishes test fairness and can be eliminated by removing or by revising the item.

DIF[2] (Douglas et al., 1996) items are not supposed to be detected. Dichotomously scored items and polytomously scored items were analyzed with the M-3PL model and a multidimensional version of the partial credit model, respectively, with 3,000 examinees for each group. Six DIF items were included separately in 16-, 26-, and 36-item tests. Six DIF conditions were considered: (1) non-DIF, (2) benign DIF only, (3) adverse uniform DIF only (UNIF), (4) adverse nonuniform DIF only (NONUNIF), (5) benign DIF with adverse uniform DIF (UNIFBOTH), and (6) benign DIF with adverse nonuniform DIF (NONUNIFBOTH). The Type I error rates of Lord's Wald test were zero for all conditions. For Lord's Wald test and Raju's volume test, UNIFBOTH and NONUNIFBOTH showed higher Type II error rates than UNIF and NONUNIF. For Lord's Wald test, in general, the Type II error rates were not influenced by the test length and the correlation between the two dimensions for the UNIF and NONUNIF conditions. However, Type II error rates tended to decrease as the test length increased, and the correlation increased for UNIFBOTH and NONUNIFBOTH conditions. In summary, they found that Raju's volume test was preferred although Lord's Wald test and Raju's volume test provided comparable results. However, the problem associated with the volume test is that the uncertainty of the cutoff values for Raju's volume test, which makes the volume test less convenient. The cutoff values rely on the discrimination and intercept parameter values, ability distributions, and null condition, and thus vary.

2.4 Multidimensional IRT Models

The primary assumption of the IRT model is that the underlying latent ability measured is unidimensional. However, when this assumption is not met, the estimation of

[2] Benign DIF (Bolt & Stout, 1996; Douglas et al., 1996) is the DIF caused by the auxiliary (secondary) dimension. Benign DIF occurs when the secondary dimension does not contribute to the test unfairness.

item parameters and examinee abilities may be detrimentally affected (Ansley & Forsyth, 1985; Camilli, Wang, & Fesq, 1995; Finch, 2010, 2011; Reckase, 1985; Reckase, Carlson, Ackerman, & Spray, 1986). Moreover, most educational and psychological tests in practical settings are somewhat multidimensional (Bolt & Lall, 2003; Zhang & Stone, 2008). MIRT models reflect the interaction between examinees and test items more thoroughly than UIRT models (Zhang & Stone, 2008); MIRT models provide subscale scores and diagnostic value for the examinees (Wainer, Vera, Camacho, Reeve, Nelson, et al., 2001). When a more informative test is required, the test is more likely to be multidimensional (Yao & Li, 2010).

When a multidimensional model was chosen to analyze multidimensional data in which items measure multiple latent abilities, a clear distinction was frequently drawn between two types of multidimensional models: *compensatory* and *noncompensatory* MIRT models (Ackerman, 1989; Embretson & Reise, 2000; Way, Ansley, & Forsyth, 1988). In compensatory MIRT models, the latent abilities interact, and thus, a deficiency in one ability can be compensated for by other abilities. In the noncompensatory MIRT model, in contrast, sufficient levels of each measured latent ability are necessary, and a deficiency in one ability cannot be offset through an increase in others (Bolt & Lall, 2003). Compensatory MIRT models often assume that the probability of responding to an item correctly is influenced by a weighted linear combination of latent abilities (McDonald, 1997; Reckase, 1997b). In addition, estimation software is available for compensatory MIRT models in exploratory and confirmatory applications. In contrast, estimating noncompensatory MIRT models is challenging in exploratory situations. The practitioner may select a MIRT model for a multidimensional test based on prior

knowledge about the item response, so that the description of the ability interaction best suits the data.

When a multidimensional test is evaluated, two multidimensional tests are categorized by measuring unintended-to-be-defined or intended-to-be-defined latent traits: the multidimensional between-item test and the multidimensional within-item test (Oshima et al., 1997; Wang et al., 1995, 1997). According to Wang et al. (1995,1997), the multidimensional between-item test is a test such as high-stakes licensure tests that measure multiple subsets of skills (latent traits), whereas the multidimensional within-item test contains items that require compound multiple skills (latent traits), such as test items for two-dimensional data to measure two intended-to-be-defined skills (latent traits), θ_1 and θ_2 throughout the test (e.g., Oshima et al., 1997; Snow & Oshima, 2009).

Although several compensatory and noncompensatory models exist, this study utilizes two popular IRT models: the compensatory multidimensional extensions of the two-parameter logistic (M-2PL; Embretson & Reise, 2000) model and the compensatory M-3PL (Reckase, 1997) model with the multidimensional within-item test. The M-2PL model is given by

$$P(U_{ij}=1|\boldsymbol{\theta}_i,\mathbf{a}_j,d_j)=\frac{e^{\mathbf{a}_j\boldsymbol{\theta}_i'-d_j}}{1+e^{\mathbf{a}_j\boldsymbol{\theta}_i'-d_j}}, \tag{18}$$

where $P(U_{ij}=1|\boldsymbol{\theta}_i,\mathbf{a}_j,d_j)$ is the probability of an examinee *i* answering item *j* correctly with ability $\boldsymbol{\theta}_i$. In Equation 18, $\mathbf{a}_j\boldsymbol{\theta}_i'-d_j=a_{j1}\theta_{i1}+a_{j2}\theta_{i2}+\ldots+a_{jm}\theta_{im}-d_j=\sum_{l=1}^{m}a_{jl}\theta_{il}-d_j$ can be the expansion of the exponent, *e*. The expression of the exponent denotes a linear

function of $\boldsymbol{\theta}$ with the intercept parameter, the *d* parameter, and the elements of slope parameters, the **a** vector.

In the real test situation, it is natural for examinees to guess the response item correctly especially on objective tests such as multiple-choice items. When guessing is present and following the notation of the M-2PL model in Equation 18, the M-3PL model can be expressed as:

$$P(U_{ij}=1|\boldsymbol{\theta}_i,\mathbf{a}_j,d_j,g_j)=g_j+(1-g_j)\frac{e^{\mathbf{a}_j\boldsymbol{\theta}_i'-d_j}}{1+e^{\mathbf{a}_j\boldsymbol{\theta}_i'-d_j}}, \tag{19}$$

where g_j is the pseudo-guessing parameter for item *j*.

2.5 Lord's Wald Test in Multidimensional IRT Models

Lord's Wald test (Lord, 1977, 1980) has been widely used for unidimensional IRT-based methods for detecting DIF in practice. The unidimensional IRT-based methods compare item parameter estimates from different groups. A generalized Lord's Wald test statistic was introduced by Kim et al. (1995), and using discrimination and intercept parameters for the M-2PL model, the statistic is adjusted as follows:

$$\chi^2=\hat{\mathbf{v}}'\sum{}^{-1}\hat{\mathbf{v}}, \tag{20}$$

where $\hat{\mathbf{v}}=[\hat{a}_{f1}-\hat{a}_{r1},\hat{a}_{f2}-\hat{a}_{r2},\hat{d}_f-\hat{d}_r]$ represents the vector differences for all item parameters between the reference and focal groups, and $\sum^{-1}$ is the error variance-covariance matrix.

$$\sum{}^{-1}=(\sum\nolimits_F+\sum\nolimits_R)^{-1} \tag{21}$$

where $\sum_F$ is the focal group's error variance-covariance matrix and $\sum_R$ is the reference group's error variance-covariance matrix. For example, the error variance-covariance matrix for the focal group under M-2PL is written below:

$$\sum_F = \begin{pmatrix} \text{var}(a_{f1}) & \text{cov}(a_{f1}, a_{f2}) & \text{cov}(a_{f1}, d_f) \\ \text{cov}(a_{f2}, a_{f1}) & \text{var}(a_{f2}) & \text{cov}(a_{f2}, d_f) \\ \text{cov}(d_f, a_{f1}) & \text{cov}(d_f, a_{f2}) & \text{var}(d_f) \end{pmatrix}_{3\times 3} . \quad (22)$$

The χ^2 statistic follows a χ^2 distribution with M +1 degrees of freedom (i.e., df = 3) for the null hypothesis of no DIF for M-2PL:

$$H_0 : a_{f1} = a_{r1}; a_{f2} = a_{r2}; d_f = d_r. \quad (23)$$

The generalized Lord's Wald test using discrimination, intercept, and guessing parameters for the M-3PL model can be expressed as:

$$\chi^2 = \hat{\mathbf{v}}' \sum^{-1} \hat{\mathbf{v}}, \quad (24)$$

where $\hat{\mathbf{v}} = [\hat{a}_{f1} - \hat{a}_{r1}, \hat{a}_{f2} - \hat{a}_{r2}, \hat{d}_f - \hat{d}_r, \hat{g}_f - \hat{g}_r]$ is the vector of differences for all item parameters between the focal and reference groups. The notation for the error variance-covariance matrix in Equation 21 are the same as for M-2PL. However, the error variance-covariance matrix is different from M-2PL, for example, for the focal group,

$$\sum_F = \begin{pmatrix} \text{var}(a_{f1}) & \text{cov}(a_{f1}, a_{f2}) & \text{cov}(a_{f1}, d_f) & \text{cov}(a_{f1}, g_f) \\ \text{cov}(a_{f2}, a_{f1}) & \text{var}(a_{f2}) & \text{cov}(a_{f2}, d_f) & \text{cov}(a_{f2}, g_f) \\ \text{cov}(d_f, a_{f1}) & ... & \text{var}(d_f) & ... \\ ... & ... & \text{cov}(g_f, d_f) & \text{var}(g_f) \end{pmatrix}_{4\times 4} . \quad (25)$$

The χ^2 statistic follows a χ^2 distribution with M +2 degrees of freedom (i.e., df = 4) under the null hypothesis of no DIF for M-3PL:

$$H_0 : a_{f1} = a_{r1}; a_{f2} = a_{r2}; d_f = d_r; g_f = g_r. \quad (26)$$

2.6 Estimation Methods

In the current study, the Wald-1 test by Cai et al. (2011) was investigated using flexMIRT (Cai, 2012) for detecting DIF in the MIRT framework and compared with Lord's Wald test using BMIRT (Yao, 2003, 2010) via a Monte Carlo study. The Wald-1 test from flexMIRT is carried out with the SEM algorithm under the MML estimation approach. In contrast, Lord's Wald test from BMIRT is obtained from the MCMC (Gamerman, 1997) algorithm under the Bayesian approach. McLaughlin and Drasgow (1987) found that JML estimation resulted in inflated Type I error rates for item and ability parameters. Yao and Boughton (2005b, 2007) reported the MCMC estimation method was empirically comparable to and better than the ML estimation method in estimating multidimensional item and ability parameters, and found that the performance of the MCMC parameter estimation was better than that of the ML parameter estimation. Following are brief summaries of three estimation methods, JML, MML, and MCMC, in terms of DIF studies.

2.6.1 Joint Maximum Likelihood Estimation. In Lord's (1977, 1980) work on the Wald test for evaluating DIF, SE estimates obtained with the joint maximum likelihood implemented in LOGIST (Wood et al., 1976) were calculated with θ as a fixed latent variable. According to McLaughlin and Drasgow (1987), SE estimates from JML are not precise in the modern sense of the θ concept as a latent random variable. In a simulation study, McLaughlin and Drasgow (1987) found that inaccurate SE could lead to seriously inflated Type I error rates at least 10 times higher than the nominal α level. When the performance of the MML item parameter estimation was compared to that of JML, Drasgow (1989) observed the MML estimation was superior. Lim and Drasgow

(1990) noted that although many IRT approaches in DIF studies have used JML, it can lead to incorrect and misleading results. For example, the average biases in item parameters for JML estimation were much larger than when the MML estimation method was used. Thus, utilizing the MML estimation method has become the accepted method for implementing Lord's Wald test for detecting DIF (Drasgow, 1989; Langer, 2008).

2.6.2 Marginal Maximum Likelihood Estimation. According to McLaughlin and Drasgow (1987), the MML estimation method is distinguished from JML estimation method because the item parameters are not estimated simultaneously with ability parameters. The Wald-1 test is expected to improve on Lord's (1980) original test because the covariance matrix is estimated using the SEM algorithm (Langer, 2008). This is mainly because the SEM algorithm is designed as a strategy for calculating the information matrix used for estimating the SEs of the item parameter estimates, whereas an EM algorithm is used for estimating parameters (Woods et al., 2013). Consequently, undesirable SE estimates, which are the fundamental problem with Lord's original test for detecting DIF, can be circumvented by utilizing the SEM algorithm. Calculating SEs is not straightforward with EM algorithms because the full parameter information matrix is not a by-product of the estimation as it is with non-EM maximum likelihood estimation (Cai, 2008). Based on the SEM algorithm, Wald-1 and Wald-2 link the metric across groups simultaneously with item parameter estimation and DIF testing and should therefore improve on ad hoc linking. Wald-1 and Wald-2 can be implemented in flexMIRT and use SEM estimation for the covariance matrix.

2.6.3 Bayesian Markov Chain Monte Carlo Estimation. In the MIRT context, parameter estimation can be challenging due to the number of additional parameters, compared to the unidimensional case. The item and person parameters are estimated using the Metropolis-Hasting algorithm that samples from the joint posterior probability. The Metropolis-Hastings algorithm is a general term for Markov chain simulation methods and has been useful for selecting samples from appropriate distributions. When the Markov chain is applied in IRT, point estimates of the model parameters are often selected as the means of the marginal posterior distributions (Bolt & Lall, 2003).

MCMC methods provide much potential for estimating complex statistical models (Gilks, Richardson, & Spiegelhalter, 1996) and have received increasing attention in IRT (Albert, 1992; Baker, 1998; Béguin & Glas, 2001; Bolt & Lall, 2003; Kim, 2001; Pats & Junker, 1999a, 1999b; Wollack, Bolt, Cohen, & Lee, 2002). Glas and Meijer (2003) remarked that Bayesian estimates are not better than those produced by the MML estimation, but the Bayesian MCMC approach can be used when the complexity of the model renders finding the derivatives difficult if not impossible. Using Bayesian estimation methods, the tendency of ML estimates to shift beyond a reasonable range of values is avoidable (Mislevy, 1986). Moreover, Mislevy (1986) found that when small samples (e.g., usually fewer than 500) or short tests (e.g., the number of test items is fewer than 20) are used, an estimation method such as MBE is recommended. Thus, under such conditions, the selection of estimation method may play a part in detecting DIF using Lord's Wald test (Kim et al., 1994). A Bayesian formulation of MIRT is implemented in BMIRT (Yao, 2003) and uses the MCMC method to estimate the M-2PL and M-3PL models.

CHAPTER 3

METHODOLOGY

A simulation study was conducted to investigate the performance of Lord's Wald test for detecting DIF in terms of the comparison of two estimation methods, MML and MCMC, in the multidimensional IRT framework. This was necessary since comparisons of estimation methods with Lord's Wald test for DIF detection using a multidimensional framework are rare. In this chapter, the simulation design, DIF conditions, evaluation criteria, and empirical distributions of the χ^2 statistics are explained.

3.1 Simulation Design

The simplest two-dimensional MIRT model in terms of the number of dimensions was considered in the simulation study. Using the compensatory multidimensional M-2PL model in Equation 18, the 24-item test and the 46-item test data were generated. The performance of the DIF detection methods is influenced by major factors such as test length, sample size, DIF magnitude, and the number of DIF items in the test (Mazor, Clauser, & Hambleton, 1991; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Four factors that were of primary interest in the study and were important in practice were manipulated: (a) DIF type, (b) DIF magnitude, (c) test length, and (d) sample size differences for the reference group and the focal group.

For the first factor, DIF type, three conditions were considered: non-DIF, uniform DIF, and nonuniform DIF. The non-DIF condition occurs when the data are generated without DIF items. The uniform DIF condition occurs when data are generated with uniform DIF items. When one group is advantaged consistently over the other group across all levels of ability, DIF is in the uniform type. The nonuniform DIF condition

occurs when data are generated with nonuniform DIF items. When one group is advantaged over the other group to a different degree at different locations on the ability scales, nonuniform DIF occurs. In other words, the nonuniform DIF condition is defined as a difference in the *a* parameters (a_1 and/or a_2) between the focal and reference groups along with or without a difference in the intercept, the *d* parameter. Various combinations of uniform DIF and nonuniform DIF can be found in the previous literature (e.g., Oshima et al., 1997; Suh & Cho, 2014; Swaminathan & Rogers, 1990).

For the second factor, DIF magnitude, two different combinations of DIF were considered under a uniform DIF condition (low and medium) and a nonuniform DIF condition (low and medium). In this study, two levels of DIF magnitude similar to Oshima et al.'s study (1997) were chosen: (1) low ($\Delta_{Fa_{1j}} = 0.25, \Delta_{Fa_{2j}} = 0.25$ or $\Delta_{Fa_{2j}} = 0.3, \Delta_{Fd_j} = 0.25$) and (2) medium ($\Delta_{Fa_{1j}} = 0.5, \Delta_{Fa_{2j}} = 0.5$ or $\Delta_{Fa_{2j}} = 0.6$, $\Delta_{Fd_j} = 0.5$). There were two DIF conditions in the uniform DIF and eight DIF conditions in the nonuniform DIF for each DIF magnitude. Table 3.1 summarizes the simulation factors in the study.

The third factor was test length. Test length has been manipulated in many DIF simulation studies (e.g., French & Maller, 2007; Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1996; Paek & Wilson, 2011; Rogers & Swaminathan, 1993; Woods, 2009; Yao & Li, 2010). Previous study results have shown that statistical power increased as test length increased (Narayanan & Swaminathan, 1996). Lim and Drasgow (1990) used a 20-item test, and Snow and Oshima (2009) selected a 40-item test. Swaminathan and Rogers (1990) used three test lengths (40, 60, and 80) to detect DIF in logistic regression and MH procedures. Kim et al. (1994) investigated Lord's Wald test for DIF with a 50-

item test. Kim and Cohen (1995) used data based on a 28-item test. Wang and Ye (2003) used a 25-item test, Zhang (2012) used 30 and 46 items, and Woods (2009) selected 24 and 48 items in the simulation study. In this study, 24 and 46 items were chosen.

The fourth factor considered four combinations of sample sizes. Considerable attention to sample size and its effect on item parameter estimates has been devoted to develop IRT models in numerous previous studies (e.g., Allen & Donoghue, 1996; Goodman, Willse, Allen, & Klaric, 2011; Swaminathan & Gifford, 1986; Swaminathan et al., 2003) and thus to applying MIRT models for investigating DIF detection (e.g., Oshima et al., 1997). Three balanced sample designs and one unbalanced sample design were simulated: (a) reference and focal groups of 1,000 examinees (R1,000/F1,000), (b) reference and focal groups of 3,000 examinees (R3,000/F3,000), (c) reference and focal groups of 5,000 examinees (R5,000/F5,000), and (d) a reference group of 4,000 examinees and a focal group of 2,000 examinees (R4,000/F2,000). The total number of examinees in the focal and reference groups was 2,000 for the small, 6,000 for the medium, and 10,000 for the large sample. A medium sample of 6,000 was used to make a comparison between the M-2PL and M-3PL models using MML and MCMC. Two dimensions (abilities) were generated from a bivariate normal distributions with means of 0, variances of 1 for both dimensions, and the correlation between the two dimensions of 0 for both groups across all simulation conditions. The number of replications was 100 for each condition. In addition to all four factors, the performance of the M-3PL model was examined with influential conditions (e.g., bias and RMSE of item parameters) using MML and MCMC for comparison with the M-2PL analysis.

Table 3.1

Summary of the Simulation Study Factors

Factors in the Study			**Details**			
1. DIF Type			Non-DIF		n/a	
			Uniform DIF		Low	
					Medium	
			Nonuniform DIF		Low	
					Medium	
2. DIF Size			Number of items with DIF	a_1	a_2	d
	Uniform DIF	Low	4 DIF	n/a	n/a	$\Delta_{Fd_j} = 0.25$
			8 DIF			
		Med	4 DIF	n/a	n/a	$\Delta_{Fd_j} = 0.5$
			8 DIF			
	Nonuniform DIF	Low	4 DIF	$\Delta_{Fa_{1j}} = 0.25$	n/a	$\Delta_{Fd_j} = 0.25$
			8 DIF			
			4 DIF	$\Delta_{Fa_{1j}} = 0.25$	$\Delta_{Fa_{2j}} = 0.25$	$\Delta_{Fd_j} = 0.25$
			8 DIF			
			4 DIF	n/a	$\Delta_{Fa_{2j}} = 0.3$	$\Delta_{Fd_j} = 0.25$
			8 DIF			
			4 DIF	n/a	$\Delta_{Fa_{2j}} = 0.3$	n/a
			8 DIF			
		Med	4 DIF	$\Delta_{Fa_{1j}} = 0.5$	n/a	$\Delta_{Fd_j} = 0.5$
			8 DIF			
			4 DIF	$\Delta_{Fa_{1j}} = 0.5$	$\Delta_{Fa_{2j}} = 0.5$	$\Delta_{Fd_j} = 0.5$
			8 DIF			
			4 DIF	n/a	$\Delta_{Fa_{2j}} = 0.6$	$\Delta_{Fd_j} = 0.5$
			8 DIF			
			4 DIF	n/a	$\Delta_{Fa_{2j}} = 0.6$	n/a
			8 DIF			

3. Test Length	Total Item	Number of DIF Items (Proportion of DIF Items)	
	24	4 (.167)	8 (.333)
	46	4 (.086)	8 (.174)
4. Sample Size	Balanced	$N_R = 1,000, N_F = 1,000$	
		$N_R = 3,000, N_F = 3,000$	
		$N_R = 5,000, N_F = 5,000$	
	Unbalanced	$N_R = 4,000, N_F = 2,000$	

3.2 DIF Test Conditions

To assess the performance of Lord's Wald test for detecting DIF, 24-item and 46-item test data sets were generated under four conditions: (a) 20 anchor items with four DIF items, (b) 16 anchor items with eight DIF items, (c) 42 anchor items with four DIF items, and (d) 38 anchor items with eight DIF items. The first two conditions were simulated for the 24-item test, and the latter two conditions were simulated for the 46-item test. Tables 3.2 through 3.5 show the item parameters used to generate anchor items that are the items for the DIF test in a Type I error study (or in other words, the anchor item set) with a two-dimensional framework.

The item parameters of the 24- and 46-item tests were selected to resemble values used in two previous studies (see Reckase, 2009, p. 204; Suh & Cho, 2014), and were used as true item parameters. Motivated by and resembling the design in Suh and Cho's (2014) work, three item clusters were considered, as shown in Tables 3.2 through 3.5. For example, Table 3.2 shows three loading patterns: (a) The first 11 items (i.e., items 1 through 11) were selected to measure predominant loadings along the first dimension (i.e., larger a_1 than a_2), (b) the next 11 items (i.e., items 12 through 22) were selected to measure predominant loadings along the second dimension (i.e., larger a_2 than a_1), and (c) the last 20 items were selected to measure approximately equal balanced loading for both dimensions. The means for the three anchor item clusters under each test set are provided in Tables 3.2 through 3.5. Total means and standard deviations (SDs) across the four test conditions present a very similar pattern; the overall range of the a_1 and a_2 parameters is from 0.59 to 0.62 and the range for the *d* parameter is from –0.12 to –0.03.

Table 3.2

Item Parameters Used in the Two-Dimensional M-2PL Model with 42 Non-DIF (Anchor) Items for 46-Item Test

Item	a_1	a_2	d
1	1.04	0.00	-0.09
2	0.88	0.13	0.27
3	1.05	0.04	1.23
4	1.17	0.02	-0.23
5	1.02	0.23	0.84
6	0.98	0.08	-0.77
7	1.01	0.20	-0.86
8	0.87	0.21	0.02
9	0.97	0.19	-0.22
10	0.98	0.02	-0.12
11	0.92	0.08	-0.77
12	0.09	1.03	0.09
13	0.00	0.96	0.90
14	0.24	0.92	-0.47
15	0.21	0.94	-1.09
16	0.19	0.83	0.41
17	0.04	0.97	-0.58
18	0.06	1.00	-0.88
19	0.16	1.01	1.14
20	0.15	1.13	1.15
21	0.14	0.95	-0.38
22	0.15	0.81	-1.26
23	0.74	0.75	0.29
24	0.66	0.84	-0.52
25	0.81	0.73	-0.62
26	0.67	0.59	-0.44
27	0.70	0.73	-0.91
28	0.71	0.72	-0.47
29	0.55	0.70	-0.75
30	0.68	0.62	0.77
31	0.80	0.76	0.01
32	0.69	0.69	0.10
33	0.60	0.84	1.16
34	0.73	0.68	-0.18
35	0.56	0.74	0.69
36	0.72	0.59	-0.93

37	0.77	0.69	0.58
38	0.67	0.63	-0.33
39	0.82	0.68	-0.16
40	0.72	0.76	0.67
41	0.63	0.66	1.39
42	0.64	0.72	0.04
M(1-11)	0.99	0.11	-0.06
M(12-22)	0.13	0.96	-0.09
M(23-42)	0.69	0.71	0.02
M (Total)	0.62	0.62	-0.03
SD (Total)	0.33	0.33	0.71

Note. M indicates the mean; *SD* indicates the standard deviation.

Table 3.3

Item Parameters Used in the Two-Dimensional M-2PL Model with 38 Non-DIF (Anchor) Items for 46-Item Test

Item	a_1	a_2	d
1	1.04	0.00	-0.09
2	0.88	0.13	0.27
3	1.05	0.04	1.23
4	1.17	0.02	-0.23
5	1.02	0.23	0.84
6	0.98	0.08	-0.77
7	1.01	0.20	-0.86
8	0.87	0.21	0.02
9	0.97	0.19	-0.22
10	0.98	0.02	-0.12
11	0.92	0.08	-0.77
12	0.09	1.03	0.09
13	0.00	0.96	0.90
14	0.24	0.92	-0.47
15	0.21	0.94	-1.09
16	0.19	0.83	0.41
17	0.04	0.97	-0.58
18	0.06	1.00	-0.88
19	0.16	1.01	1.14
20	0.15	1.13	1.15
21	0.14	0.95	-0.38
22	0.15	0.81	-1.26

23	0.74	0.75	0.29
24	0.66	0.84	-0.52
25	0.81	0.73	-0.62
26	0.67	0.59	-0.44
27	0.70	0.73	-0.91
28	0.71	0.72	-0.47
29	0.55	0.70	-0.75
30	0.68	0.62	0.77
31	0.80	0.76	0.01
32	0.69	0.69	0.10
33	0.60	0.84	1.16
34	0.73	0.68	-0.18
35	0.56	0.74	0.69
36	0.72	0.59	-0.93
37	0.77	0.69	0.58
38	0.67	0.63	-0.33
M(1-11)	0.99	0.11	-0.06
M(12-22)	0.13	0.96	-0.09
M(23-38)	0.69	0.71	-0.10
M (Total)	0.62	0.61	-0.08
SD (Total)	0.35	0.35	0.70

Note. M indicates the mean; *SD* indicates the standard deviation.

Table 3.4

Item Parameters Used in the Two-Dimensional M-2PL Model with 20 Non-DIF (Anchor) Items for 24-Item Test

Item	a_1	a_2	*d*
1	1.04	0.00	-0.09
2	0.88	0.13	0.27
3	1.17	0.02	-0.23
4	0.97	0.19	-0.22
5	0.98	0.02	-0.12
6	0.92	0.08	-0.77
7	0.09	1.03	0.09
8	0.00	0.96	0.90
9	0.04	0.97	-0.58
10	0.06	1.00	-0.88
11	0.15	1.13	1.15
12	0.14	0.95	-0.38

13	0.74	0.75	0.29
14	0.70	0.73	-0.91
15	0.71	0.72	-0.47
16	0.80	0.76	0.01
17	0.69	0.69	0.10
18	0.73	0.68	-0.18
19	0.67	0.63	-0.33
20	0.64	0.72	0.04
M(1-6)	0.99	0.07	-0.19
M(7-12)	0.08	1.00	0.05
M(13-20)	0.71	0.71	-0.18
M (Total)	0.61	0.61	-0.12
SD (Total)	0.37	0.37	0.51

Note. M indicates the mean; *SD* indicates the standard deviation.

Table 3.5

Item Parameters Used in the Two-Dimensional M-2PL Model with 16 Non-DIF (Anchor) Items for 24- Item Test

Item	a_1	a_2	*d*
1	1.04	0.00	-0.09
2	0.88	0.13	0.27
3	1.17	0.02	-0.23
4	0.97	0.19	-0.22
5	0.98	0.02	-0.12
6	0.92	0.08	-0.77
7	0.09	1.03	0.09
8	0.00	0.96	0.90
9	0.04	0.97	-0.58
10	0.06	1.00	-0.88
11	0.15	1.13	1.15
12	0.14	0.95	-0.38
13	0.74	0.75	0.29
14	0.70	0.73	-0.91
15	0.71	0.72	-0.47
16	0.80	0.76	0.01
M(1-6)	0.99	0.07	-0.19
M(7-12)	0.08	1.00	0.05
M(13-16)	0.74	0.74	-0.27
M (Total)	0.59	0.59	-0.12

SD (Total)	0.42	0.43	0.58

Note. M indicates the mean; *SD* indicates the standard deviation.

The anchor items parameters for the M-3PL were identical to those for the M-2PL in terms of the two discrimination parameters and intercept parameter. The guessing parameter was fixed to 0.2 across all anchor items and studied items. Four and eight DIF items were generated under each of the four DIF conditions modified from Suh and Cho's (2014) and Oshima et al.'s (1997, p. 264) DIF study designs: low and medium uniform DIF conditions and low and medium nonuniform DIF conditions. For example, in the case of the 46-item test, items 1–42 in Table 3.2 were used as anchor items, and the item parameters for the last four items (items 43–46) in Table 3.6. When items 1–38 in Table 3.3 were used as anchor items, the item parameters for the last eight items (items 39–46) in Table 3.7 were simulated to represent different DIF types for each DIF condition.

When the magnitude and types of DIF items simulated were manipulated, four DIF item patterns were considered in the item discrimination parameters (a_1, a_2) and the intercept parameter (*d*) for the focal group as shown in Tables 3.6 and 3.7. The intercept parameters for all items were increased by either 0.25 or 0.5 for the focal group therefore creating items that were more difficult in the uniform and the nonuniform DIF conditions. The 0.25 difference in the *d* parameters represented a low uniform DIF magnitude. A medium DIF magnitude was simulated with a 0.5 difference in the *d* parameters. For the nonuniform DIF conditions, each of the four items had a different nonuniform DIF pattern. A low nonuniform DIF was introduced at a shift size of 0.25 or 0.3 in the a_1 and/or a_2 parameter, such that the *a* parameters were set 0.25 or 0.3 higher for the focal

group than for the reference group. For example, in the low nonuniform DIF of the 46-item test with eight DIF items condition in Table 3.7, a shift size 0.25 of the a_1 parameter and the *d* parameter was introduced for items 39–40. For items 41–42, a 0.25 shift size for the a_1 parameter, a_2 parameter, and *d* parameter were introduced. For items 43–44, a 0.3 shift size for the a_2 parameter and a 0.25 shift size for the *d* parameter were introduced. For last two items, 45–46, a 0.3 shift size for only the a_2 parameter was introduced. The medium nonuniform DIF was simulated by 0.5 (for items 39–42) for the a_1 parameter, and .6 (for items 43–46) for the a_2 parameter in the same manner as the low nonuniform DIF along with the 0.5 difference in the *d* parameter.

Table 3.6

Item Parameters Used in the Generating DIF Conditions for the Last Four Items

Uniform DIF									
Reference Group				Focal Group					
				Low DIF			Medium DIF		
Item*	a_1	a_2	*d*	a_1	a_2	*d*	a_1	a_2	*d*
43 (21)	1.0	0.1	0.0	1.0	0.1	**0.25**	1.0	0.1	**0.5**
44 (22)	0.1	1.0	0.0	0.1	1.0	**0.25**	0.1	1.0	**0.5**
45 (23)	0.7	0.7	0.0	0.7	0.7	**0.25**	0.7	0.7	**0.5**
46 (24)	1.0	0.6	0.0	1.0	0.6	**0.25**	1.0	0.6	**0.5**
Nonuniform DIF									
Reference Group				Focal Group					
				Low DIF			Medium DIF		
Item*	a_1	a_2	*d*	a_1	a_2	*d*	a_1	a_2	*d*
43 (21)	1.0	0.1	0.0	**1.25**	0.1	**0.25**	**1.5**	0.1	**0.5**
44 (22)	1.0	0.1	0.0	**1.25**	**0.35**	**0.25**	**1.5**	**0.6**	**0.5**
45 (23)	0.7	0.7	0.0	0.7	**1.0**	**0.25**	0.7	**1.3**	**0.5**
46 (24)	0.7	0.7	0.0	0.7	**1.0**	0.0	0.7	**1.3**	0.0

*The number outside of the parentheses indicates the item number for the 46-item test, and the number inside the parentheses represents the item number for the 24-item test, etc.

Table 3.7

Item Parameters Used in the Generating DIF Conditions for the Last Eight Items

Uniform DIF									
Reference Group				Focal Group					
				Low DIF			Medium DIF		
Item*	a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
39 (17)	1.0	0.1	0.0	1.0	0.1	**0.25**	1.0	0.1	**0.5**
40 (18)	1.0	0.1	0.0	1.0	0.1	**0.25**	1.0	0.1	**0.5**
41 (19)	0.1	1.0	0.0	0.1	1.0	**0.25**	0.1	1.0	**0.5**
42 (20)	0.1	1.0	0.0	0.1	1.0	**0.25**	0.1	1.0	**0.5**
43 (21)	0.7	0.7	0.0	0.7	0.7	**0.25**	0.7	0.7	**0.5**
44 (22)	0.7	0.7	0.0	0.7	0.7	**0.25**	0.7	0.7	**0.5**
45 (23)	1.0	0.6	0.0	1.0	0.6	**0.25**	1.0	0.6	**0.5**
46 (24)	1.0	0.6	0.0	1.0	0.6	**0.25**	1.0	0.6	**0.5**
Nonuniform DIF									
Reference Group				Focal Group					
				Low DIF			Medium DIF		
Item*	a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
39 (17)	1.0	0.1	0.0	**1.25**	0.1	**0.25**	**1.5**	0.1	**0.5**
40 (18)	1.0	0.1	0.0	**1.25**	0.1	**0.25**	**1.5**	0.1	**0.5**
41 (19)	1.0	0.1	0.0	**1.25**	**0.35**	**0.25**	**1.5**	**0.6**	**0.5**
42 (20)	1.0	0.1	0.0	**1.25**	**0.35**	**0.25**	**1.5**	**0.6**	**0.5**
43 (21)	0.7	0.7	0.0	0.7	**1.0**	**0.25**	0.7	**1.3**	**0.5**
44 (22)	0.7	0.7	0.0	0.7	**1.0**	**0.25**	0.7	**1.3**	**0.5**
45 (23)	0.7	0.7	0.0	0.7	**1.0**	0.0	0.7	**1.3**	0.0
46 (24)	0.7	0.7	0.0	0.7	**1.0**	0.0	0.7	**1.3**	0.0

*The number outside of the parentheses indicates the item number for the 46-item test, and the number inside the parentheses represents the item number for the 24 item test, etc.

3.3 Item Parameter Estimation from flexMIRT and BMIRT

As with the other estimation methods, model identification constraints were necessary for specific parameters in the M-2PL and M-3PL applications that used MML and MCMC algorithms. For M-2PL and M-3PL, the θ axes were constrained to be orthogonal to address the rotational indeterminacy problem. And the first item of the second dimension was fixed to zero (i.e., $a_{12} = 0$) for both the focal and reference groups. To resolve the metric indeterminacy, θ_1 and θ_2 were fixed to means of 0 and variances of 1 from a bivariate normal distribution for all generating conditions for the reference group. The correlation between the two dimensions was also fixed at 0 for both groups to deal with the model identification problem for the within-item test structure in which all items were loaded on both dimensions. The correlation between the two dimensions cannot be estimated in flexMIRT due to the model identification problem. These identification constraints were also used in Bolt and Lall's (2003) MIRT study. The same constraints used in flexMIRT were imposed in BMIRT to make fair comparisons between the two estimation approaches. The means of the focal group were free to be estimated, whereas the variances of both ability parameters were fixed at 1[3].

FlexMIRT allows the user to specify constraints and prior distributions for item parameters. Beta distribution priors were applied in flexMIRT to add the item

[3] The variances in the focal group were fixed to 1 following other IRT software such as MULTILOG (Thissen, 1991) and it may be unnecessary for identification purposes. Other software, such as NOHARM (Fraser, 1987), also theoretically assumes a multivariate standard normal distribution for latent traits. Oshima et al. (1997) used the multivariate standard normal distribution instead of using estimated theta values when they calculated DIF and DTF indices.

uniquenesses (Bock, Gibbons, & Muraki, 1988) in the form of a beta distribution (α , β) for the slope parameters; flexMIRT has a user-specific α parameter and the β parameter fixed at 1. In this study, α = 1.6 and thus a prior distribution of the form beta (1.6, 1.0) was imposed in the item uniquenesses of items 1 through 24 for the 24-item test and 46 for the 46-item test, respectively. The prior chosen was a normal distribution (–1.4, 0.1) (i.e., $\mu = -1.4$ and $\sigma^2 = 0.1$) on the guessing parameters for all items in M-3PL, for the focal and reference groups. For the guessing parameter, the normal prior applied a normal distribution prior on the logit of the guessing parameter. A normal prior (–1.4, 0.1) was an appropriate prior for the guessing parameter, with a mode around 0.2 in the typical *g* metric using flexMIRT 2 (Houts & Cai, 2013).

It was not possible to check each outcome for convergence in this simulation study for MCMC. Although many statistic tools help select sufficient burn-in and chain length, there is no agreed-upon selection method (Wollack et al., 2002). Geyer (1992) and Patz and Junker (1999b) suggested that the appropriate burn-in length is the number of lags needed to obtain negligible autocorrelations. Wollack et al. (2002) found the autocorrelations were nearly zero for all parameters when the lag between draws was set at 50. A combination of these previous suggestions, a length of 500 iterations, was sufficient for burn-in for all but extremely difficult items (Wollack et al., 2002). In this study, 10,000 iterations, with 5,000 burn-ins as the number of MCMCs to be thrown away were used across all conditions and replications. For M-3PL, the priors for the item parameters in BMIRT were as follows:

$$d_{1j} \sim N(\mu_{d_{1j}}, \sigma^2_{d_{1j}}), \tag{27}$$

$$\log(a_{1jl}) \sim N(\log(\mu_{a_{1jl}}), \sigma^2_{a_{1jl}}) \tag{28}$$

for l = 1 and 2.

$$g_j \sim beta(\alpha, \beta) \tag{29}$$

$\mu_{d_{1j}} = 0$, $\mu_{a_{1j}} = 1$, $\sigma^2_{d_{1j}} = 1$, and $\sigma^2_{a_{1jl}} = 1.5$. Normal priors were assigned to the d parameters, with a mean of 0 and a variance of 1. Lognormal priors were assigned to the slope parameters, with a mean of 1 and a variance of 1.5. Beta priors were assigned to the guessing parameters, with $\alpha = 100$ and $\beta = 400$. All values were default values with BMIRT.

3.4 Evaluation Criteria

For each condition, the Type I error rates and the power of Lord's Wald tests from MML and MCMC were evaluated. To examine the effect of the M-2PL and M-3PL model differences on the item parameter estimation for the focal and reference groups, the bias and root mean square errors (RMSEs) were calculated and compared for non-DIF conditions in this study. The bias and the RMSE were obtained with

$$\text{Bias } (\hat{\eta}) = \frac{\sum_{r}^{R} (\hat{\eta}_r - \eta)}{R} \tag{30}$$

$$\text{RMSE } (\hat{\eta}) = \sqrt{\frac{\sum_{r}^{R} (\hat{\eta}_r - \eta)^2}{R}} \tag{31}$$

where R is the number of replications (i.e., 100), $\hat{\eta}_r$ is the estimated item parameter η at the r^{th} replication, and η is the true item parameter. Smaller RMSEs and bias indicate better estimates of the item parameters. The Type I error of Lord's Wald test is defined as the proportion of times a non-DIF item is erroneously detected as a DIF item across replications, whereas the power is described as the proportion of times a DIF item known

to exhibit DIF is correctly identified across replications. Thus, the Type I error rate was calculated as the proportion of the number of significant χ^2 test statistics (at α = .05) out of total replications (i.e., 100) for each non-DIF item. The power rate was calculated as the proportion of the number of significant χ^2 test statistics out of the total replications for each uniform and nonuniform DIF item. In addition, when the observed power rates of DIF detection were interpreted, Type I error rates must be prudently considered because an uncontrolled Type I error can lead to inadequate power rates. For example, high Type I error rates (higher than the expected value) can inflate the power of the test.

3.5 Empirical Distributions of the χ^2 Statistics

Researchers often observe greater power due to an inflated Type I error rate for the DIF detection method in application. Because the power is affected by the Type I error rate, without corrections, the power can be overestimated or underestimated (de la Torre & Lee, 2013). Thus, in addition to theoretical distributions of the χ^2 statistics, we considered the empirical distributions of the χ^2 statistics obtained from the non-DIF conditions (i.e., by selecting the 95^{th} empirical χ^2 statistics as critical values) to determine the empirical power rates if the Type I error rates were not controlled. Thus, when such corrections were made using the empirical distributions, adjusted power rates were reported.

CHAPTER 4

SIMULATION RESULTS

4.1 Analysis and Comparison of the Estimation Methods

Each DIF condition was analyzed with the MML and MCMC estimation methods. Several evaluative measures are considered for the comparison of the Lord's Wald test estimation methods: the bias and RMSE of item parameter estimates and the overall contribution of the factors to outcomes in terms of the Type I error rate and power rate of the Wald test using two estimation methods. The estimation accuracy was compared and discussed across different DIF conditions; specifically, the impact of the inclusion of the guessing parameter on the estimation accuracy will be discussed at the end of section 4.2 in the M-3PL model simulation results. The estimated accuracy of the parameters for the M-2PL and M-3PL models is also compared. For the recovery study, all three balanced sample sizes for M-2PL and the medium R3,000/F3,000 sample for M-3PL in the 24- and 46-item test with four and eight DIF item conditions were used.

The criterion of interest in the current study is given in three parts: (a) The first part summarizes the parameter recovery results based on the bias and RMSE for the M-2PL and M-3PL models using MML and MCMC estimation methods. (b) The second part summarizes the Type I error rates for Lord's Wald test in the MML and the MCMC estimation methods under various DIF test conditions. (c) The third part summarizes the results of the power study for Lord's Wald test using the MML and MCMC estimation methods to detect uniform and nonuniform DIF items under various DIF detection conditions for the M-2PL and M-3PL models.

4.2 Bias and RMSE of the M-2PL and M-3PL Results

Tables 4.1 to 4.4 report the biases and RMSEs for all M-2PL parameters with the R1,000/F1,000, R3,000/F3,000, and the R5,000/F5,000 samples and M-3PL parameters with the R3,000/F3,000 medium sample in the non-DIF conditions. For the M-3PL, only the medium sample condition was considered, for the comparison purpose with the M-2PL. Tables 4.1 to 4.3 report the M-2PL estimation results in terms of the bias and the RMSE for each parameter type of the focal group. Biases and RMSEs of the focal group were calculated as the average across the items and replications for each parameter type in the non-DIF conditions. For example, in the case of the 24-item test for each item parameter type, $\eta = a_1, a_2,$ or *d*, bias is calculated as $\sum_{j=1}^{24}\sum_{r=1}^{100}\left(\hat{\eta}_{jr} - \eta_j\right)/2400$ and the RMSE was calculated as $\sqrt{\sum_{j=1}^{24}\sum_{r=1}^{100}\left(\hat{\eta}_{jr} - \eta_j\right)^2/2400}$, where $j = 1, \ldots, 24$ denotes the item, $r = 1, \ldots, 100$ denotes the replication, and $\hat{\eta}_{jr}$ is the item parameter estimate. For the 46-item test conditions, the RMSEs are reported the same way as for the 24-item test conditions, except for $j = 1, \ldots, 46$.

When the bias and RMSE of the item discriminations or intercept parameter estimates were examined across the test length and the number of DIF items for the three balanced sample size conditions with M-2PL (Tables 4.1 to 4.3), the biases for the a_1, a_2, and *d* parameter estimates that used MML resulted in consistently higher biases than those used MCMC across all DIF conditions. This result does not agree with the results of Wollack et al.'s (2002) and Kieftenbeld and Natesan's (2012) studies in which the recovery of discrimination parameters for MML were better than those for MCMC using the nominal response model and the graded response model, respectively.

In terms of the test length results, the 24-item test conditions tended to be less biased than the 46-item test conditions for the a_1 and a_2 parameters using the MML and MCMC estimations. However, the bias of the *d* parameter estimates decreased as the test length increased using MML and MCMC. The biases of the a_1 and a_2 parameters from the 24-item test with four DIF items (.17 DIF) appeared to be lower than those of other conditions, whereas the bias of the *d* parameter from the 46-item test with four DIF items (.10 DIF) appeared to be lower than that from other conditions. In other words, the RMSEs of the short test (24) with four DIF items was more accurate for a_1 and a_2 parameters with all sample sizes, whereas the RMSEs of long test (46) for estimating the *d* parameter were more accurate than the short test.

Regarding the number of DIF item results, the biases of the item parameter estimates from the four-DIF item conditions were slightly lower than those for the eight-DIF item conditions for the a_1 and a_2 parameters using MML and MCMC. For the *d* parameter estimates, the bias increased as the number of DIF items increased when MCMC was used for all test length conditions, whereas the bias decreased for the 24-item test and increased for the 46-item test as the number of DIF item increased when MML was used. These patterns of bias and RMSE of the a_1, a_2, and d parameters using MML and MCMC were found regardless of the sample size.

For the sample size effect on the recovery based on Tables 4.1 to 4.3, the biases and RMSEs tended to decrease as the sample size increased for MML and MCMC on average. This pattern was more apparent with the slope parameters that used MML. For easier interpretation of the sample size effect, the patterns of the RMSEs in Tables 4.1 to 4.3 are plotted as Figures 4.1 through 4.3.

Table 4.1

Parameter Recovery (Bias and Root Mean Square Errors) for the Multidimensional Two-Parameter Logistic Model (M-2PL) Focal group of Sample Size 1000 ($\alpha = 0.05$)

		M-2PL							
		24_4_1000		24_8_1000		46_4_1000		46_8_1000	
Parameter		MML	MCMC	MML	MCMC	MML	MCMC	MML	MCMC
a_1	Bias	**0.317**	**0.085**	0.359	0.118	0.397	0.105	0.411	0.109
	RMSE	**0.318**	**0.089**	0.359	0.120	0.397	0.107	0.411	0.111
a_2	Bias	**0.269**	**0.093**	0.316	0.124	0.371	0.119	0.387	0.109
	RMSE	**0.269**	**0.097**	0.316	0.126	0.372	0.121	0.387	0.111
d	Bias	0.214	0.058	0.201	0.078	**0.056**	**0.013**	0.149	0.024
	RMSE	0.215	0.062	0.202	0.082	**0.057**	**0.029**	0.150	0.033
Average (Bias)		**0.267**	**0.079**	0.292	0.107	0.275	0.079	0.316	0.081
Average (RMSE)		**0.267**	**0.083**	0.292	0.109	0.275	0.086	0.316	0.085

Note. Bold-face numbers represent the smallest values for MML and MCMC in the rows.

Table 4.2

Parameter Recovery (Bias and Root Mean Square Errors) for the Multidimensional Two-Parameter Logistic Model (M-2PL) Focal group of Sample Size 3000 ($\alpha = 0.05$)

		M-2PL							
		24_4_3000		24_8_3000		46_4_3000		46_8_3000	
Parameter		MML	MCMC	MML	MCMC	MML	MCMC	MML	MCMC
a_1	Bias	**0.162**	**0.069**	0.192	0.094	0.208	0.093	0.226	0.097
	RMSE	**0.163**	**0.070**	0.192	0.095	0.209	0.094	0.226	0.099
a_2	Bias	**0.127**	**0.081**	0.160	0.104	0.191	0.102	0.199	0.103
	RMSE	**0.127**	**0.082**	0.160	0.105	0.192	0.104	0.199	0.104
d	Bias	0.215	0.035	0.201	0.039	**0.055**	**0.005**	0.148	0.030
	RMSE	0.216	0.037	0.201	0.043	**0.055**	**0.019**	0.148	0.032
Average (Bias)		0.168	**0.062**	0.184	0.079	**0.151**	0.067	0.191	0.077

Average (RMSE)	0.168	**0.063**	0.184	0.081	**0.151**	0.072	0.191	0.078

Note. Bold-face numbers represent the smallest values for MML and MCMC in the rows.

Table 4.3

Parameter Recovery (Bias and Root Mean Square Errors) for the Multidimensional Two-Parameter Logistic Model (M-2PL) Focal group of Sample Size 5000 ($\alpha = 0.05$)

		M-2PL							
		24_4_5000		24_8_5000		46_4_5000		46_8_5000	
Parameter		MML	MCMC	MML	MCMC	MML	MCMC	MML	MCMC
a_1	Bias	**0.116**	**0.065**	0.140	0.089	0.147	0.096	0.161	0.102
	RMSE	**0.116**	**0.067**	0.140	0.090	0.147	0.097	0.161	0.103
a_2	Bias	**0.088**	**0.082**	0.113	0.103	0.134	0.109	0.137	0.104
	RMSE	**0.089**	**0.084**	0.113	0.103	0.134	0.110	0.137	0.105
d	Bias	0.215	0.054	0.199	0.072	**0.055**	**0.002**	0.148	0.027
	RMSE	0.215	0.055	0.199	0.073	**0.056**	**0.016**	0.148	0.033
Average (Bias)		0.140	**0.067**	0.151	0.088	**0.112**	0.069	0.149	0.078
Average (RMSE)		0.140	**0.069**	0.151	0.089	**0.112**	0.074	0.149	0.080

Note. Bold-face numbers represent the smallest values for MML and MCMC in the rows.

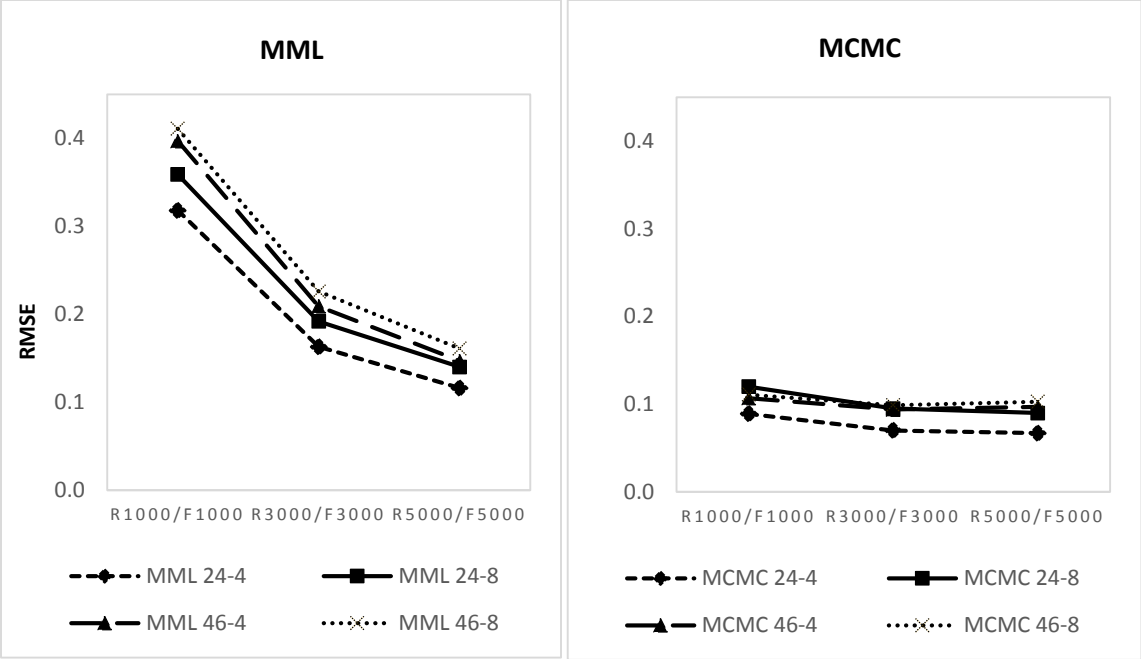

Figure 4.1. RMSE of a_1 parameter for Sample Sizes in the M-2PL

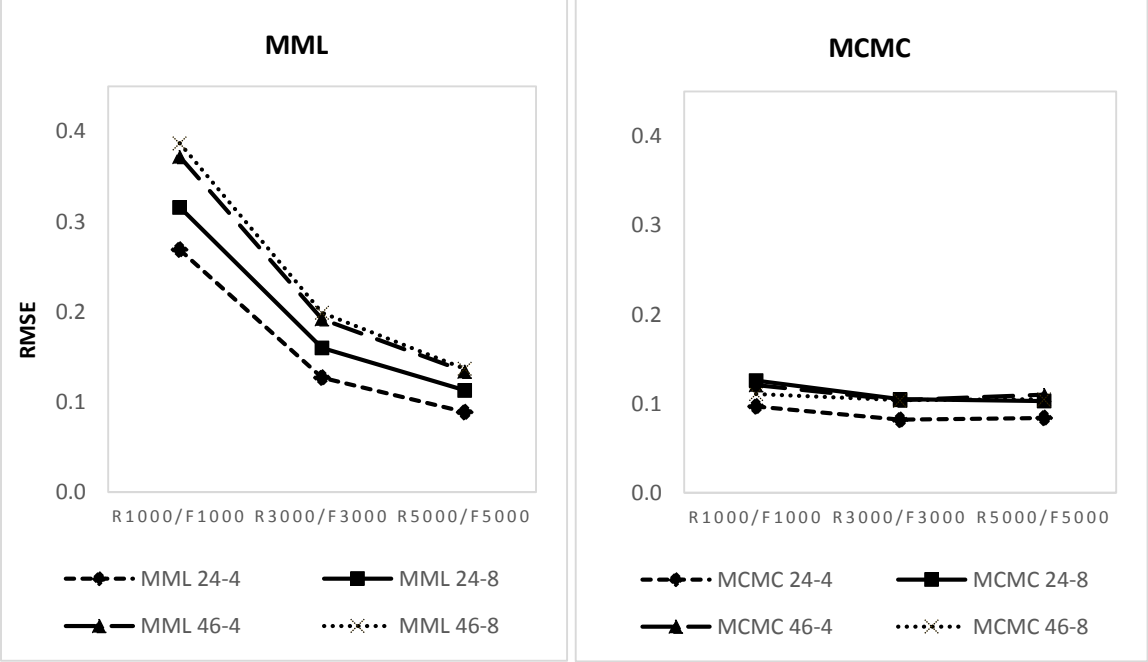

Figure 4.2. RMSE of a_2 parameter for Sample Sizes in the M-2PL

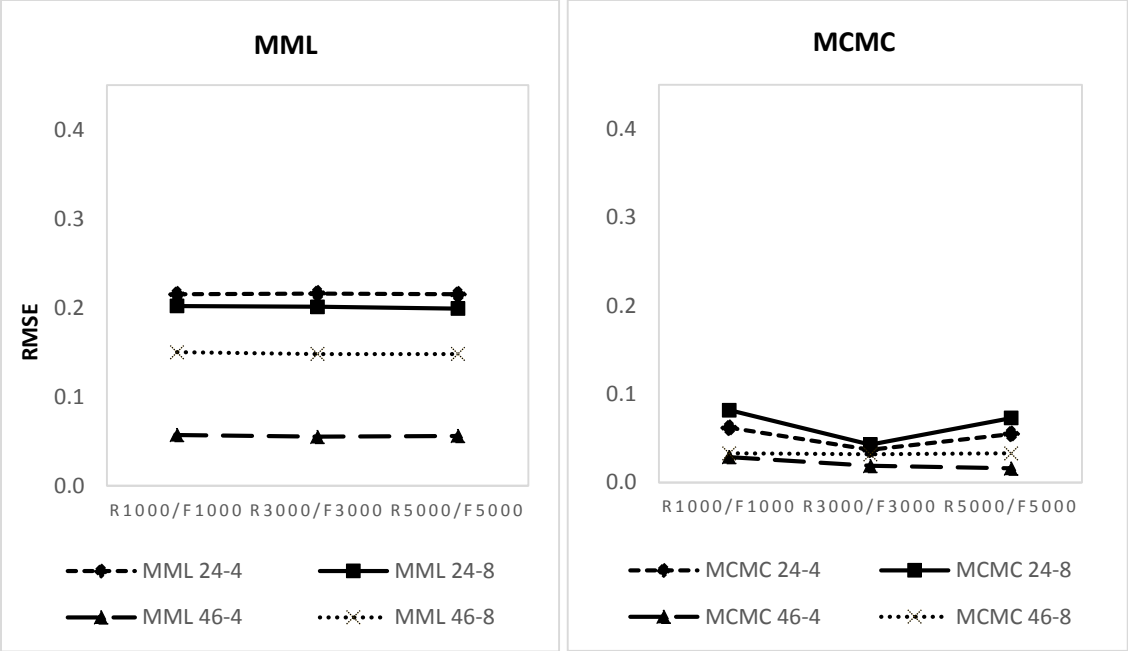

Figure 4.3. RMSE of *d* parameter for Sample Sizes in the M-2PL

Table 4.4 reports the bias and RMSE results for the focal group for M-3PL. The similar bias and RMSE patterns of the a_1 and a_2 parameters when MML was used from M-2PL were observed in M-3PL (the bias and RMSE of the guessing parameter will be reported in Table 4.5). However, one interesting finding for M-3PL compared to M-2PL is that MML was no longer inferior to MCMC in the 46-item test with four-DIF item condition, in which the bias and RMSE of the a_2 and *d* parameters for MML were lower than those for MCMC.

For the test lengthwise results, the bias and RMSE of the slope parameters increased as the test length increased for MML and MCMC except for the eight-DIF item conditions when MCMC was used. These results for the a_1 and a_2 parameters differ from those of previous item parameter recovery studies for the nominal response model (Lee, 1997; Wollack & Cohen, 1997; Wollack et al., 2002), which showed that, as the

test length increased (i.e., a long test, 40 items or more), the RMSE of the slope parameter estimates decreased. Similar to M-2PL, the bias and RMSE of the *d* parameter decreased as the test length increased when MML was used. The estimation of the *d* parameter differed from the results for the a_1 and a_2 parameters. The bias and RMSE of the a_1 parameter from the 24-item test with four DIF items was smaller than other test length conditions for MML and MCMC. The bias and RMSE of the a_2 parameter for MML was the smallest in the 24-item test with four DIF items, whereas those for MCMC were the smallest on the 46-item test with eight DIF items. Similar to M-2PL, the bias and RMSE of the *d* parameter from the 46-item test with four DIF items (.10 DIF) appeared to be the lowest among all conditions when MML was used. Unlike M-2PL, the bias and RMSE from the 46-item test with eight DIF items (.17 DIF) appeared to be the lowest of that from other conditions when MCMC was used.

Biases and RMSEs are also reported for the guessing parameters of M-3PL in Table 4.5. Consistent with Table 4.4, the estimation of the guessing, *g*, parameters for MML tended to be slightly worse than those for MCMC. However, regardless of the estimation methods, the parameter recovery for guessing tended to be reasonably well approximated across all simulation conditions. The RMSE values of MML were close to each other across all DIF conditions, whereas the RMSE values of MCMC were close to each other with the exception of the 46-item test with four DIF items, on which the highest value was observed.

The results in Table 4.4 are somewhat similar to those from a previous parameter estimation recovery study that used MML for M-3PL (Zhang, 2012). From Zhang's (2012) simulation study, the conditions of 30- and 46-item tests with a 3,000 sample and

a correlation between the subscales of zero were selected to compare the results to the similar simulation conditions in the present study, such the 24-item and 46-item tests with a 3,000 sample that used the MML estimation method. Zhang (2012) reported that the average RMSE of the a_1 and a_2 parameters was 0.093 and 0.089 for the 30-item test and 0.098 and 0.092 for the 46-item test, respectively. These values are smaller than those in Table 4.4. For the intercept (or difficulty) and guessing parameters, more similar results were observed. The average RMSE of the *d* (intercept) parameter was 0.091 for the 30-item test and 0.105 for the 46-item test. The average RMSE of the guessing parameter was 0.049 for the 30-item test and 0.050 for the 46-item test.

For the results for the number of DIF items, a pattern similar to that for M-2PL was found in M-3PL for the a_1, a_2, and *d* parameters when MML was used on the 24-item test. For the MCMC estimation method, a clear pattern in M-3PL was observed. For the 24-item tests, the small number of DIF items (i.e., four DIF items) conditions resulted in less bias and RMSE than the large number of DIF items (i.e., eight DIF items) conditions for the a_1, a_2, and *d* parameters. For the 46-item tests, the opposite results were found: The large number of DIF items (i.e., eight DIF items) conditions resulted in less bias and RMSE than the small number of DIF items (i.e., four DIF items) conditions for the a_1, a_2, and *d* parameters.

Table 4.4

Parameter Recovery (Bias and Root Mean Square Errors) for the Multidimensional Three-Parameter Logistic Model (M-3PL) Focal group of Sample Size 3000 ($\alpha = 0.05$)

		M-3PL							
		24_4_3000		24_8_3000		46_4_3000		46_8_3000	
Parameter		MML	MCMC	MML	MCMC	MML	MCMC	MML	MCMC
a_1	Bias	**0.191**	**0.065**	0.220	0.095	0.234	0.205	0.251	0.071
	RMSE	**0.192**	**0.067**	0.220	0.097	0.234	0.235	0.251	0.073
a_2	Bias	**0.168**	0.083	0.197	0.111	0.221	0.212	0.229	**0.081**
	RMSE	**0.168**	0.084	0.197	0.112	0.221	0.243	0.229	**0.083**
d	Bias	0.157	0.029	0.139	0.035	**-0.003**	0.137	0.090	**0.005**
	RMSE	0.158	0.032	0.140	0.038	**0.009**	0.181	0.093	**0.019**
Average (Bias)		0.172	0.059	0.185	0.080	**0.151**	0.185	0.190	**0.052**
Average (RMSE)		0.172	0.061	0.186	0.082	**0.155**	0.220	0.191	**0.058**

Note. Bold-face numbers represent the smallest values for MML and MCMC in the rows.

Table 4.5

Guessing Parameter Recovery (Bias and Root Mean Square Errors) for the Multidimensional Three-Parameter Logistic Model (M-3PL) Focal group of Sample Size 3000 ($\alpha = 0.05$)

		M-3PL							
		24_4_3000		24_8_3000		46_4_3000		46_8_3000	
Parameter		MML	MCMC	MML	MCMC	MML	MCMC	MML	MCMC
g	Bias	0.012	-0.004	0.012	-0.003	0.011	-0.022	0.010	-0.004
	RMSE	0.012	0.004	0.012	0.004	0.011	0.026	0.010	0.004

Figures 4.4 through 4.6 are graphical representations of the summary in Tables 4.2 and 4.4 that show the RMSE values of the three parameters, a_1, a_2, and d, respectively, under M-2PL and M-3PL for MCMC and MML across the test length and number of DIF items. The RMSEs increased from M-2PL to M-3PL with MML for the a_1 and a_2 parameter estimation. Unlike other parameters, the RMSEs of the *d* parameter when MML was used decreased from M-2PL to M-3PL regardless of the test length. However, for the RMSE results for MCMC, a different pattern for MML was found in the parameter estimation. The RMSEs of the a_1, a_2, and *d* parameters either decreased or were invariant from M-2PL to M-3PL except for the 46-item test with four DIF item condition, which increased from M-2PL to M-3PL.

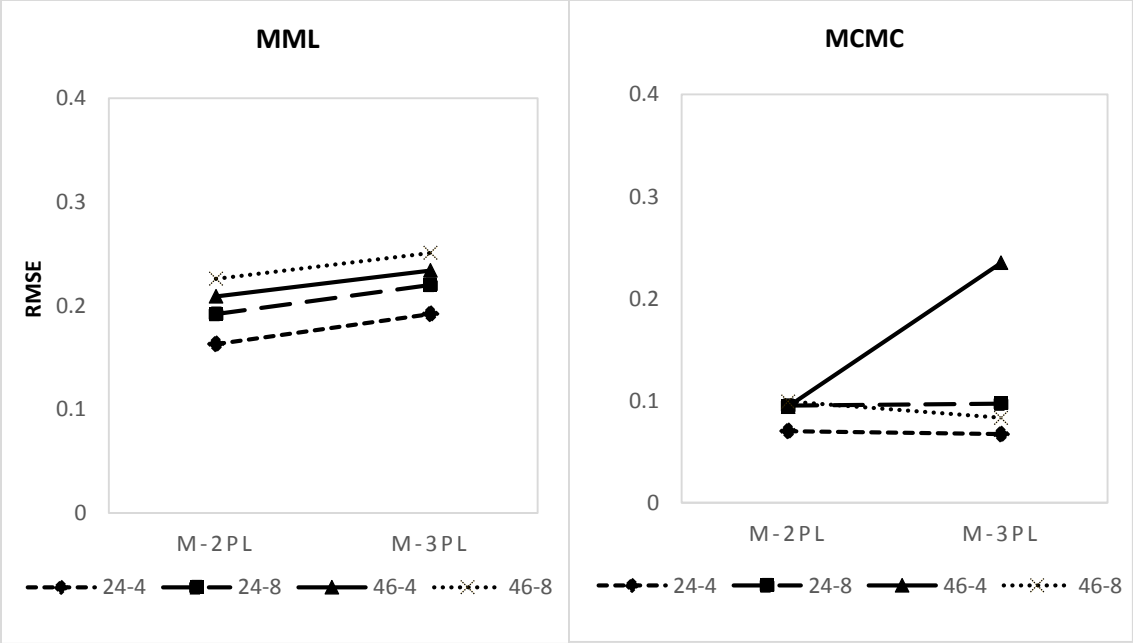

Figure 4.4. RMSE of a_1 parameter in the M-2PL and the M-3PL

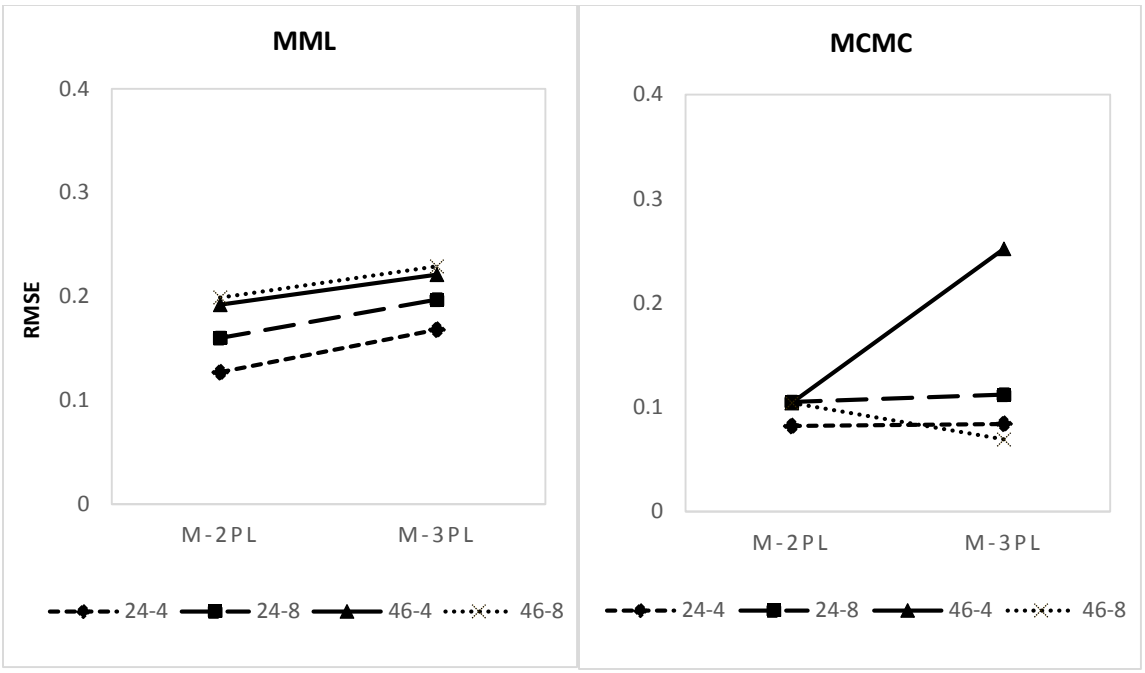

Figure 4.5. RMSE of a_2 parameter in the M-2PL and the M-3PL

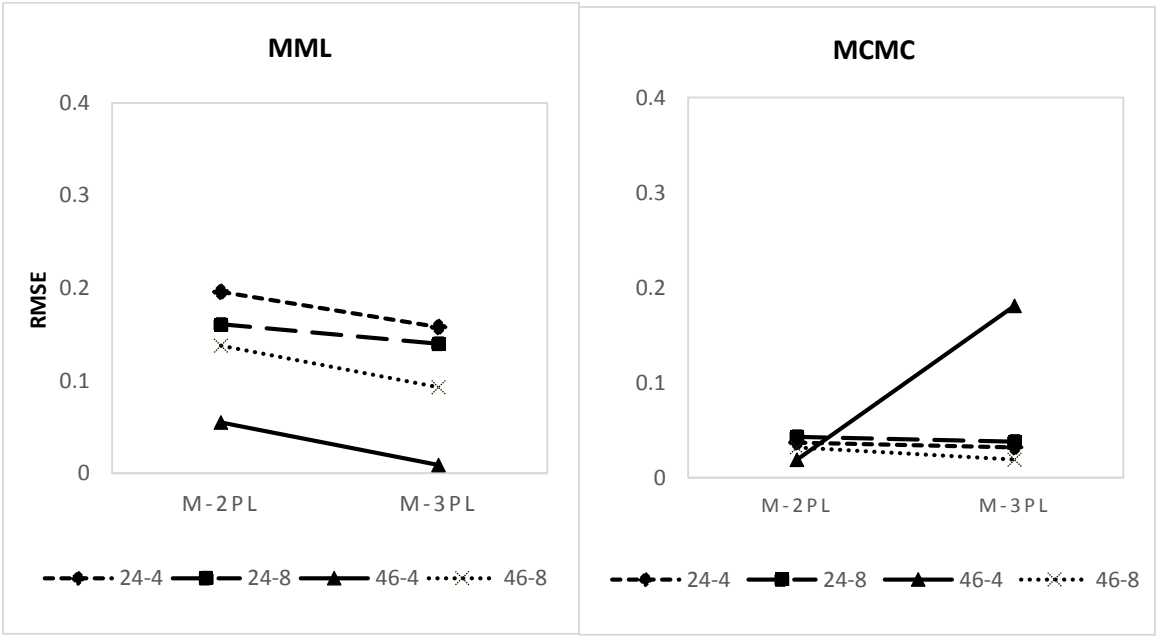

Figure 4.6. RMSE of *d* parameter in the M-2PL and the M-3PL

4.3 The Type I Error Study

4.3.1 The M-2PL Results. The Type I error rate for this study was calculated and defined as follows. For each item, the percentage of times the item was detected as DIF out of the 100 replications under each non-DIF condition was calculated. A Type I error rate close to 5% was ideally expected given the significance level of 0.05. If the Type I error rate was less than 0.05, it was conservative in control; if the Type I error rate was greater than 0.05, it was inflated. The average Type I error rates across all items are reported separately for the MML and the MCMC estimation methods for each non-DIF condition.

Table 4.6 shows the M-2PL Type I error rates of the four items under the 24-item and the 46-item test conditions as a function of sample size for separately identifying DIF at $\alpha = 0.05$ of the MML and MCMC estimation approaches. In Table 4.6, the average Type I error rates for MCMC are substantially higher than those of the MML method irrespective of the sample size factor. The average error rates from MML ranged from 0.0 to 0.08, whereas those from MCMC ranged from 0.23 to 0.43 at the nominal alpha level of 0.05. The unbalanced sample condition yielded the highest Type I error rates with both estimation methods. That is, the unbalanced sample design in Table 4.6 has larger Type I error rates than the balanced sample design; the Type I error rates for MML vs. MCMC were 0.07 vs. 0.43 for the 24-item test and 0.08 vs. 0.43 for the 46-item test, respectively. Within the balanced sample design, the largest Type I error rates for MML were observed in the large sample (R5,000/F5,000) condition. In contrast, the largest Type I error rates with MCMC were observed in the small-sample (R1,000/F,1000) condition.

Inflated Type I error rates for MCMC were found in all sample sizes. As the sample size increased under the balanced sample size designs, the Type I error rates with MCMC tended to be less inflated. In addition, the Type I error rates for MCMC decreased as the test length increased (from the 24-item test to the 46-item test) with one exception in the small-sample (R1,000/F,1000) condition. The patterns regarding sample size and test length in MML were the opposite for those in MCMC. The Type I error rates increased slightly as the sample size increased in the balanced sample design. As the test length increased, MML slightly increased from the 24-item test with four DIF items to the 46-item test with four DIF items. However, all Type I error rates fell below the nominal level of 0.05 (e.g., between 0.00 and 0.03) except for the unbalanced design condition. There was no clear pattern regarding different DIF items.

Table 4.6

Type I Error Results for the M-2PL in the 24-item Test and 46-item Test with 4 Studied Items

Sample Size	Balanced						Unbalanced		Average	
	R1000/F1000		R3000/F3000		R5000/F5000		R4000/F2000			
DIF Item	MML	MCMC	MML	MCMC	MML	MCMC	MML	MCMC	MML	MCMC
21	0.00	0.41	0.01	0.42	0.00	0.36	0.07	0.43	0.02	0.41
22	0.01	0.36	0.00	0.31	0.03	0.34	0.08	0.37	0.03	0.35
23	0.01	0.39	0.03	0.37	0.03	0.37	0.05	0.46	0.03	0.40
24	0.00	0.49	0.01	0.39	0.02	0.44	0.08	0.44	0.03	0.44
Average	0.00	0.41	0.01	0.37	0.02	0.38	**0.07**	**0.43**	0.03	0.40
43	0.00	0.40	0.01	0.26	0.04	0.21	0.07	0.39	0.03	0.32
44	0.01	0.46	0.00	0.30	0.02	0.18	0.08	0.40	0.03	0.34
45	0.00	0.44	0.02	0.40	0.03	0.31	0.06	0.45	0.02	0.40
46	0.00	0.39	0.03	0.41	0.04	0.22	0.09	0.46	0.04	0.37
Average	0.00	0.42	0.02	0.34	0.03	0.23	**0.08**	**0.43**	0.03	0.36

Note. Bold-face numbers represent the largest values for MML and MCMC in the rows.

Table 4.7 shows the M-2PL Type I error rates of eight items under the 24-item and the 46-item test conditions as a function of sample size for separately identifying DIF at $\alpha = 0.05$ of the MML and the MCMC estimation approaches. The patterns were similar to the Type I error rates shown in Table 4.4. The average Type I error rates for MML ranged from 0.01 to 0.13, and those for MCMC ranged from 0.39 to 0.49, which are slightly higher than the values in Table 4.6 on average. This implies that the Type I error from the two estimation methods increased slightly as the number of DIF items increased. The unbalanced design again provided the largest error rates for both test length conditions regardless of the estimation method. Similar to the four DIF item conditions in Table 4.6, the largest difference between the two estimation methods was 0.01 for MML and 0.42 for MCMC in the small-sample condition (R1,000/F,1000). Although the general patterns regarding the sample size and test length were similar to those in Table 4.6, such patterns were less apparent in Table 4.7. One interesting finding with the unbalanced sample design in eight DIF item conditions was that the Type I error rates for MML on the short test (24) were higher than those for MML on the long test (46).

In summary, regardless of all test factors (length of 24-item vs. 46-item, the number of DIF items, four DIF vs. eight DIF, and sample sizes of R1,000/F1,000 vs. R3,000/F3,000 vs. R5,000/F5,000), the Type I error rates for the unbalanced sample design (R4,000/F2,000) were notably the highest among all simulation conditions that used MML and MCMC, and the MML provided more accurate Type I error rates than MCMC.

Table 4.7

Type I Error Results for the M-2PL in the 24-item Test and 46-item Test with 8 Studied Items

	Balanced						Unbalanced			
Sample Size	R1000/F1000		R3000/F3000		R5000/F5000		R4000/F2000		Average	
DIF Item	MML	MCMC	MML	MCMC	MML	MCMC	MML	MCMC	MML	MCMC
17	0.00	0.37	0.01	0.34	0.00	0.26	0.13	0.48	0.03	0.36
18	0.00	0.42	0.00	0.35	0.02	0.37	0.24	0.47	0.07	0.40
19	0.00	0.39	0.01	0.34	0.02	0.35	0.17	0.44	0.05	0.38
20	0.02	0.37	0.03	0.37	0.00	0.45	0.10	0.45	0.04	0.41
21	0.01	0.45	0.01	0.41	0.01	0.36	0.10	0.48	0.03	0.46
22	0.01	0.36	0.00	0.46	0.00	0.50	0.11	0.52	0.03	0.46
23	0.01	0.52	0.01	0.46	0.02	0.35	0.09	0.52	0.03	0.46
24	0.03	0.41	0.01	0.41	0.02	0.49	0.07	0.54	0.03	0.46
Average	0.01	0.41	0.01	0.39	0.01	0.39	**0.13**	**0.49**	0.04	0.42
39	0.01	0.38	0.00	0.32	0.04	0.21	0.15	0.42	0.05	0.33
40	0.00	0.45	0.00	0.32	0.03	0.32	0.09	0.41	0.03	0.38
41	0.01	0.33	0.01	0.35	0.09	0.29	0.09	0.43	0.05	0.35
42	0.00	0.46	0.01	0.36	0.02	0.23	0.09	0.47	0.03	0.38
43	0.01	0.49	0.01	0.45	0.08	0.28	0.08	0.53	0.03	0.44
44	0.01	0.39	0.00	0.46	0.03	0.31	0.09	0.57	0.03	0.43
45	0.01	0.42	0.00	0.35	0.05	0.41	0.07	0.53	0.03	0.43
46	0.01	0.44	0.02	0.41	0.02	0.30	0.08	0.50	0.03	0.41
Average	0.01	0.42	0.01	0.38	0.05	0.29	**0.09**	**0.48**	0.04	0.39

Note. Bold-face numbers represent the largest values for MML and MCMC in the rows.

4.3.2 The M-3PL Results. Table 4.8 reports the M-3PL results in terms of Type I error rates for all test lengths and the number of DIF items in the medium sample (R3,000/F3,000) conditions. The average Type I error rates for MML ranged from 0.02 to 0.07, and the average Type I error rates for MCMC increased, ranging from 0.34 to 0.45 at the significance level of 0.05. As the test length increased, the Type I error rates for MCMC were less inflated, whereas the Type I error rates for MML increased to higher

than the nominal level. As the number of DIF items increased, the Type I error rates for MCMC and MML tended to increase.

Table 4.8

Type-I Error Results for the M-3PL in the 24-item Test and 46-Item Test

Sample Size	R3000/F3000						
Test Length	24-item test			46-item test			
DIF Item	MML	MCMC	Average	DIF Item	MML	MCMC	Average
21	0.03	0.39	0.21	43	0.05	0.31	0.18
22	0.00	0.41	0.20	44	0.07	0.31	0.19
23	0.02	0.43	0.23	45	0.06	0.37	0.22
24	0.03	0.52	0.28	46	0.06	0.38	0.22
Average	0.02	0.44	**0.23**		0.06	0.34	**0.20**
17	0.04	0.35	0.20	39	0.04	0.40	0.22
18	0.05	0.50	0.28	40	0.04	0.34	0.19
19	0.10	0.38	0.24	41	0.04	0.27	0.16
20	0.04	0.48	0.26	42	0.08	0.46	0.27
21	0.08	0.49	0.29	43	0.07	0.39	0.23
22	0.04	0.50	0.27	44	0.07	0.41	0.24
23	0.06	0.51	0.29	45	0.10	0.39	0.25
24	0.09	0.40	0.25	46	0.08	0.41	0.25
Average	0.06	0.45	**0.25**		0.07	0.38	**0.23**

4.3.3 A Comparison of the M-2PL and M-3PL Results. Figures 4.7 and 4.8 illustrate the comparison results for M-2PL and M-3PL. The average Type I error rates for MML increased slightly from M-2PL to M-3PL on the 24-item test and the 46-item test regardless of the number of DIF items (four DIF vs. eight DIF items). The Type I error rates for M-2PL tended to be invariant regardless of the test length. However, the

degree of increase in the Type I error rates for MML with M-3PL tended to increase when the test length increased from the 24-item test to the 46-item test. Unlike MML, the Type I error rates for MCMC from the M-3PL decreased slightly when the test length increased from the 24-item test to the 46-item test. The Type I error rates for both MML and MCMC increased from M-2PL to M-3PL with 24-item test. The Type I error rates for MML increased, whereas MCMC decreased slightly from M-2PL to M-3PL when the test length was 46 items.

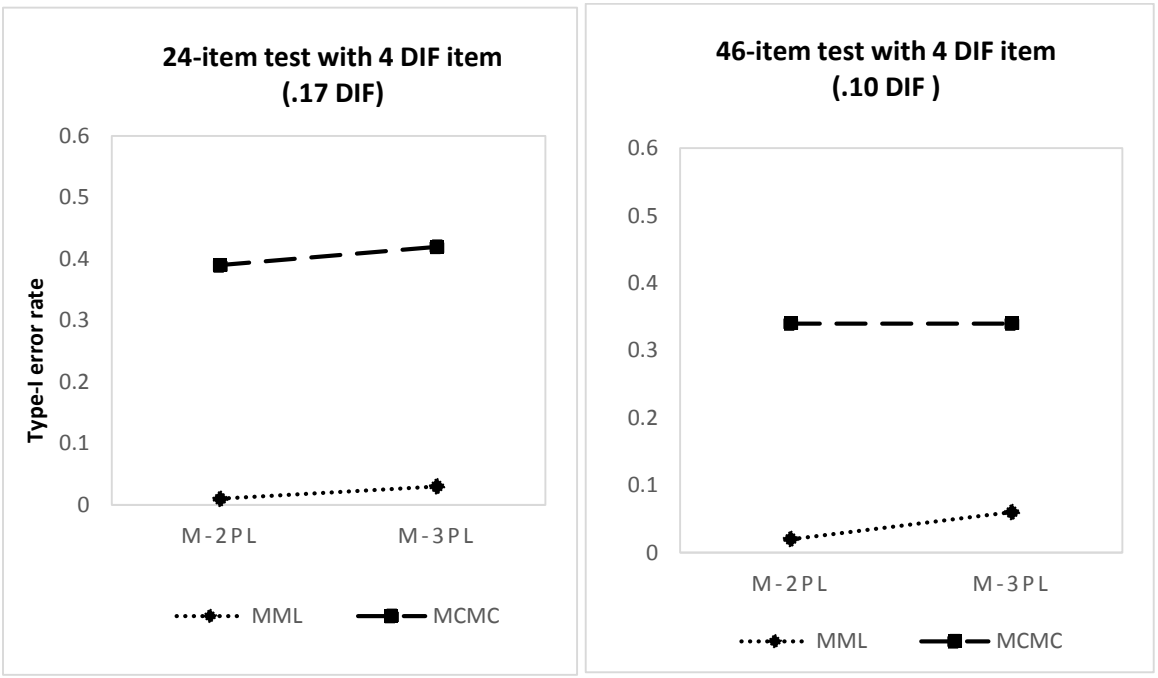

Figure 4.7. Type-I Error Rates of 4-Studied Items Conditions in the M-2PL and the M-3PL (R3000/F3000)

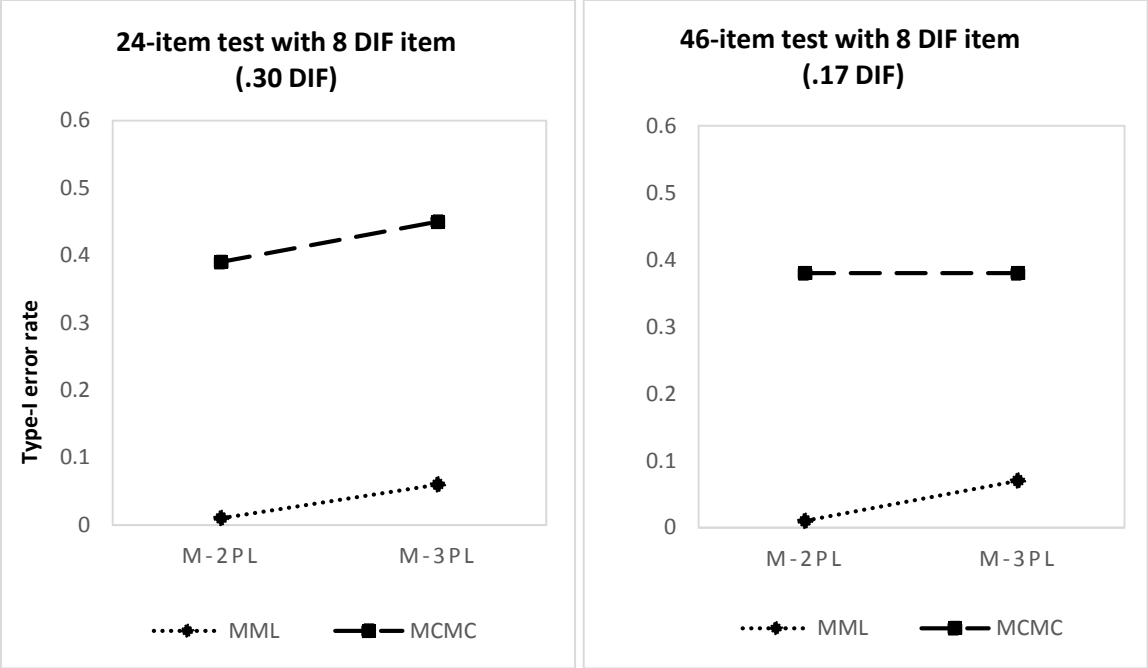

Figure 4.8. Type-I Error Rates of 8-Studied Items Conditions in the M-2PL and the M-3PL (R3000/F3000)

4.4 The Power Study

4.4.1 The M-2PL Results. For the simulation study, the power rate of Lord's Wald test was calculated in the similar way as the Type I error rate was calculated. For each studied item, the percentage of times the item was detected as DIF out of 100 replications was calculated for each DIF condition. In general, a widely used power cutoff value of at least 0.80 is acceptable for evaluating power results (Cohen, 1977).

The interpretation of power rates depends on the Type I error rate control for a given significance level. Because Type I error rates vary under different conditions and power rates are affected by uncontrolled Type I error rates, without correction, power can be overestimated or underestimated (de la Torre & Lee, 2013). Consequently, a direct comparison of the power rates from the theoretical χ^2 distribution of the Lord's Wald

test would be misleading. Thus, comparable power rates across the simulation conditions by controlling for Type I error rates were computed using the 95th percentiles from the empirical χ^2 distributions of the Lord's Wald test. The empirical distributions of the Lord's Wald test under the null hypothesis of no DIF were obtained for each item in the non-DIF conditions.

Uniform DIF Results. *Theoretical vs. empirical power rates.* Tables 4.9 through 4.12 list the power results for uniform DIF conditions based on the theoretical and the empirical power rates of MML and MCMC calculated using the χ^2 (df = 3) distribution for the DIF magnitude, test length, number of DIF items, and sample size. Due to the Type I error control issue, comparing Lord's Wald test statistic to the theoretical distribution can lead to underestimated or overestimated power. Thus, empirical power rates were computed selecting the 95th percentiles of the empirical χ^2 statistics as critical values and were compared with the theoretical power rates on Tables 4.9 through 4.12. Values in parentheses refer to the empirical power rates.

The theoretical and empirical power rates of the 24-item test with four DIF items in low and medium uniform DIF conditions in Table 4.9 were substantially different for the two estimation methods, MML and MCMC. The theoretical power rates for MCMC in uniform DIF conditions were higher than the empirical power rates especially with the small sample (R1,000/F1,000). For example, in the case of the low DIF condition with the 24-item test in R1,000/F1,000, the average theoretical power rates vs. the average empirical power rates were 0.90 vs. 0.50 for MCMC in the uniform DIF condition. Tables 4.10 through 4.12 report the power rates for the 24-item test with eight DIF items (.30 DIF), the 46-item test with four DIF items (.10 DIF), and the 46-item test with eight

DIF items (.17 DIF) in terms of the DIF magnitude for the uniform DIF conditions. The average theoretical power rates vs. the average empirical power rates for MCMC in the low uniform DIF condition with the small sample (R1,000/F,1000) were 0.90 vs. 0.54, 0.91 vs. 0.51, and 0.91 vs. 0.59, respectively. A similar pattern was found for the medium DIF conditions shown in Tables 4.9 through 4.12. For example, the average theoretical power rates vs. the average empirical power rates for R1,000/F1,000 from MCMC for the 24-item test with four DIF items, the 24-item test with eight DIF items, the 46-item test with four DIF items, and the 46-item test with eight DIF items were 1.0 vs. 0.96, 1.0 vs. 0.97, 1.0 vs. 0.99, and 1.0 vs. 0.97, respectively. However, the degree of increase was much smaller than that for the low DIF conditions from MCMC.

However, the power rates for MML increased from the theoretical to the empirical power rates, because the Type I error rates for MML were often smaller than the expected value. For example, for the 24-item test with four DIF items (.17 DIF), the 24-item test with eight DIF items (.30 DIF), the 46-item test with four DIF items (.10 DIF), and the 46-item test with eight DIF items (.17 DIF) for the low uniform DIF conditions, the average theoretical power rates vs. the average empirical power rates for MML with the small sample (R1,000.F1,000) were 0.17 vs. 0.51, 0.20 vs. 0.51, 0.14 vs. 0.38, and 0.12 vs. 0.42, respectively. For the medium DIF conditions, the pattern was similar to the low DIF conditions shown in Tables 4.9 through 4.12. For example, the average theoretical power rates vs. the average empirical power rates in the small sample (R1,000/F1,000) for the 24-item test with four DIF items, the 24-item test with eight DIF items, the 46-item test with four DIF items, and the 46-item test with eight DIF items were 0.89 vs. 0.98, 0.90 vs. 0.99, 0.88 vs. 0.96, and 0.84 vs. 0.90, respectively. Similar to

the MCMC, the degree of increase from MML was also much smaller than that of the low DIF conditions. Overall, the difference between the theoretical power rates and the empirical power rates for MML and MCMC decreased as the sample size and DIF magnitude increased.

Empirical power results. The graphical analyses of effects by test length, number of DIF items, sample size, and DIF magnitude on the average empirical power rates of MML and MCMC are shown in Figures 4.9 through 4.12. Figures 4.9 and 4.10 show that sample size affected the empirical power rates especially with the low DIF conditions. The figures show that the small sample (R1,000/F1,000) had lower empirical power rates than other sample sizes did in the low DIF magnitude condition. However, the large sample (R5,000/F5,000) had higher empirical power rates than other sample size conditions in the low DIF magnitude; all test conditions for MML and MCMC produced excellent power rates (1.0).

For the medium DIF magnitude, the empirical power rates of all test conditions for MML and MCMC resulted in close to perfect power rates (above 0.90–1.0) regardless of sample size. Under the medium uniform with four DIF conditions (shown in Figure 4.9), the empirical power rates of small sample design (R1,000/F1,000) with both test lengths conditions for MML and MCMC were close to perfect (1.0). For all other medium uniform with eight DIF conditions (shown in Figure 4.10), the empirical power rates for all DIF conditions were 1.0 except for the 24-item test in medium (R3,000/F3,000 and R4,000/F2,000) and large (R5,000/F5,000) sample size conditions.

Table 4.9

Power Results for the M-2PL in the 24- item Test Conditions with 4 Uniform DIF Items (.17 DIF)

Sample Size		Balanced							Unbalanced		
		R1000/F1000		R3000/F3000		R5000/F5000			R4000/F2000		
		$\alpha = .05$		$\alpha = .05$		$\alpha = .05$			$\alpha = .05$		
DIF Type	DIF Item	MML[a]	MCMC[a]	MML[a]	MCMC[a]	MML[a]	MCMC[a]	Average	MML[a]	MCMC[a]	Average
Low Uniform	21	.24 (.52)	.94 (.45)	.84 (.95)	.99 (.94)	1.0 (1.0)	1.0 (1.0)	.84 (.81)	.95 (.92)	1.0 (.95)	.98 (.94)
	22	.13 (.53)	.93 (.51)	.87 (.99)	1.0 (.96)	1.0 (1.0)	1.0 (1.0)	.82 (.83)	.95 (.89)	1.0 (.97)	.98 (.93)
	23	.18 (.56)	.89 (.58)	.91 (.97)	.99 (.97)	1.0 (1.0)	1.0 (1.0)	.83 (.85)	.98 (.98)	.99 (.88)	.99 (.93)
	24	.12 (.43)	.83 (.45)	.85 (.96)	1.0 (.87)	1.0 (1.0)	1.0 (.99)	.80 (.78)	.94 (.91)	1.0 (.85)	.97 (.88)
Average		.17 (.51)	.90 (.50)	.87 (.97)	1.0 (.94)	1.0 (1.0)	1.0 (1.0)		.96 (.93)	1.0 (.91)	
Medium Uniform	21	.90 (.99)	1.0 (.93)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.98 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	22	.84 (.96)	1.0 (.97)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.97 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	23	.96 (.98)	1.0 (.97)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.99 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	24	.87 (.97)	1.0 (.96)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.98 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
Average		.89 (.98)	1.0 (.96)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)		1.0 (1.0)	1.0 (1.0)	

Note. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning.
[a]The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parenthesis indicates the proportion of significant Wald test statistics from the empirical distribution.

Table 4.10

Power Results for the M-2PL in the 24- item Test Conditions with 8 Uniform DIF Items (.30 DIF)

Sample Size		Balanced							Unbalanced		
		R1000/F1000		R3000/F3000		R5000/F5000			R4000/F2000		
		$\alpha = .05$		$\alpha = .05$		$\alpha = .05$			$\alpha = .05$		
DIF Type	DIF Item	MML[a]	MCMC[a]	MML[a]	MCMC[a]	MML[a]	MCMC[a]	Average	MML[a]	MCMC[a]	Average
Low Uniform	17	.22 (.57)	.93 (.77)	.88 (.97)	1.0 (.92)	1.0 (1.0)	1.0 (1.0)	.84 (.87)	.96 (.95)	.99 (.76)	.99 (.90)
	18	.15 (.43)	.91 (.33)	.93 (1.0)	1.0 (.95)	1.0 (1.0)	1.0 (1.0)	.83 (.79)	.99 (.95)	1.0 (.85)	.99 (.94)
	19	25 (.50)	.88 (.57)	.82 (.99)	1.0 (.95)	1.0 (1.0)	1.0 (1.0)	.83 (.84)	1.0 (1.0)	1.0 (.82)	1.0 (.96)
	20	.19 (.57)	.85 (.47)	.83 (.96)	1.0 (.88)	1.0 (1.0)	1.0 (1.0)	.81 (.81)	1.0 (1.0)	.99 (.86)	.99 (.91)
	21	.23 (.48)	.90 (.55)	.84 (.91)	.99 (.95)	1.0 (1.0)	1.0 (1.0)	.83 (.82)	1.0 (.99)	.99 (.88)	1.0 (.89)
	22	.21 (.61)	.98 (.79)	.90 (.99)	.99 (.97)	1.0 (1.0)	1.0 (.99)	.85 (.89)	1.0 (.99)	1.0 (.90)	.99 (.89)
	23	.16 (.42)	.84 (.43)	.91 (.98)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	.82 (.80)	.90 (.84)	.98 (.56)	.97 (.84)
	24	.21 (.53)	.92 (.41)	.90 (.96)	.99 (.89)	1.0 (1.0)	1.0 (.99)	.84 (.80)	.85 (.83)	.98 (.45)	.94 (.84)
Average		.20 (.51)	.90 (.54)	.88 (.97)	1.0 (.94)	1.0 (1.0)	1.0 (1.0)		.96 (.94)	.99 (.76)	
Medium Uniform	17	.90 (1.0)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.98 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0))
	18	.91 (.98)	1.0 (.94)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.99 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	19	.96 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.99 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	20	.95 (.99)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.99 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	21	.91 (1.0)	1.0 (.97)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.99 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	22	.90 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.98 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	23	.85 (.95)	1.0 (.91)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.98 (.98)	1.0 (1.0)	1.0 (.99)	1.0 (1.0)
	24	.84 (.98)	1.0 (.93)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.97 (.99)	1.0 (1.0)	1.0 (.99)	1.0 (1.0)
Average		.90 (.99)	1.0 (.97)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)		1.0 (1.0)	1.0 (1.0)	

Note. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning.

[a]The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parenthesis indicates the proportion of significant Wald test statistics from the empirical distribution.

Table 4.11

Power Results for the M-2PL in the 46- item Test Conditions with 4 Uniform DIF Items (.10 DIF)

Sample Size		Balanced							Unbalanced		
		R1000/F1000		R3000/F3000		R5000/F5000			R4000/F2000		
		$\alpha = .05$		$\alpha = .05$		$\alpha = .05$			$\alpha = .05$		
DIF Type	DIF Item	MML[a]	MCMC[a]	MML[a]	MCMC[a]	MML[a]	MCMC[a]	Average	MML[a]	MCMC[a]	Average
Low Uniform	43	.13 (.38)	.98 (.45)	.92 (.97)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	.84 (.80)	.99 (.99)	1.0 (.88)	1.0 (.94)
	44	.16 (.45)	.89 (.62)	.88 (1.0)	1.0 (.98)	1.0 (1.0)	1.0 (1.0)	.82 (.84)	1.0 (1.0)	1.0 (.92)	1.0 (.96)
	45	.17 (.40)	.91 (.52)	.92 (.94)	1.0 (.94)	1.0 (1.0)	1.0 (1.0)	.83 (.80)	.99 (.99)	.98 (.85)	.99 (.92)
	46	.11 (.30)	.85 (.46)	.85 (.86)	.99 (.89)	1.0 (1.0)	1.0 (1.0)	.80 (.75)	.90 (.89)	1.0 (.80)	.95 (.85)
Average		.14 (.38)	.91 (.51)	**.89 (.94)**	1.0 (.95)	1.0 (1.0)	1.0 (1.0)		.97 (.97)	1.0 (.86)	
Medium Uniform	43	.94 (.97)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.99 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	44	.89 (.97)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.98 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	45	.88 (.96)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.98 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	46	.80 (.92)	1.0 (.97)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.97 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
Average		.88 (.96)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)		1.0 (1.0)	1.0 (1.0)	

Note. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning.
[a]The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parenthesis indicates the proportion of significant Wald test statistics from the empirical distribution.

Table 4.12

Power Results for the M-2PL in the 46- item Test Conditions with 8 Uniform DIF Items (.17 DIF)

Sample Size		Balanced							Unbalanced		
		R1000/F1000		R3000/F3000		R5000/F5000			R4000/F2000		
		$\alpha = .05$		$\alpha = .05$		$\alpha = .05$			$\alpha = .05$		
DIF Type	DIF Item	MML[a]	MCMC[a]	MML[a]	MCMC[a]	MML[a]	MCMC[a]	Average	MML[a]	MCMC[a]	Average
Low Uniform	39	.11 (.47)	.89 (.57)	.86 (.96)	1.0 (.94)	1.0 (1.0)	1.0 (1.0)	.81 (.82)	.99 (.96)	.98 (.84)	.99 (.90)
	40	.13 (.42)	.88 (.52)	.94 (.97)	1.0 (.97)	1.0 (1.0)	1.0 (1.0)	.83 (.81)	.99 (.98)	.99 (.89)	.99 (.94)
	41	.13 (.43)	.89 (.82)	.90 (.99)	1.0 (.96)	.99 (.98)	1.0 (1.0)	.82 (.86)	1.0 (1.0)	.99 (.92)	1.0 (.96)
	42	.12 (.53)	.93 (.47)	.86 (.93)	1.0 (.85)	1.0 (1.0)	1.0 (.99)	.82 (.80)	1.0 (.99)	.97 (.82)	.99 (.91)
	43	.11 (.24)	.90 (.60)	.88 (.99)	1.0 (.95)	1.0 (1.0)	1.0 (1.0)	.82 (.80)	1.0 (.98)	.99 (.79)	1.0 (.89)
	44	.15 (.52)	.95 (.66)	.89 (.98)	1.0 (.96)	1.0 (1.0)	1.0 (1.0)	.83 (.85)	.99 (.98)	.99 (.79)	.99 (.89)
	45	.08 (.38)	.92 (.62)	.82 (.95)	1.0 (.91)	1.0 (1.0)	1.0 (1.0)	.80 (.81)	.94 (.91)	.99 (.76)	.97 (.84)
	46	.13 (.34)	.90 (.43)	.85 (.99)	1.0 (.95)	1.0 (1.0)	1.0 (1.0)	.81 (.79)	.88 (.80)	.99 (.87)	.94 (.84)
Average		.12 (.42)	.91 (.59)	.88 (.97)	1.0 (.94)	1.0 (1.0)	1.0 (1.0)		.97 (.95)	.99 (.84)	
Medium Uniform	39	.84 (.98)	1.0 (.97)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.97 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0))
	40	.89 (.99)	1.0 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.98 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	41	.84 (.94)	1.0 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.97 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	42	.84 (.97)	.99 (.92)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.97 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	43	.86 (.94)	1.0 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.98 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	44	.90 (.97)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.98 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	45	.82 (.99)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.97 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	46	.76 (.92)	1.0 (.95)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.96 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
Average		.84 (.90)	1.0 (.97)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)		1.0 (1.0)	1.0 (1.0)	

Note. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning.

[a]The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parenthesis indicates the proportion of significant Wald test statistics from the empirical distribution.

Figures 4.9 and 4.10 also show that test length affected the empirical power rates for certain sample size conditions such as R1,000/F,1000 and R4,000/F2,000. In the R1,000/F1,000 sample conditions, lower empirical power rates from MML are shown as the test length increased, whereas higher empirical power rates from MCMC are shown as the test length increased. In the R4,000/F2,000 unbalanced sample conditions, there is an inconsistency across the number of DIF item conditions. For the four DIF item conditions, the empirical power rates of the 46-item test were higher than that of the 24-item test for MML, whereas the empirical power rates of the 24-item test were higher than those of the 46-item test for MCMC in Figure 4.9. In contrast, for the eight DIF item conditions, the empirical power rates of the 46-item test were higher than those of the 24-item test for MML and MCMC, as shown in Figure 4.10.

In summary, sample size and DIF magnitude had a distinctive impact on the empirical power rates as expected. Regardless of all other factors, the average empirical power rates of medium DIF magnitude were higher than the average empirical power rates of low DIF magnitude across all uniform DIF conditions. Furthermore, the graphs in Figures 4.9 through 4.10 illustrate that the test length noticeably affects the results of two sample size conditions, R1,000/F1,000 and R4,000/F2,000 in low uniform conditions.

Regarding the total sample size comparison of 6,000, the balanced sample design (R3,000/F3,000) with the (1:1) ratio condition and the unbalanced sample design (R4,000/F2,000) with the (2:1) ratio condition are compared in Figures 4.11 and 4.12. There is no clear pattern between the balanced and unbalanced sample designs in terms of the empirical power rates of the MML and MCMC estimation methods under uniform DIF conditions. High (i.e., above 0.85) empirical power rates for all sample design

conditions were found across all DIF conditions for MML and MCMC except for the 24-item test with the eight DIF item condition for the unbalanced sample design (R4,000/F2,000) using MCMC in Figure 4.12, which shows the lowest empirical power rates (around 0.75) of all other low uniform DIF conditions. Overall, when the sample size ratio was balanced (1:1), excellent power rates (above 0.90) were acquired from MML and MCMC for detecting DIF across uniform DIF conditions. However, when the sample size ratio was unbalanced (2:1), sufficient power rates (above 0.80) were obtained from MML and MCMC for detecting DIF with the exception with the 24-item test with eight DIF condition from MCMC. MML performed better than MCMC on average. Considering these results of uniform DIF conditions in terms of the sample design effect, MML seems to be a better choice for detecting DIF for balanced and unbalanced sample size DIF situations.

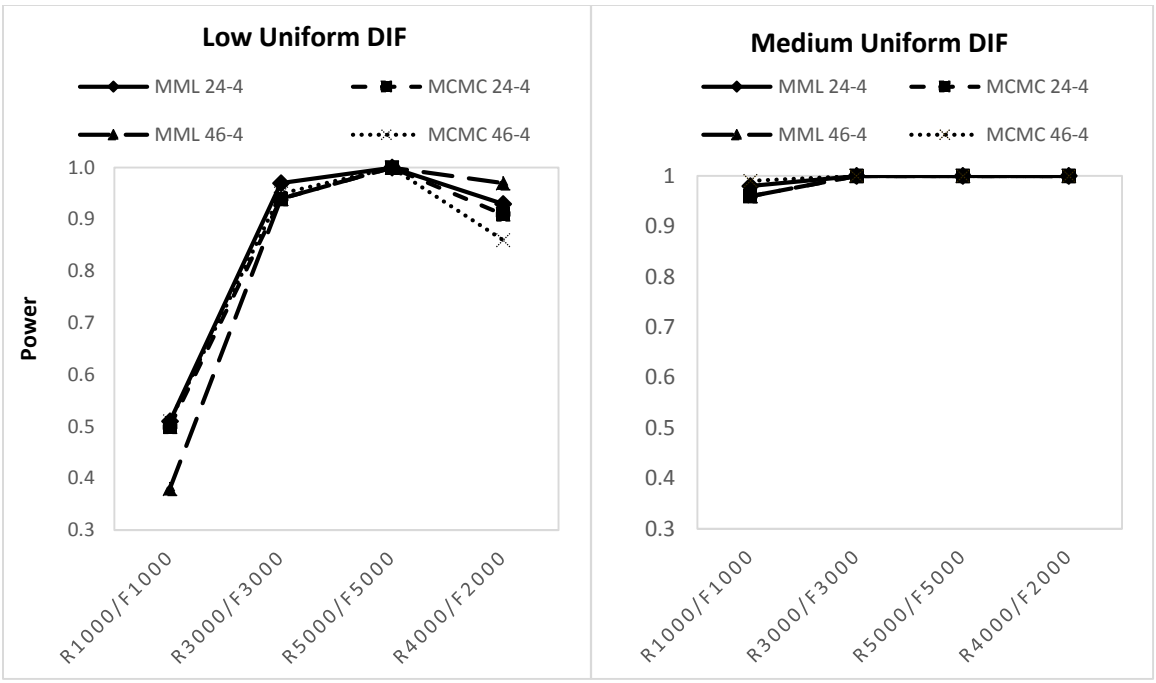

Figure 4.9. Average Power Rates of M-2PL by Test Length, the Number of DIF Items (4 DIF), and Sample Size with Low and Medium Uniform DIF Conditions

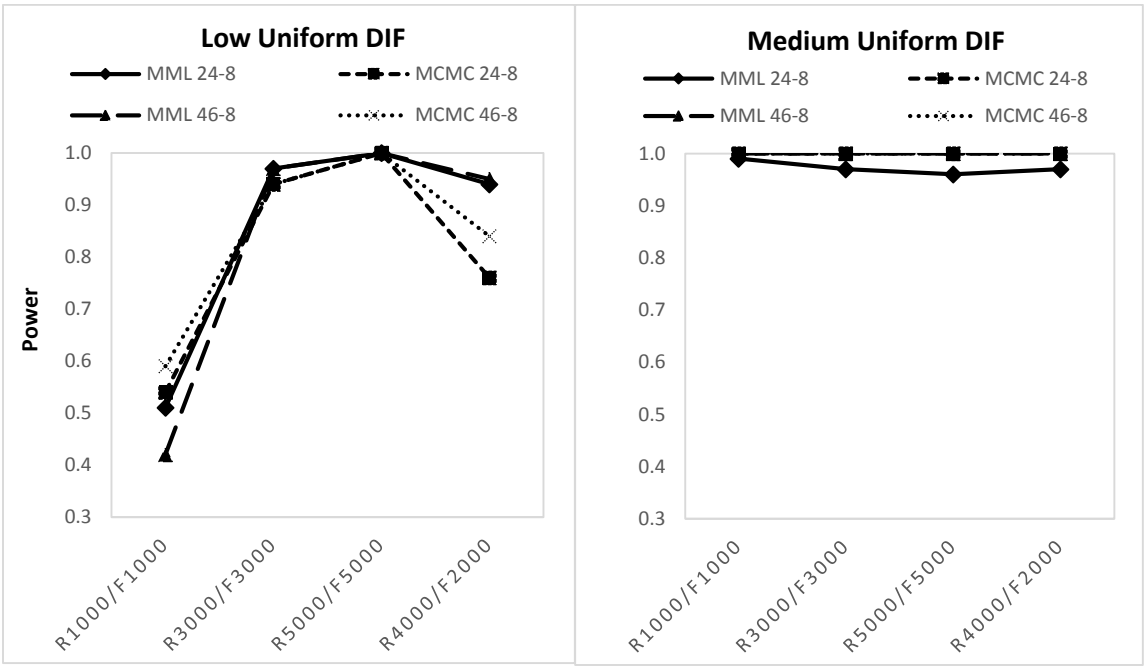

Figure 4.10. Average Power Rates of M-2PL by Test Length, the Number of DIF Items (8 DIF), and Sample Size with Low and Medium Uniform DIF Conditions

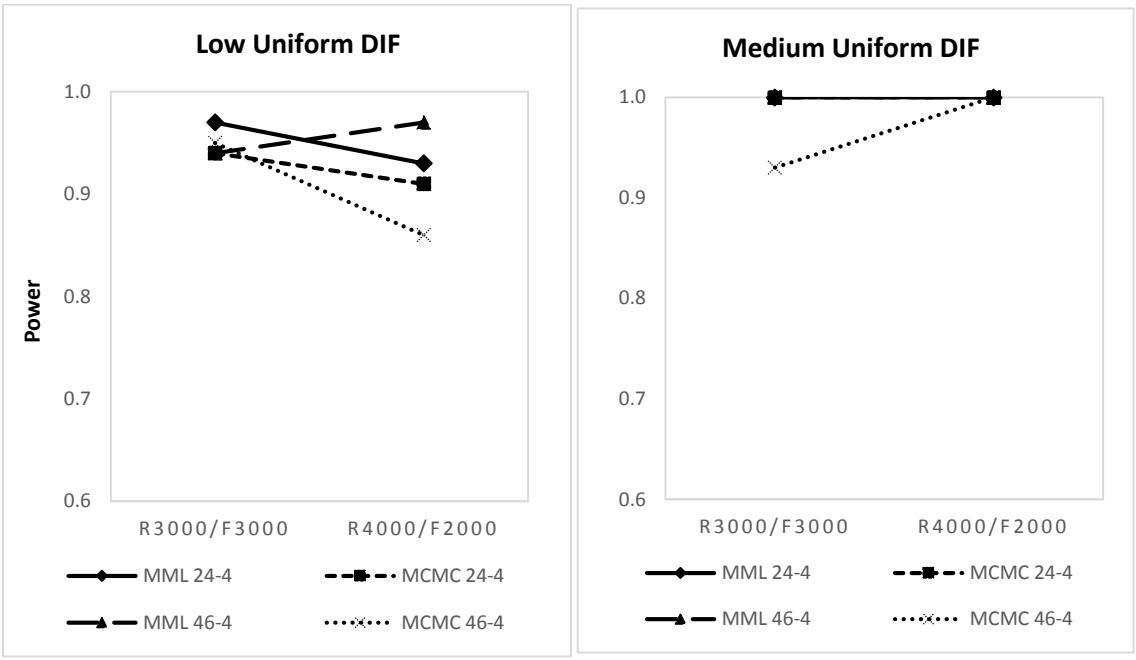

Figure 4.11. Average Power Rates of Low and Medium Uniform DIF with 4 DIF Items Conditions of the M-2PL between (R3000/F3000) and (R4000/F2000)

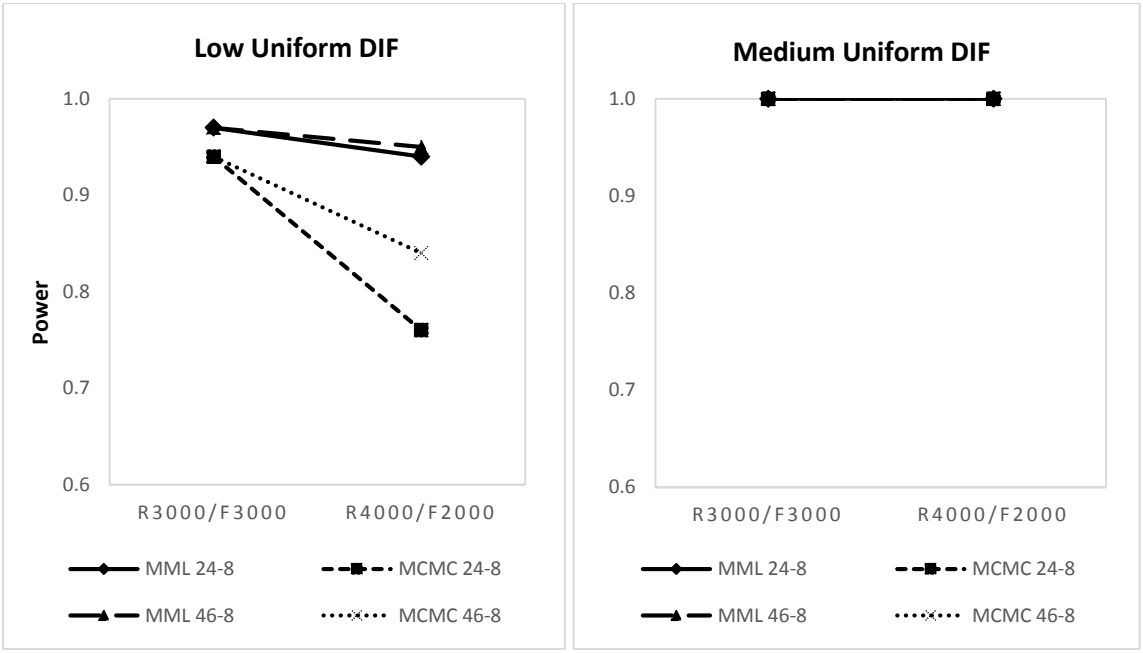

Figure 4.12. Average Power Rates of Low and Medium Uniform DIF with 8 DIF Items form the M-2PL between (R3000/F3000) and (R4000/F2000)

Nonuniform DIF Results*. Theoretical vs. empirical power rates.* Tables 4.13 through 4.16 list the power results for nonuniform DIF conditions determined with the theoretical and empirical power rates calculated using χ^2 (*df* = 3) for the DIF magnitude, test length, number of DIF items, and sample size conditions, respectively. Similar to the performance of the uniform DIF detection, the average empirical power rates for Lord's Wald test from MCMC were lower than the theoretical power rates for detecting nonuniform DIF. However, the average empirical power rates from MML were higher than the theoretical power rates.

The theoretical and empirical power rates of the 24-item test with four DIF items in low and medium nonuniform DIF conditions shown in Table 4.13 were substantially different for two estimation methods, MML and MCMC. Similar to the uniform DIF conditions, the theoretical power rates of MCMC in the nonuniform DIF conditions were higher than the empirical rates especially with the small sample (R1,000/F1,000). For example, in the case of the low nonuniform DIF condition with the 24-item test in R1,000/F1,000, as shown in Table 4.13, the average theoretical power rates vs. the average empirical power rates were 0.83 vs. 0.37 for MCMC. Tables 4.14 through 4.16 report power rates for the 24-item test with eight DIF items (.30 DIF), the 46-item test with four DIF items (.10 DIF), and the 46-item test with eight DIF items (.17 DIF) in terms of DIF magnitude for nonuniform DIF conditions. The average theoretical power rates vs. the average empirical power rates from MCMC in the low nonuniform DIF condition with a small sample (R1,000/F,1000) were 0.81 vs. 0.39, 0.89 vs. 0.38, and 0.87 vs. 0.45, respectively. A similar pattern was found in the medium DIF conditions shown in Tables 4.13 through 4.16. For example, the average theoretical power rates vs.

the average empirical power rates from MCMC for the 24-item test with four DIF items, the 28-item test with eight DIF items, the 46-item test with four DIF items, and the 46-item test with eight DIF items were 0.98 vs. 0.81, 0.96 vs. 0.78, 0.99 vs. 0.82, and 0.98 vs. 0.86, respectively. The average empirical power rates of the medium DIF magnitude conditions were higher than those of the low DIF magnitude conditions irrespective of all other factors. Similar to the uniform DIF conditions, the degree of increase was much smaller than that of the low DIF conditions from MCMC.

In contrast, the theoretical power rates of MML tended to be underestimated under nonuniform DIF conditions, compared to the empirical power rates. For example, for the 24-item test with four DIF items (.17 DIF), the 24-item test with eight DIF items (.30 DIF), the 46-item test with four DIF items (.10 DIF), and the 46-item test with eight DIF items (.17 DIF) for low nonuniform DIF conditions, the average theoretical power rates vs. the average empirical power rates of R1,000/F1,000 from MML were 0.33 vs. 0.63, 0.35 vs. 0.67, 0.39 vs. 0.64, and 0.39 vs. 0.70, respectively. For the medium nonuniform DIF conditions, the pattern was similar to the low nonuniform DIF conditions shown in Tables 4.13 through 4.16. For example, the average theoretical power rates vs. the average empirical power rates of R1,000/F1,000 from MML for the 24-item test with four DIF items, the 24-item test with eight DIF items, the 46-item test with four DIF items, and the 46-item test with eight DIF items were 0.93 vs. 0.97, 0.92 vs. 0.98, 0.93 vs. 0.98, and 0.96 vs.0.99, respectively. Similar to MCMC, the degree of increase from MML was also much smaller than that of the low nonuniform DIF conditions. Overall, the difference between the theoretical power rates and the empirical

power rates of MML and MCMC decreased as the sample size and DIF magnitude increased.

Empirical power results. For easier interpretation of Tables 4.13 to 4.16, Figures 4.13 through 4.16 illustrate the average empirical power results for nonuniform DIF of M-2PL that used MML and MCMC. As shown in Figures 4.13 and 4.14, the sample sizes affected the empirical power rates. The small sample (R1,000/F1,000) had lower empirical power rates than the other sample size in the low DIF magnitude condition. The large sample (R5,000/F5,000) had higher empirical power rates than the other sample size conditions in low DIF magnitude. However, unlike the low uniform DIF in the large sample conditions, not all test conditions for MML and MCMC produced excellent power rates (1.0). For example, the empirical power rate in the large sample (R5,000/F5,000) of the 24-item test with four DIF items in the low nonuniform condition shown in Figure 4.13 was slightly insufficient (below 0.80). Compared to the performance of the empirical power rates for uniform DIF conditions, the empirical power rates for nonuniform DIF conditions for medium and large samples (e.g., R3,000/F3,000, R4,000/F2,000, and R5,000/F5,000) were relatively low, especially from the MCMC estimation method in low DIF magnitude. For example, in the case of the empirical power rates of the 24-item and 46-item test with four DIF items with medium samples of R3,000/F3,000 and R4,000/F2,000 conditions for MCMC, the empirical power rates were all below 0.80 for low nonuniform DIF condition (Figure 4.13), whereas the empirical power rate was around 0.95 for low uniform DIF condition (Figure 4.11).

Table 4.13

Power Results for the M-2PL in the 24- item Test Conditions with 4 Nonuniform DIF Items (.17 DIF)

Sample Size		Balanced							Unbalanced		
		R1000/F1000		R3000/F3000		R5000/F5000			R4000/F2000		
		$\alpha = .05$		$\alpha = .05$		$\alpha = .05$			$\alpha = .05$		
DIF Type	DIF Item	MML[a]	MCMC[a]	MML[a]	MCMC[a]	MML[a]	MCMC[a]	Average	MML[a]	MCMC[a]	Average
Low Non-uniform	21	.25 (.61)	.87 (.36)	1.0 (1.0)	1.0 (.95)	1.0 (1.0)	1.0 (1.0)	.85 (.82)	.97 (.95)	1.0 (.80)	.99 (.88)
	22	.52 (.76)	.86 (.36)	1.0 (1.0)	1.0 (.89)	1.0 (1.0)	1.0 (1.0)	.90 (.84)	1.0 (1.0)	1.0 (.90)	1.0 (.95)
	23	.37 (.78)	.93 (.62)	1.0 (1.0)	1.0 (.97)	1.0 (1.0)	1.0 (1.0)	.88 (.90)	1.0 (1.0)	1.0 (.82)	1.0 (.91)
	24	.16 (.36)	.67 (.13)	.83 (.93)	.78 (.15)	1.0 (1.0)	.92 (.15)	.73 (.45)	.78 (.71)	.77 (.24)	.78 (.48)
Average		.33 (.63)	.83 (.37)	.96 (.98)	.95 (.74)	1.0 (1.0)	.98 (.79)		.94 (.92)	.94 (.69)	
Medium Non-uniform	21	.96 (.99)	1.0 (.90)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.99 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	22	.99 (.99)	1.0 (.89)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.99 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	23	.99 (1.0)	1.0 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	24	.78 (.88)	.91 (.45)	1.0 (1.0)	1.0 (.87)	1.0 (1.0)	1.0 (.85)	.95 (.84)	1.0 (1.0)	1.0 (.83)	.99 (.81)
Average		.93 (.97)	.98 (.81)	1.0 (1.0)	1.0 (.97)	1.0 (1.0)	1.0 (.96)		1.0 (1.0)	1.0 (.96)	

Note. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning.
[a]The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parenthesis indicates the proportion of significant Wald test statistics from the empirical distribution.

Table 4.14

Power Results for the M-2PL in the 24- item Test Conditions with 8 Nonuniform DIF item (.30 DIF)

Sample Size		Balanced							Unbalanced		
		R1000/F1000		R3000/F3000		R5000/F5000			R4000/F2000		
		$\alpha = .05$		$\alpha = .05$		$\alpha = .05$			$\alpha = .05$		
DIF Type	DIF Item	MML[a]	MCMC[a]	MML[a]	MCMC[a]	MML[a]	MCMC[a]	Average	MML[a]	MCMC[a]	Average
Low Non-uniform	17	.35 (.72)	.91 (.63)	.94 (1.0)	1.0 (.85)	1.0 (1.0)	1.0 (1.0)	.87 (.87)	.96 (.95)	.99 (.81)	1.0 (.96)
	18	.32 (.58)	.83 (.22)	.93 (.98)	1.0 (.86)	1.0 (1.0)	1.0 (1.0)	.85 (.77)	.99 (.95)	.99 (.82)	1.0 (.91)
	19	.48 (.81)	.85 (.52)	.99 (1.0)	.99 (.90)	1.0 (1.0)	1.0 (1.0)	.89 (.87)	1.0 (1.0)	1.0 (.76)	1.0 (.94)
	20	.48 (.87)	.91 (.42)	1.0 (1.0)	1.0 (.78)	1.0 (1.0)	1.0 (1.0)	.90 (.85)	1.0 (1.0)	.98 (.91)	1.0 (.91)
	21	.46 (.82)	.95 (.50)	.97 (.99)	1.0 (.91)	1.0 (1.0)	1.0 (1.0)	.90 (.87)	1.0 (.99)	.99 (.84)	1.0 (.92)
	22	.41 (.75)	.91 (.72)	1.0 (1.0)	1.0 (.96)	1.0 (1.0)	1.0 (1.0)	.89 (.91)	1.0 (.99)	.99 (.83)	1.0 (.94)
	23	.14 (.35)	.48 (.04)	.74 (.90)	.76 (.32)	1.0 (1.0)	.96 (.56)	.68 (.53)	.90 (.84)	.83 (.11)	.94 (.61)
	24	.12 (.43)	.61 (.06)	.75 (.92)	.80 (.29)	1.0 (1.0)	.88 (.10)	.69 (.47)	.85 (.83)	.73 (.08)	.90 (.64)
Average		.35 (.67)	.81 (.39)	.92 (.97)	.94 (.73)	1.0 (1.0)	.98 (.83)		.96 (.94)	.94 (.65)	
Medium Non-uniform	17	.94 (.99)	.99 (.97)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.99 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	18	.97 (1.0)	1.0 (.83)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (.97)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	19	1.0 (1.0)	1.0 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	20	1.0 (1.0)	1.0 (.96)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	21	.99 (1.0)	1.0 (.96)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	22	.98 (1.0)	.99 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	23	.72 (.94)	.88 (.27)	1.0 (1.0)	.99 (.79)	1.0 (1.0)	1.0 (.98)	.93 (.83)	1.0 (1.0)	.99 (.58)	1.0 (.91)
	24	.75 (.89)	.84 (.31)	1.0 (1.0)	.99 (.78)	1.0 (1.0)	1.0 (.83)	.93 (.80)	1.0 (1.0)	1.0 (.50)	1.0 (.93)
Average		.92 (.98)	.96 (.78)	1.0 (1.0)	1.0 (.95)	1.0 (1.0)	1.0 (.98)		1.0 (1.0)	1.0 (.89)	

Note. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning.

[a]The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parenthesis indicates the proportion of significant Wald test statistics from the empirical distribution.

Table 4.15

Power Results for the M-2PL in the 46- item Test Conditions with 4 Nonuniform DIF Items (.10 DIF)

Sample Size		Balanced							Unbalanced		
		R1000/F1000		R3000/F3000		R5000/F5000			R4000/F2000		
		$\alpha = .05$		$\alpha = .05$		$\alpha = .05$			$\alpha = .05$		
DIF Type	DIF Item	MML[a]	MCMC[a]	MML[a]	MCMC[a]	MML[a]	MCMC[a]	Average	MML[a]	MCMC[a]	Average
Low Non-uniform	43	.28 (.56)	.94 (.39)	.96 (1.0)	1.0 (.95)	1.0 (1.0)	1.0 (1.0)	.86 (.82)	.99 (.99)	.99 (.92)	.99 (.96)
	44	.52 (.73)	.95 (.48)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.91 (.87)	1.0 (1.0)	.99 (.93)	1.0 (.97)
	45	.45 (.72)	.95 (.45)	.99 (.99)	1.0 (.96)	1.0 (1.0)	1.0 (1.0)	.90 (.85)	.99 (.99)	.99 (.88)	.99 (.94)
	46	.31 (.53)	.70 (.18)	.82 (.88)	.80 (.24)	1.0 (1.0)	.86 (.51)	.75 (.56)	.90 (.89)	.86 (.19)	.88 (.54)
Average		.39 (.64)	.89 (.38)	.94 (.97)	.95 (.79)	1.0 (1.0)	.97 (.88)		.97 (.97)	.96 (.73)	
Medium Non-uniform	43	.89 (.96)	1.0 (.90)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.98 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	44	1.0 (1.0)	1.0 (.98)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	45	.98 (1.0)	1.0 (.93)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	46	.86 (.96)	.95 (.47)	1.0 (1.0)	.99 (.73)	1.0 (1.0)	.99 (.91)	.97 (.85)	1.0 (1.0)	.98 (.61)	.99 (.81)
Average		.93 (.98)	.99 (.82)	1.0 (1.0)	1.0 (.93)	1.0 (1.0)	1.0 (.98)		1.0 (1.0)	1.0 (.90)	

Note. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning.
[a]The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parenthesis indicates the proportion of significant Wald test statistics from the empirical distribution.

Table 4.16

Power Results for the M-2PL in the 46-item Test Conditions with 8 Nonuniform DIF item (.17 DIF)

Sample Size		Balanced							Unbalanced		
		R1000/F1000		R3000/F3000		R5000/F5000			R4000/F2000		
		$\alpha = .05$		$\alpha = .05$		$\alpha = .05$			$\alpha = .05$		
DIF Type	DIF Item	MML[a]	MCMC[a]	MML[a]	MCMC[a]	MML[a]	MCMC[a]	Average	MML[a]	MCMC[a]	Average
Low Non-uniform	39	.35 (.68)	.88 (.52)	.99 (1.0)	.99 (.96)	1.0 (1.0)	1.0 (1.0)	.87 (.86)	1.0 (1.0)	1.0 (.91)	1.0 (.96)
	40	.29 (.62)	.89 (.46)	.99 (.99)	.99 (.95)	1.0 (1.0)	1.0 (.99)	.86 (.67)	1.0 (1.0)	.99 (.82)	1.0 (.91)
	41	.49 (.78)	.91 (.57)	1.0 (1.0)	1.0 (.97)	1.0 (1.0)	1.0 (.99)	.90 (.89)	1.0 (1.0)	1.0 (.88)	1.0 (.94)
	42	.43 (.86)	.94 (.36)	1.0 (1.0)	1.0 (.88)	1.0 (1.0)	1.0 (1.0)	.90 (.85)	1.0 (1.0)	.99 (.81)	1.0 (.91)
	43	.53 (.72)	.99 (.79)	1.0 (1.0)	1.0 (.95)	1.0 (1.0)	1.0 (1.0)	.92 (.91)	1.0 (1.0)	1.0 (.83)	1.0 (.92)
	44	.46 (.85)	.95 (.57)	1.0 (1.0)	1.0 (.98)	1.0 (1.0)	1.0 (1.0)	.90 (.90)	1.0 (1.0)	1.0 (.87)	1.0 (.94)
	45	.27 (.52)	.67 (.29)	.85 (.97)	.83 (.38)	1.0 (1.0)	.88 (.47)	.75 (.61)	1.0 (1.0)	.88 (.22)	.94 (.61)
	46	.26 (.59)	.69 (.07)	.89 (.93)	.91 (.34)	1.0 (1.0)	.86 (.42)	.77 (.56)	1.0 (1.0)	.80 (.28)	.90 (.64)
Average		.39 (.70)	.87 (.45)	.97 (.97)	.95 (.80)	1.0 (1.0)	.97 (.86)		1.0 (1.0)	.96 (.70)	
Medium Non-uniform	39	.96 (1.0)	1.0 (.95)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.99 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	40	.95 (1.0)	1.0 (.95)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	.99 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	41	1.0 (1.0)	1.0 (.95)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (.99)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	42	1.0 (1.0)	1.0 (.90)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (.98)	1.0 (1.0)	1.0 (.99)	1.0 (1.0)
	43	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	44	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
	45	.91 (.95)	.96 (.66)	1.0 (1.0)	1.0 (.77)	1.0 (1.0)	1.0 (.93)	.98 (.89)	1.0 (1.0)	.99 (.81)	1.0 (.91)
	46	.86 (.97)	.90 (.49)	1.0 (1.0)	.99 (.78)	1.0 (1.0)	1.0 (.91)	.96 (.86)	1.0 (1.0)	1.0 (.86)	1.0 (.93)
Average		.96 (.99)	.98 (.86)	1.0 (1.0)	1.0 (.94)	1.0 (1.0)	1.0 (.98)		1.0 (1.0)	.96 (.96)	

Note. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning.

[a]The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parenthesis indicates the proportion of significant Wald test statistics from the empirical distribution.

For the medium DIF magnitude, the empirical power rates of all test conditions for MML and MCMC resulted in sufficient power rates (above 0. 80) regardless of sample size. However, one exception was shown in the empirical power rate of the 24-item test with eight DIF items condition of R1,000/F1,000 from MCMC and was below 0.80, as shown in Figure 4.14.

In general, the empirical power rates of the eight DIF item conditions were slightly lower than the four DIF item conditions in nonuniform low and medium DIF conditions. Most apparent from Figures 4.13 and 4.14 is that MML provided substantially higher power rates than MCMC. Similar to the uniform DIF conditions, the effect of test length was also associated with the higher power rates on two sample size conditions, R1,000/F,1000 and R4,000/F2,000 in low and medium nonuniform DIF conditions for MML and MCMC. However, the overall effect of the number of DIF items on the nonuniform conditions was inconsistent. As also shown in Tables 4.13 through 4.16, the power rates for detecting nonuniform DIF using MML slightly increased as the number of DIF items increased from four to eight. However, the power rates for MCMC slightly decreased from four DIF items to eight DIF items in the 46-item test condition of R5,000/F5,000,whereas the power rates increased in the 24-item test condition of R5,000/F5,000. As expected, the results showed that power increased as the DIF magnitude increased from low to medium for MML and MCMC in nonuniform conditions.

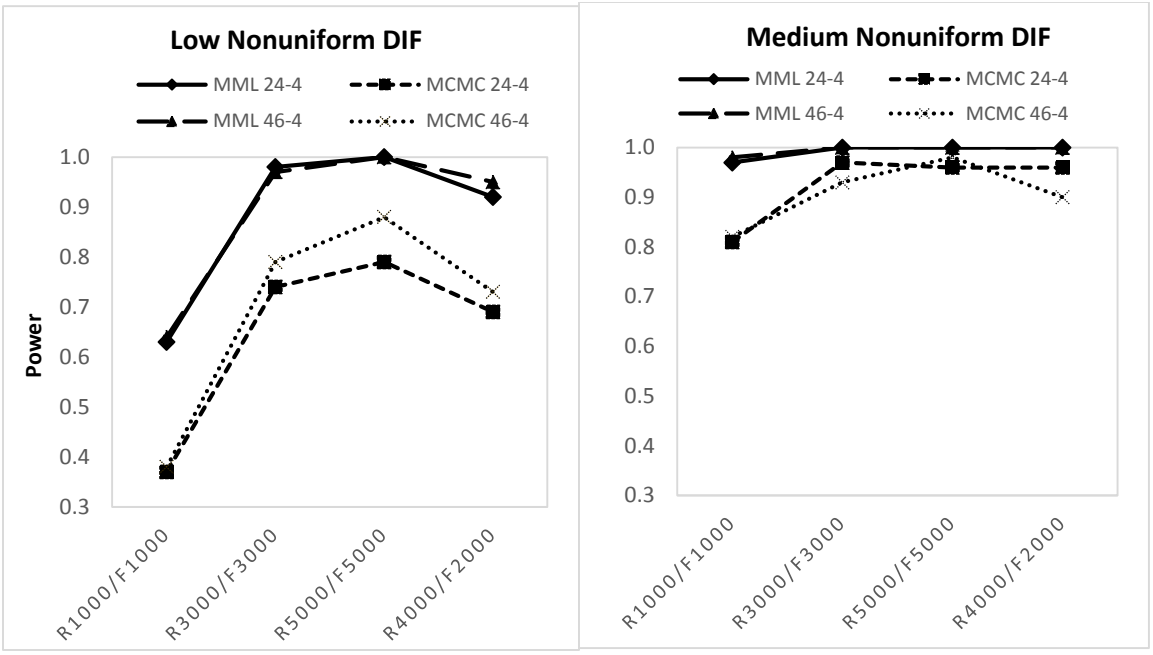

Figure 4.13. Average Power Rates of M-2PL by Test Length, the Number of DIF Items (4 DIF), and Sample Size with Low and Medium Nonuniform DIF Condition

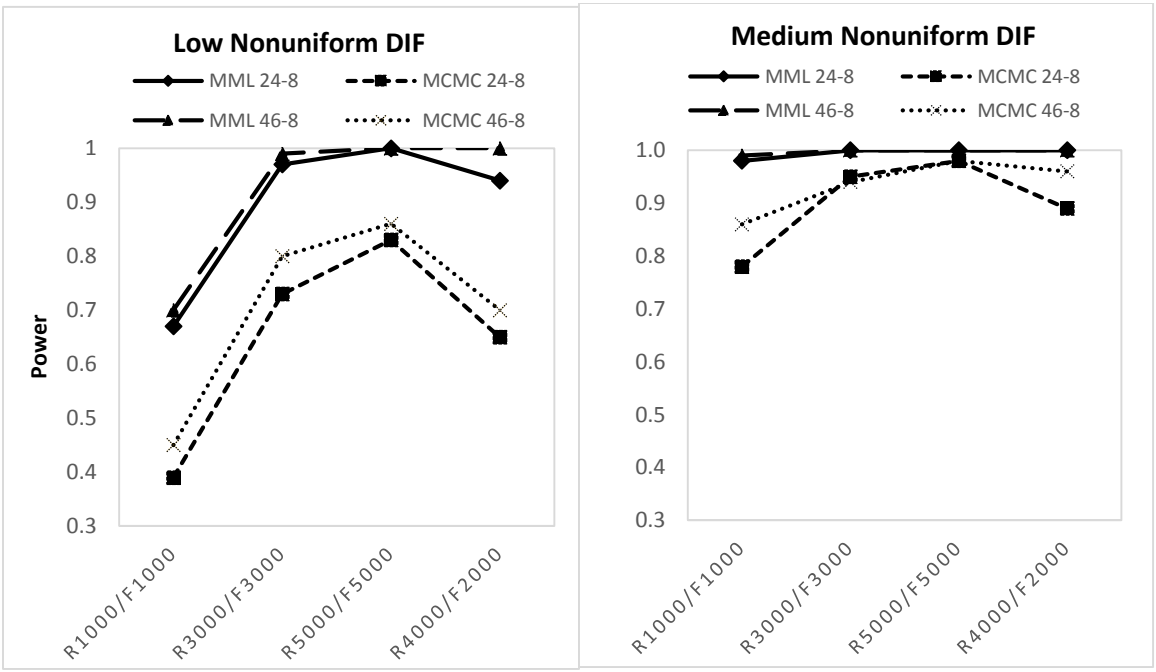

Figure 4.14. Average Power Rates of M-2PL by Test Length, the Number of DIF Items (8 DIF), and Sample Size with Low and Medium Nonuniform DIF Conditions

Regarding the total sample comparison of 6,000, the balanced sample design (R3,000/F3,000) with a (1:1) ratio and the unbalanced sample design (R4,000/F2,000) with (2:1) ratio conditions were compared and are shown in Figures 4.15 and 4.16. In general, higher power rates were observed in the balanced design than in the unbalanced design especially with low nonuniform DIF with MCMC. Among all nonuniform DIF conditions, the 24-item test with eight DIF item condition in the unbalanced sample design (R4,000/F2,000) using MCMC had the lowest power rates (around 0.65). Similar to the low uniform condition, the 46-item test with four DIF item condition in the unbalanced sample design (R4,000/F2,000) using MCMC had the lowest empirical power rates in low nonuniform DIF, but was very powerful (over 0.85). Unlike the medium uniform DIF condition, the lowest empirical power rates of medium nonuniform DIF were found in the R4,000/F2,000 sample. The larger DIF magnitude (medium DIF) resulted in higher power rates across all uniform and nonuniform DIF conditions regardless of the sample size ratio.

Overall, when the sample size ratio was balanced (1:1) or unbalanced (2:1), excellent power rates (above 0.95) were acquired from MML across nonuniform DIF conditions. However, when the sample size was balanced (1:1), slightly insufficient power rates (below 0.80) were obtained from MCMC, and when the sample size was unbalanced (2:1), insufficient power rates (below 0.70) were obtained from MCMC in nonuniform DIF conditions. Considering these results of nonuniform DIF conditions in terms of the total sample size effect, similar to the uniform DIF conditions, MML seems to be a better choice for detecting DIF for balanced and unbalanced sample size DIF situations.

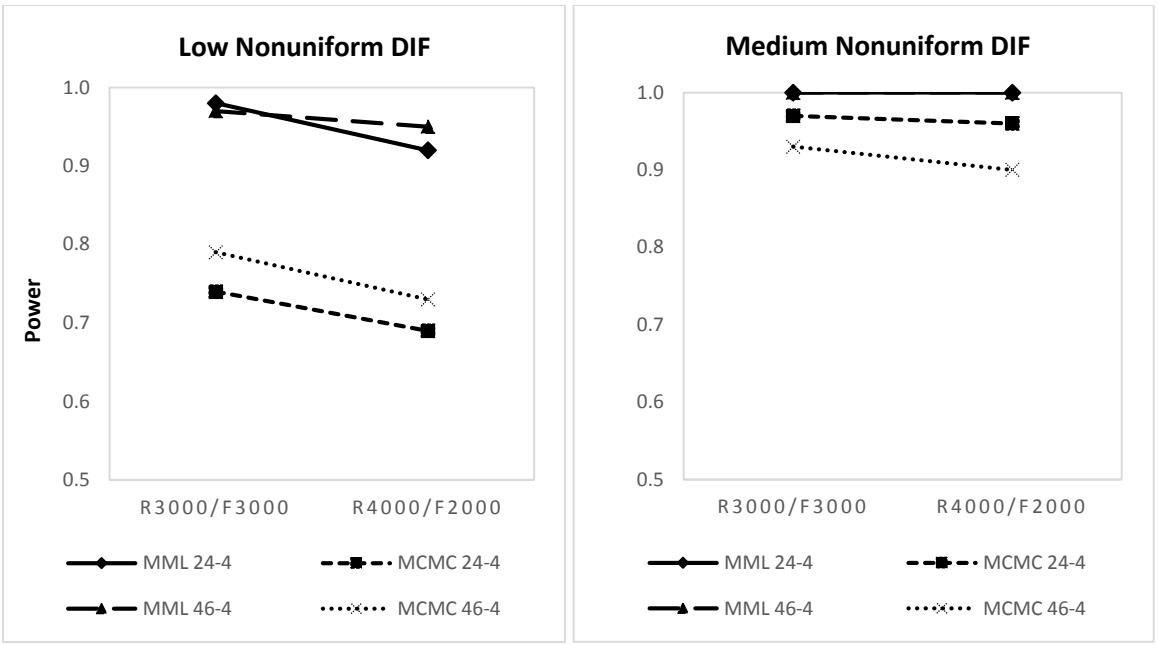

Figure 4.15. Average Power Rates of Low and Medium Nonuniform DIF Conditions with 4 DIF Items in M-2PL

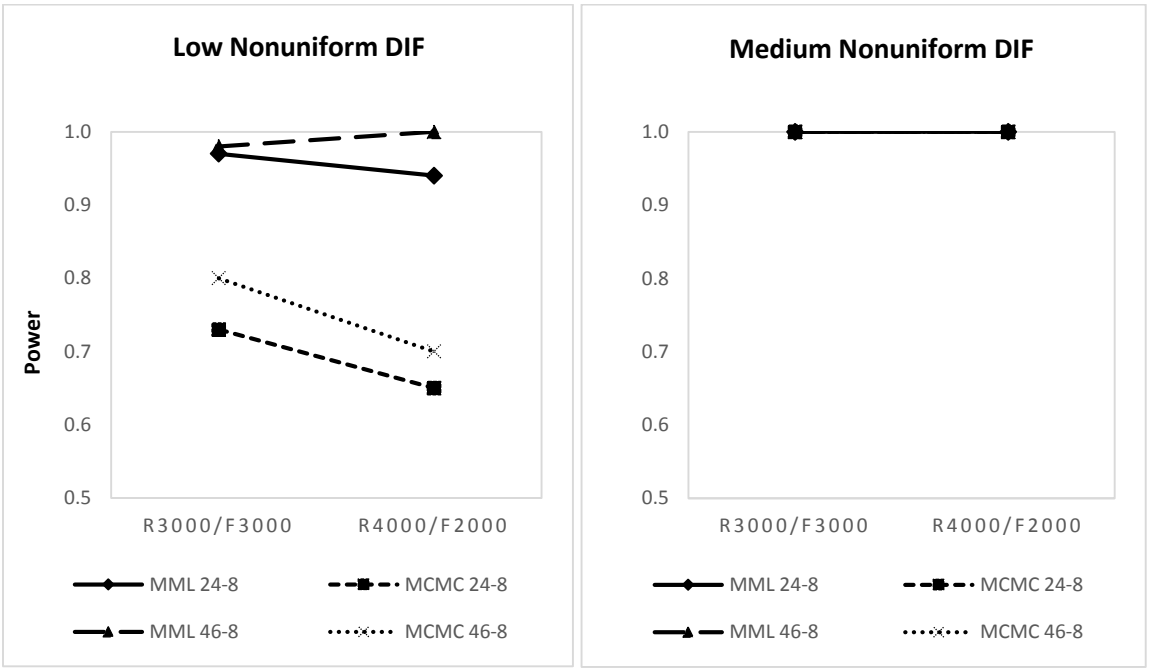

Figure 4.16. Average Power Rates of Low and Medium Nonuniform DIF Conditions with 8 DIF Items in M-2PL

4.4.2 The M-3PL Results.

Uniform DIF Results. *Theoretical vs. empirical power rates.* Tables 4.17 through 4.18 list the power results for uniform DIF conditions for M-3PL determined with theoretical and the empirical power rates calculated using χ^2 (df = 4) for the DIF magnitude, test length, and number of DIF items, respectively. Similar to the performance of uniform DIF detection for M-2PL, the average empirical power rates of Lord's Wald test for MCMC were lower than that for MML. The theoretical and empirical power rates of the 24-item test with four DIF items in low and medium uniform DIF conditions shown in Table 4.17 were substantially different for two estimation methods, MML and MCMC. Unlike the M-2PL, the theoretical power rates for MML were higher than the empirical power rates except for the 24-item test with four DIF items in low uniform DIF conditions. The theoretical power rates for MCMC in the uniform DIF conditions were higher than the empirical power rates, especially with the low DIF magnitude. For example, in the case of the 24-item test with four DIF items in the R3,000/F3,000 condition, as shown in Table 4.17, the average theoretical power rates vs. the average empirical power rates of low DIF from MCMC were 0.94 vs. 0.62, whereas the average power rates of medium DIF were 0.94 vs. 0.92. Unlike the low DIF conditions, the difference between the theoretical and empirical power rates was almost zero under the medium DIF conditions from MML and MCMC. For both methods, the average empirical power rates of the medium DIF magnitude conditions were higher than those of the low DIF magnitude conditions irrespective of all other factors.

Table 4.17

Power Results for the M-3PL in the 24 & 46-item Test Conditions with 4 Uniform DIF Items

Sample Size		R3000/F3000				
		$\alpha = .05$				
DIF Type	DIF Item	MML[a]	MCMC[a]	DIF Item	MML[a]	MCMC[a]
Low Uniform	21	.80 (.95)	.88 (.29)	43	.96 (.96)	.93 (.78)
	22	.79 (.98)	.94 (.82)	44	.96 (.80)	.93 (.91)
	23	.89 (.97)	.96 (.44)	45	.98 (.97)	.91 (.58)
	24	1.0 (1.0)	.97 (.92)	46	.86 (.85)	.93 (.73)
Average		.87 (.98)	.94 (.62)		.94 (.90)	.93 (.75)
Medium Uniform	21	1.0 (1.0)	.93 (.93)	43	1.0 (1.0)	.87 (.86)
	22	1.0 (1.0)	.97 (.97)	44	1.0 (1.0)	.92 (.92)
	23	1.0 (1.0)	.91 (.87)	45	1.0 (1.0)	.96 (.96)
	24	1.0 (1.0)	.93 (.92)	46	1.0 (1.0)	.96 (.96)
Average		1.0 (1.0)	.94 (.92)		1.0 (1.0)	.93 (.93)

[a] The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parentheses indicates the proportion of significant Wald test statistics from the empirical distribution.

Table 4.18

Power Results for the M-3PL in the 24 & 46-item Test Conditions with 8 Uniform DIF Items

Sample Size		R3000/F3000				
		$\alpha = .05$				
DIF Type	DIF Item	MML[a]	MCMC[a]	DIF Item	MML[a]	MCMC[a]
Low Uniform	17	.91 (.95)	.93 (.87)	39	.92 (.98)	.93 (.81)
	18	.92 (.92)	.93 (.54)	40	.92 (1.0)	.96 (.90)
	19	.93 (.90)	.93 (.51)	41	.95 (.99)	.93 (.86)
	20	.95 (.97)	.98 (.77)	42	.96 (.36)	.94 (.77)
	21	.94 (.87)	.93 (.44)	43	.97 (.90)	.90 (.76)
	22	.92 (.94)	.92 (.86)	44	.99 (.81)	.93 (.62)

	23	.88 (.74)	.95 (.53)	45	.84 (.25)	.97 (.77)
	24	.89 (.42)	.98 (.50)	46	.93 (.78)	.94 (.69)
Average		.92 (.84)	.94 (.63)		.94 (.76)	.94 (.77)
Medium Uniform	17	1.0 (1.0)	.93 (.93)	39	1.0 (1.0)	.97 (.97)
	18	1.0 (1.0)	.92 (.92)	40	1.0 (1.0)	.90 (.89)
	19	1.0 (1.0)	.94 (.94)	41	1.0 (1.0)	.93 (.93)
	20	1.0 (1.0)	.96 (.96)	42	1.0 (1.0)	.93 (.93)
	21	1.0 (1.0)	.95 (.95)	43	1.0 (1.0)	.94 (.94)
	22	1.0 (1.0)	.96 (.96)	44	1.0 (1.0)	.95 (.95)
	23	1.0 (1.0)	.94 (.94)	45	1.0 (1.0)	.97 (.97)
	24	1.0 (1.0)	.97 (.96)	46	1.0 (1.0)	.95 (.95)
Average		1.0 (1.0)	.95 (.95)		1.0 (1.0)	.94 (.94)

[a] The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parentheses indicates the proportion of significant Wald test statistics from the empirical distribution.

Nonuniform DIF Results. *Theoretical vs. empirical power rates.* Tables 4.19 through 4.20 list the power results for nonuniform DIF conditions from M-3PL determined with theoretical and the empirical power rates calculated using χ^2 (df = 4) for the DIF magnitude, test length, and number of DIF items, respectively. Similar to the performance of uniform DIF detection from M-3PL, the average empirical power rates of Lord's Wald test from MCMC were lower than that from MML. The theoretical and empirical power rates of the 24-item test with four DIF items in low and medium nonuniform DIF conditions shown in Table 4.19 was substantially different for the two estimation methods, MML and MCMC. The theoretical power rates of MCMC in the nonuniform DIF conditions were notably higher than the empirical power rates, especially with the low DIF magnitude, as shown in Tables 4.19 and 4.20. Unlike the M-2PL results, the theoretical power rates of MML tended to be higher than the empirical power rates under the nonuniform DIF conditions except for the 24-item test with four

DIF items low condition, as shown in Table 4.19. For the medium nonuniform DIF conditions, the difference between the theoretical power rates and the empirical power rates were apparent from MCMC, whereas the difference from MML was almost zero, the same as the medium uniform conditions. For example, in the medium nonuniform DIF conditions, the average theoretical power rates vs. the average empirical power rates of R3,000/F3,000 from MML for the 24-item test with four DIF items, the 24-item test with eight DIF items, the 46-item test with four DIF items, and the 46-item test with eight DIF items were 1.0 vs. 1.0, 1.0 vs. 0.99, 1.0 vs. 1.0, and 1.0 vs. 1.0, respectively. The corresponding average theoretical power rates vs. the average empirical power rates of from MCMC were 0.93 vs. 0.67, 0.92 vs. 0.69, 0.93 vs. 0.80, and 0.92 vs. 0.65, respectively.

Overall, the average empirical power rates from M-3PL when MML was used were higher than the average empirical power rates using MCMC for detecting nonuniform DIF. The difference between the theoretical power rates and the empirical power rates of MML and MCMC became smaller as DIF magnitude increased from low to medium. Similar to the uniform DIF condition, the degrees of increase and decrease between the theoretical power rates and the empirical power rates of MML and MCMC in the medium DIF conditions were much smaller than that of the low DIF conditions.

Table 4.19

Power Results for the M-3PL in the 24 & 46-item Test Conditions with 4 Nonuniform DIF Items

Sample Size		R3000/F3000				
		$\alpha = .05$				
DIF Type	DIF Item	MML[a]	MCMC[a]	DIF Item	MML[a]	MCMC[a]
Low Non-uniform	21	.56 (.68)	.85 (.23)	43	.84 (.84)	.87 (.47)
	22	.80 (.97)	.94 (.43)	44	.97 (.91)	.86 (.61)
	23	.72 (.82)	.91 (.26)	45	.95 (.95)	.86 (.52)
	24	.25 (.33)	.61 (.12)	46	.71 (.71)	.65 (.17)
Average		.58 (.70)	.83 (.26)		.87 (.85)	.81 (.44)
Medium Non-uniform	21	1.0 (1.0)	.95 (.74)	43	1.0 (1.0)	.96 (.94)
	22	1.0 (1.0)	.97 (.94)	44	1.0 (1.0)	.94 (.93)
	23	1.0 (1.0)	.98 (.83)	45	1.0 (1.0)	.92 (.89)
	24	1.0 (1.0)	.83 (.18)	46	1.0 (1.0)	.91 (.43)
Average		1.0 (1.0)	.93 (.67)		1.0 (1.0)	.93 (.80)

[a] The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parenthesis indicates the proportion of significant Wald test statistics from the empirical distribution.

Table 4.20

Power Results for the M-3PL in the 24 & 46-item Test Conditions with 8 Nonuniform DIF Items

Sample Size		R3000/F3000				
		$\alpha = .05$				
DIF Type	DIF Item	MML[a]	MCMC[a]	DIF Item	MML[a]	MCMC[a]
Low Non-uniform	17	.63 (.68)	.93 (.70)	39	.83 (.87)	.87 (.52)
	18	.69 (.70)	.93 (.35)	40	.81 (.84)	.83 (.58)
	19	.83 (.78)	.83 (.25)	41	.92 (.96)	.87 (.49)
	20	.79 (.85)	.91 (.38)	42	.81 (.78)	.91 (.48)
	21	.82 (.77)	.94 (.39)	43	.90 (.88)	.95 (.66)

	22	.77 (.79)	.90 (.63)	44	.93 (.90)	.91 (.43)
	23	.37 (.27)	.61 (.06)	45	.71 (.51)	.52 (.09)
	24	.38 (.22)	.62 (.04)	46	.70 (.65)	.58 (.09)
Average		.66 (.63)	.83 (.35)		.83 (.80)	.81 (.42)
Medium Non-uniform	17	1.0 (1.0)	.96 (.96)	39	1.0 (1.0)	.90 (.90)
	18	1.0 (1.0)	.95 (.83)	40	1.0 (1.0)	.92 (.91)
	19	1.0 (1.0)	.97 (.79)	41	1.0 (1.0)	.92 (.91)
	20	1.0 (1.0)	.92 (.87)	42	1.0 (1.0)	.95 (.93)
	21	1.0 (1.0)	.96 (.90)	43	1.0 (1.0)	.93 (.93)
	22	1.0 (1.0)	.95 (.95)	44	1.0 (1.0)	.90 (.85)
	23	.99 (.97)	.78 (.12)	45	1.0 (1.0)	.92 (.32)
	24	1.0 (.96)	.86 (.08)	46	1.0 (1.0)	.92 (.41)
Average		1.0 (.99)	.92 (.69)		1.0 (1.0)	.92 (.65)

[a] The first value in each column represents the proportion of significant Wald test statistics from the theoretical distribution, whereas the second value in the parentheses indicates the proportion of significant Wald test statistics from the empirical distribution.

4.4.3 A Comparison of the M-2PL and M-3PL Results. Figures 4.17 through 4.20 illustrate the comparison results for M-2PL and M-3PL. Figures 4.17 and 4.18 show the comparison results for uniform DIF conditions, and Figures 4.19 and 4.20 show the comparison results for nonuniform DIF conditions separately for the medium sample (R3,000/F3,000). The power rates of MML were higher than those of MCMC from M-2PL in uniform DIF conditions. The performance of the average power rates using MCMC worsened from the M-2PL (over 0.90) to M-3PL (below 0.80) models, as shown in Figures 4.17 and 4.18. In the case of medium DIF magnitude, this pattern was less noticeable. The average power rates of MCMC slightly decreased from M-2PL (over 0.90) to M-3PL (over 0.90), especially for the 24-item test with four DIF items condition in the medium uniform DIF condition. The higher power rates were found for the short (24 items) test than for the long (46 items) test using MML, whereas the power rates of

the longer test (46 items) were higher than those of the short (24 items) test using MCMC in low uniform DIF conditions from M-3PL.

This decreasing tendency was even more apparent from MCMC in nonuniform DIF conditions than in uniform conditions. For example, the average power rates of the 24-item test with four DIF items in the uniform condition from MCMC decreased substantially from the M-2PL (over 0.70) to M-3PL (below 0.30) model. The test length factor influenced the M-3PL power rates; the power rates of short test (24-item test) with four DIF items using MCMC tended to be the lowest in all nonuniform DIF conditions from M-3PL. Unlike the uniform results, the power rates of the longer test (46 items) were higher than those of the short (24 items) test using MML and MCMC from M-3PL in the low and medium nonuniform DIF conditions. In general, across all simulation conditions, the power rates for M-2PL were much higher than those for the M-3PL model with the exception of the 24-item test with four DIF items using MML low DIF condition. The power rate of MML was much higher than those of MCMC across all DIF conditions with the exception of the 46-item test with eight DIF items using MCMC in the low uniform condition. No pattern was associated with the number of DIF items.

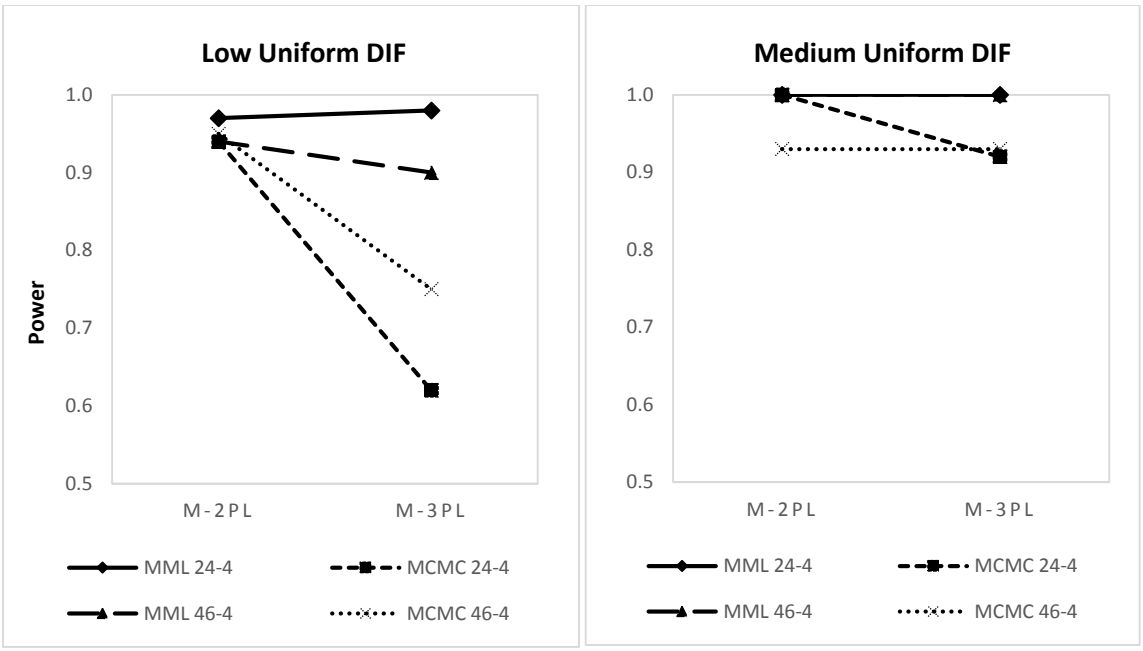

Figure 4.17. Comparison of Power Results of Uniform DIF with 4 Items Conditions between M-2PL and M-3PL (R3000/F3000)

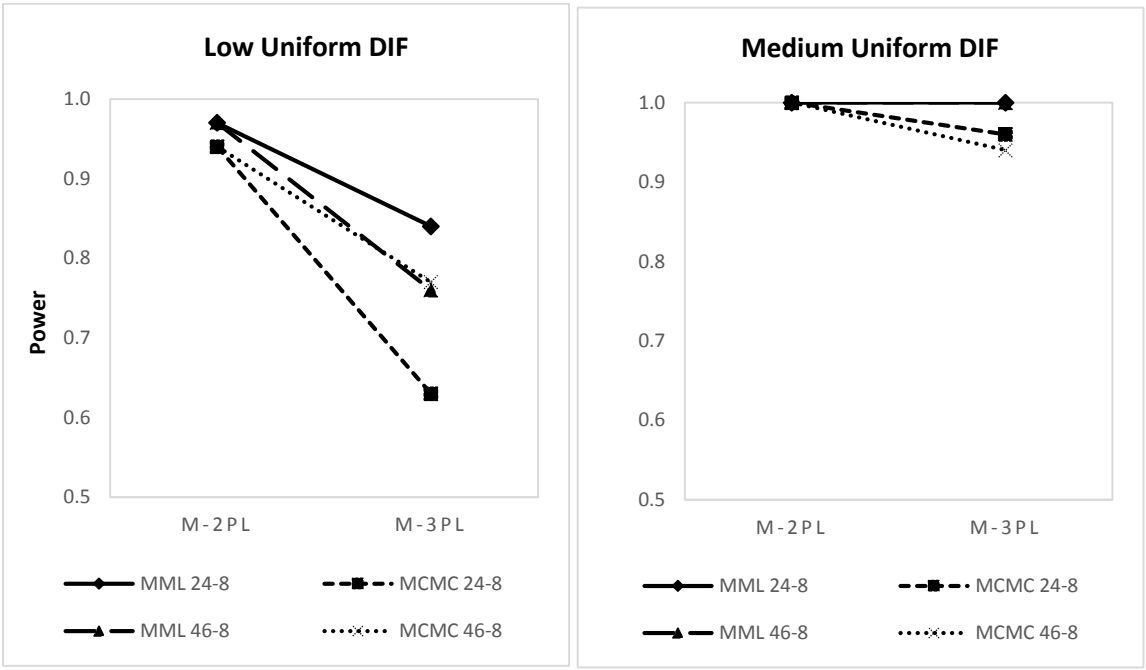

Figure 4.18. Comparison of Power Results of Uniform DIF with 8 Items Conditions between M-2PL and M-3PL (R3000/F3000)

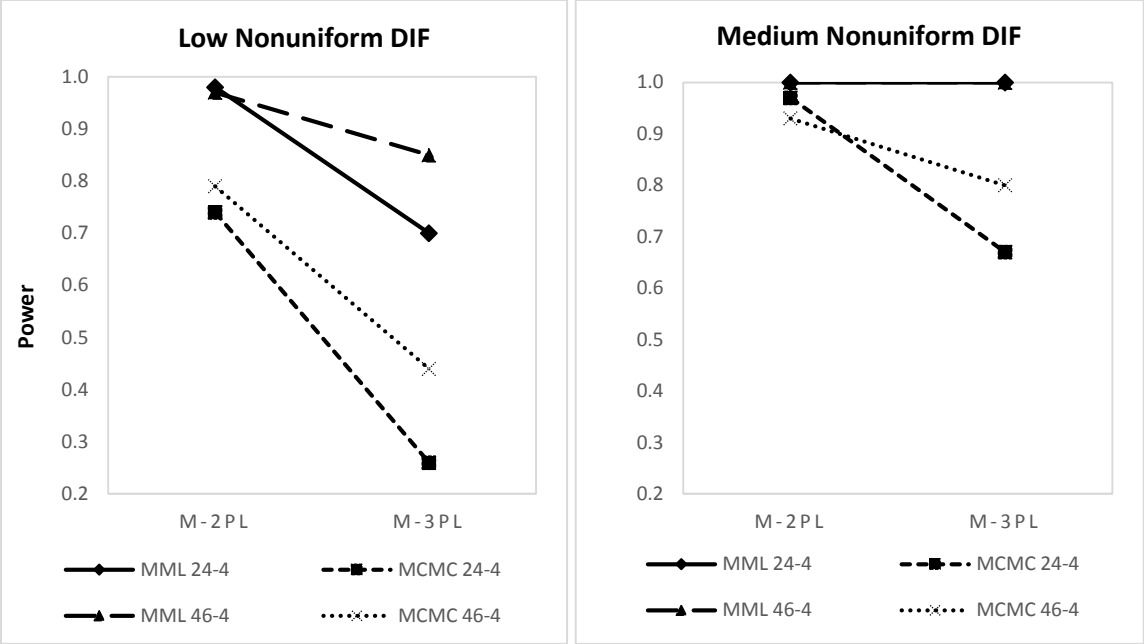

Figure 4.19. Comparison of Power Results of Nonuniform DIF with 4 Items Conditions between M-2PL and M-3PL (R3000/F3000)

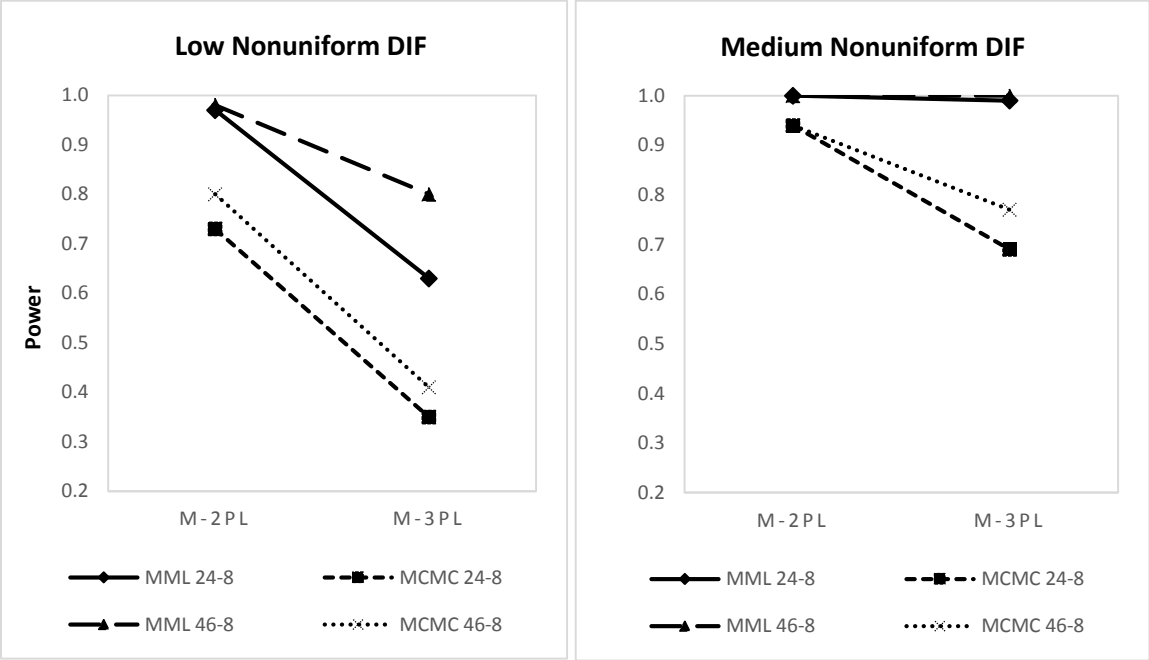

Figure 4.20. Comparison of Power Results of Nonuniform DIF with 8 Items Conditions between M-2PL and M-3PL (R3000/F3000)

4.5 Estimation Time

The sample size and the test length directly affected the running time of each computer program. The MCMC results were possibly affected by the number of iterations and burn-in for all items in all conditions (Wollack et al., 2012). For example, MML required approximately 3 hours to complete in the R1,000/F1,000 sample condition with the 24-item test, whereas MCMC required approximately 12 hours to complete 10,000 iterations with 5,000 burn-ins on a computer with 1.80 GHz processor speed. The MML required approximately 6 hours to complete the R3,000/F3,000 sample with the 24-item test, whereas the MCMC required approximately 38 hours to complete. The M-3PL required much longer time. The MML required approximately 13 hours to complete the R3,000/F3,000 sample with the 24-item test, whereas the MCMC required approximately 80 hours to complete. For both estimation methods, the computational times increased linearly as the sample size and test length increased.

CHAPTER 5

REAL DATA ANALYSIS

The real data used for this study consisted of a college-level English placement test in the University of Wisconsin (UW) system. The placement test helps enrolling students select courses and informs English college instructors about student ability levels. Only a part of the test (i.e., the 31-item English usage section) was analyzed in this study. The item specifications are shown in Table 5.1 (see the second column, "Type," in the table). Each item contains a sentence where a characteristic English usage error has been introduced. Each student must select one of the underlined options from the part of the sentence that should be corrected or the "No error" category if there is no error.

A previous dimensionality analysis of this test by Bolt, Cohen, and Wollack (2001) used a mixture nominal response model, and Bolt et al. (2001) found that two dimensions (abilities) were present in solving the English usage questions. Later, two dimensions of the test were validated again by Bolt and Lall (2003) through a multidimensional latent trait model and a M-2PL model. Based on these two studies, one dimension is related to detecting punctuation-related errors, and another dimension is associated with word usage-related errors. The misplacement of a comma or the failure to use necessary punctuation to add clarity to a sentence are examples of punctuation-related errors. The use of an incorrect verb tense or pronoun are examples of word usage–related errors. Although the item type is specifically defined in terms of the type of error introduced in the sentence (see Table 5.1), the ability to detect both types of errors is likely applied when the test taker solves any one item in the test. Students will decide one part of the

sentence containing the error while also evaluating the other underlined parts as correctly specified.

In this study, DIF analyses using Lord's Wald test were conducted for gender groups with MML and MCMC. The dataset used in this analysis included a sample of 2,800 examinees; 1,400 male and 1,400 female examinees were randomly selected based on the students' self-reported gender information. Using the total sample, the M-2PL model fitted the data with the total sample of 2,800 examinees using flexMIRT for validating dimensional structure found by Bolt et al. (2001) and Bolt and Lall (2003). The item parameter estimates under the M-2PL are shown in Table 5.1. To resolve metric indeterminacy, the discrimination parameter for the first item on the second dimension was fixed to zero (i.e., $a_{12} = 0$). The same constraints used in the simulation study were applied.

A clear association between dimensions and item type is shown in Table 5.1. For example, items 1, 3, 4, 9, 13, 18, 19, 21, 22, 23, 25, 26, 30, and 31 were predominantly loaded on the first dimension. These items mostly demonstrate punctuation clarity, run-on, comma splice, and comparison and are more likely to be selected as the correct option when the error is punctuation related. At the same time, items 5, 7, 10, 11, 20, 24, 29, 6, and 16 were predominantly loaded on the second dimension. These items mostly demonstrate verb form, subject-verb agreement, and subordination and are more likely to be selected as the correct option when the error is usage related.

Table 5.1

Item Specification and M-2PL Item Parameter Estimates for English Usage Data

The original data				
Item	Type	a_1	a_2	d[a]
1	pn clarity	0.98	0.00	1.08
2	comparison	0.74	0.79	2.93
3	tense	0.83	0.56	1.44
4	run-on	1.56	-0.08	0.50
5	verb form	0.57	0.82	1.88
6	diction/idiom	0.46	0.45	-0.50
7	s-v agree	0.46	0.87	-1.65
8	subordination	0.85	1.03	1.46
9	pn clarity	1.29	-0.18	-0.38
10	diction	0.54	0.91	-0.25
11	parallelism	0.73	1.00	-1.04
12	s-v agree	0.97	1.54	-3.17
13	comma splice	1.14	0.10	-0.04
14	pronoun agr	0.70	0.17	0.03
15	verb tense	1.11	0.83	2.24
16	adv/adj	0.68	0.69	1.05
17	pn clarity	1.38	-0.27	1.11
18	comparison	0.86	0.67	-2.46
19	diction	0.51	0.39	1.06
20	s-v agree	0.77	1.01	-1.05
21	parallelism	0.99	0.82	2.27
22	pn clarity	1.81	-0.21	0.62
23	adv/adj	0.92	0.77	-2.56
24	verb form	0.55	0.62	-0.32
25	frag	0.74	-0.32	0.28
26	tense	0.63	0.48	0.10
27	pronoun agr	0.74	0.92	-2.03
28	pn clarity	1.14	0.00	0.29
29	s-v agree	0.39	0.72	-1.98
30	run-on	1.82	-0.27	0.91
31	verb form	0.75	0.51	-0.54

[a] flexMIRT uses the M-2PL model formula with the positive sign for the intercept, whereas BMIRT uses the negative sign. Therefore, the negative sign is added to the intercept parameter estimates from flexMIRT.

To establish the anchor set of items that did not contain DIF, a "DIF sweep" analysis was conducted in flexMIRT for the MML estimation method. A previous study by Woods et al. (2013) suggested that the DIF sweep procedure may be the best method for establishing anchors for conducting a DIF analysis. The DIF sweep represents a "TestAll" procedure in which all items are tested in turn. Each item is treated as a candidate item while the other items are treated as anchor items. In the DIF sweep, all items are primarily constrained to be equal across groups to obtain conditional population distribution estimates. Then each item is freed one by one and finally tested for DIF using Lord's Wald test (Langer, 2008). From the results of the DIF sweep analysis, the anchor items are obtained. All items contained in the anchor set should have a *p* value that is larger than the nominal alpha level of 0.05. Eight items (items 2, 8, 12, 14, 15, 17, 27, and 28) were excluded from the anchor set because their *p* value was smaller than the critical value of 0.05. These eight items were tested for DIF with the MML and MCMC estimation methods.

The items were rearranged (reordered) to form three clusters as shown in Table 5.2 for the convenience of analysis, so that first two clusters were placed before the last eight studied items. Items in the first two clusters (items 1 through 23) were used as anchor items. The three clusters were as follows: (a) The first 14 items (i.e., items 1 through 14) were selected for the predominant first dimension (i.e., punctuation-related errors), (b) the next eight items (i.e., items 15 through 23) were selected for the predominant second dimension (i.e., word-usage related errors) and (c) the last eight items (i.e., items 24 through 31) were selected as the studied items. These last eight items

were identified using the DIF sweep procedure. For the next step, DIF analysis, men were designated as the reference group (R) and women as the focal group (F).

Table 5.2

Item Rearrangement and Item Specification of the English Usage Items

The rearranged Item Number	The original Item Number	Type
1	1	pn clarity
2	3	tense
3	4	run-on
4	9	pn clarity
5	13	comma splice
6	18	comparison
7	19	diction
8	21	parallelism
9	22	pn clarity
10	23	frag adv/adj
11	25	frag
12	26	tense
13	30	run-on
14	31	verb form
15	5	verb form
16	7	s-v agree
17	10	diction
18	11	parallelism
19	20	s-v agree
20	24	verb form
21	29	s-v agree
22	6	diction/idiom
23	16	adv/adj
24	2	comparison
25	8	subordination
26	12	s-v agree
27	14	pronoun agr
28	15	verb tense
29	17	pn clarity
30	27	pronoun agr
31	28	pn clarity

The DIF detection results were affected by the accuracy of the parameter recovery; thus, parameter estimates and standard errors (SEs) are examined first. The item parameters of the M-2PL model were estimated using MML and MCMC for the male and female groups. Table 5.3 reports the parameter estimates and corresponding standard errors for the anchor items using MML and MCMC separately. For MML, the SEs of the discrimination parameter estimates ranged from 0.05 to 0.12 and from 0.03 to 0.10 for a_1 and a_2, respectively. The SEs of the intercept parameter estimates ranged from 0.04 to 0.09. For MCMC, the SEs of the discrimination parameter estimates ranged from 0.00 to 0.29 and from 0.08 to 0.25 for a_1 and a_2, respectively. The SEs of the intercept parameter estimates ranged from 0.06 to 0.16.

Overall, the SEs of the discrimination and intercept parameter estimates for MML were smaller than those for MCMC, which is not consistent with the simulation recovery results. The range of the discrimination and intercept parameter estimates for the real dataset was narrower than the range of the item parameters used in the simulation study. Tables 5.4 through 5.5 show the parameter estimates and corresponding standard errors (SEs) of the eight DIF items for the female and male groups using MML and MCMC separately. Table 5.4 reports that the SEs of the discrimination parameter estimates for the female group in MML ranged from 0.07 to 0.14 and from 0.06 to 0.15 for a_1 and a_2, respectively. The SEs of the intercept parameter (*d*) estimates ranged from 0.06 to 0.18. The male group showed similar ranges but with slightly higher values than the female group. The SEs of the a_1 parameter for the female group had slightly lower SEs than the male group did with the exception of Item 30. This pattern is consistent with the SEs of the a_2 parameter for the female group across all items with exception of items 28 and 30.

The SEs of the c parameter for the female group had slightly lower SEs than the male group with exceptions of items 26, 28, and 30.

Table 5.5 lists the SEs of the discrimination and intercept parameter estimates for the female and male groups in MCMC. The SEs of the parameter estimates for the female group ranged from 0.08 to 0.30 and from 0.11 to 0.22 for a_1 and a_2, respectively. The parameter estimates for *d* ranged from 0.08 to 0.25 using MCMC. For the male group for the MCMC estimation method, the SEs of the discrimination parameter estimates ranged from 0.12 to 0.25 and from 0.13 to 0.44 for a_1 and a_2, respectively. The SEs of the intercept parameter estimates (*d*) ranged from 0.07 to 0.50 for the male group in MCMC. Similar to MML, the SEs comparison of MCMC between gender groups showed that the SEs of the female group were lower than those of the male group across all item parameters. Using MCMC, the SEs of the item parameter estimates for a_1, a_2, and *d* of the male group were larger than those of the female group across all items with the exception of items 24, 25, and 28 for a_1 parameter and items 27 and 28 for the *d* parameter.

In Tables 5.4 and 5.5, items with large $\hat{a}_1$s (e.g., items 29 and 31) are associated with word punctuation-related errors, and the differences in $\hat{a}_1$s between the female and male groups tend to be more apparent than those in $\hat{a}_2$s. Whereas items with large $\hat{a}_2$s (e.g., items 25 and 30) are associated with usage-related errors, and the differences in $\hat{a}_2$s between the two groups appear to be larger than those in $\hat{a}_1$s . For the items with similar $\hat{a}_1$ and $\hat{a}_2$ (e.g., items 27 and 28), no such clear pattern of differences in $\hat{a}_1$s or $\hat{a}_2$s is present. Instead, large differences are observed in the intercept parameters ($\hat{d}$). Similar

to the anchor items, the standard errors for the a_1, a_2, and d estimates from MML are slightly smaller than those from MCMC. Again, these results are inconsistent with the results of the simulation study, which may be attributed in part to the ranges of the item parameter estimates.

Table 5.3

Parameter Estimates for Anchor Items in the M-2PL for DIF Detection: English Usage Data

Item	Type	MML						MCMC					
		$\hat{a}_1$	SE	$\hat{a}_2$	SE	$\hat{d}$ [a]	SE	$\hat{a}_1$	SE	$\hat{a}_2$	SE	$\hat{d}$	SE
1	pn clarity	1.06	0.07	0.00	-	1.25	0.06	1.14	0.21	0.00	-	-1.34	0.06
2	tense	0.88	0.07	0.98	0.08	1.53	0.06	0.63	0.13	0.89	0.17	-1.67	0.09
3	run-on	1.74	0.10	0.67	0.06	0.74	0.06	1.69	0.24	0.55	0.19	-0.94	0.12
4	pn clarity	1.46	0.09	0.49	0.05	-0.18	0.05	1.40	0.14	0.35	0.18	0.01	0.09
5	comma splice	1.26	0.07	0.66	0.06	0.12	0.05	1.18	0.16	0.59	0.15	-0.30	0.09
6	comparison	0.90	0.08	1.03	0.08	-2.43	0.08	0.71	0.18	0.91	0.10	2.16	0.11
7	diction	0.62	0.05	0.72	0.06	1.12	0.05	0.40	0.10	0.60	0.14	-1.22	0.07
8	parallelism	0.99	0.09	1.28	0.10	2.37	0.08	0.74	0.16	1.19	0.16	-2.54	0.16
9	pn clarity	2.00	0.11	0.66	0.06	0.91	0.07	2.00	0.26	0.50	0.25	-1.13	0.12
10	adv/adj	0.90	0.08	1.15	0.09	-2.53	0.09	0.64	0.10	1.11	0.17	2.28	0.13
11	frag	0.91	0.06	0.32	0.03	0.40	0.05	0.84	0.10	0.12	0.10	-0.48	0.06
12	tense	0.68	0.05	0.81	0.06	0.13	0.04	0.49	0.09	0.75	0.14	-0.27	0.07
13	run-on	2.03	0.12	0.63	0.06	1.21	0.07	1.95	0.29	0.50	0.24	-1.40	0.14
14	verb form	0.78	0.06	0.88	0.06	-0.48	0.05	0.62	0.13	0.82	0.14	0.32	0.09
15	verb form	0.64	0.06	1.16	0.09	1.93	0.07	0.00	0.00	1.13	0.25	-2.06	0.10
16	s-v agree	0.52	0.05	1.10	0.08	-1.71	0.06	0.23	0.15	1.02	0.09	1.50	0.07
17	diction	0.54	0.05	1.18	0.08	-0.27	0.05	0.27	0.14	1.09	0.15	0.07	0.09
18	parallelism	0.69	0.06	1.36	0.09	-1.06	0.06	0.46	0.17	1.27	0.16	0.81	0.12
19	s-v agree	0.73	0.06	1.32	0.08	-1.04	0.06	0.45	0.14	1.30	0.13	0.80	0.09
20	verb form	0.60	0.05	0.92	0.06	-0.30	0.05	0.42	0.11	0.81	0.10	0.14	0.08
21	s-v agree	0.52	0.05	0.93	0.08	-2.06	0.07	0.21	0.10	0.82	0.08	1.82	0.10
22	diction/idiom	0.54	0.05	0.72	0.06	-0.49	0.04	0.35	0.10	0.60	0.08	0.35	0.06
23	adv/adj	0.72	0.06	1.02	0.07	1.10	0.05	0.48	0.13	0.92	0.12	-1.23	0.09

Note. Dashes indicate that there are no standard errors for these estimates. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning; SE = Standard Errors.

[a] flexMIRT uses the M-2PL model formula with the positive sign for the intercept, whereas BMIRT uses the negative sign. Therefore, the negative sign is added to the intercept parameter estimates from flexMIRT.

Table 5.4

Item Parameter Estimates of the M-2PL model, MML for Focal (female) and Reference (male) group ($N_{male} = N_{female} = 1,400$): English Usage Data

DIF item	Type	Parameter Estimate					
		MML (Focal group)					
		$\hat{a}_1$	SE	$\hat{a}_2$	SE	$\hat{d}$ [a]	SE
24	comparison	0.90	0.11	1.10	0.13	2.81	0.13
25	subordination	0.84	0.09	1.34	0.12	1.30	0.08
26	s-v agree	0.92	0.10	1.79	0.17	-3.33	0.18
27	pronoun agr	0.81	0.07	0.63	0.06	0.26	0.06
28	verb tense	1.21	0.14	1.35	0.15	2.64	0.13
29	pn clarity	1.43	0.12	0.62	0.07	1.16	0.08
30	pronoun agr	0.89	0.09	1.45	0.13	-2.22	0.12
31	pn clarity	1.24	0.10	0.72	0.07	0.32	0.07
Average			0.10		0.11		0.11

DIF item	Type	Parameter Estimate					
		MML (Reference group)					
		$\hat{a}_1$	SE	$\hat{a}_2$	SE	$\hat{d}$ [a]	SE
24	comparison	1.01	0.13	1.47	0.17	3.46	0.18
25	subordination	0.94	0.10	1.55	0.14	1.74	0.10
26	s-v agree	0.95	0.10	1.85	0.18	-2.88	0.16
27	pronoun agr	0.89	0.08	0.66	0.07	-0.04	0.06
28	verb tense	1.06	0.11	1.29	0.13	2.11	0.11
29	pn clarity	1.79	0.14	0.62	0.07	1.56	0.10
30	pronoun agr	0.72	0.08	1.06	0.11	-1.96	0.09
31	pn clarity	1.39	0.11	0.62	0.07	0.61	0.07
Average			0.11		0.12		0.11

Note. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning; SE = Standard Errors.

[a] flexMIRT uses the M-2PL model formula with the positive sign for the intercept, whereas BMIRT uses the negative sign. Therefore, the negative sign is added to the intercept parameter estimates from flexMIRT.

Table 5.5

Item Parameter Estimates of the M-2PL model, MCMC for Focal (female) and Reference (male) group ($N_{male} = N_{female} = 1,400$): English Usage Data

DIF item	Type	Parameter Estimate					
		MCMC (focal group)					
		$\hat{a}_1$	SE	$\hat{a}_2$	SE	$\hat{d}$	SE
24	comparison	0.50	0.17	0.81	0.15	-2.80	0.22
25	subordination	0.47	0.30	1.20	0.20	-1.48	0.09
26	s-v agree	0.50	0.12	1.73	0.22	2.97	0.25
27	pronoun agr	0.59	0.08	0.39	0.11	-0.40	0.08
28	verb tense	0.87	0.18	1.13	0.21	-2.77	0.20
29	pn clarity	1.24	0.12	0.34	0.18	-1.30	0.11
30	pronoun agr	0.57	0.15	1.31	0.13	1.90	0.12
31	pn clarity	1.05	0.15	0.50	0.20	-0.50	0.10
Average			0.16		0.18		0.15

DIF item	Type	Parameter Estimate					
		MCMC (reference group)					
		$\hat{a}_1$	SE	$\hat{a}_2$	SE	$\hat{d}$	SE
24	comparison	0.45	0.12	1.28	0.25	-3.38	0.50
25	subordination	0.63	0.25	1.45	0.23	-1.95	0.19
26	s-v agree	0.32	0.21	1.96	0.44	2.52	0.37
27	pronoun agr	0.78	0.21	0.46	0.13	-0.09	0.07
28	verb tense	0.73	0.13	1.21	0.24	-2.23	0.18
29	pn clarity	1.88	0.23	0.34	0.29	-1.71	0.17
30	pronoun agr	0.44	0.25	0.96	0.15	1.72	0.10
31	pn clarity	1.42	0.23	0.48	0.23	-0.78	0.11
Average			0.20		0.25		0.21

Note. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning; SE = Standard Errors.

Table 5.6 presents the results of Lord's Wald test for detecting DIF items in the English usage data. The critical value of the χ^2 distributions of Lord's Wald statistic with 3 degrees of freedom (i.e., $df = 3$) was used to determine if an item was identified as DIF. As can be seen in Table 5.6, seven items for each estimation method, items 24 through 29 and 31 for MML and items 25 through 31 for MCMC, were identified as DIF items at the nominal alpha level of 0.05. If an alpha level of 0.01 had been used, the results would be different. Only three items, items 25, 27, and 28, would be detected as DIF for MML, whereas six items, items 25 through 27 and 29 through 31, would be detected as DIF items. Thus, MCMC is likely to detect more DIF items than MML, which may be attributed to the inflation of Type I error rate of MCMC and/or due to the following reason. Based on the inspection of the item parameter estimates for the DIF items from the male and female groups, there were larger differences[4] in the *d* parameters than in the two *a* parameters between the two groups, implying DIF items were likely to be uniform DIF. Based on the simulation study results, MCMC performed better than MML when the low uniform DIF occurred with the 46-item test and small sample size condition (R1,000/F1,000).

[4] The average differences in the a_1, a_2, and d parameter estimates of the 7 DIF items from MML between the reference group and the focal group were, 0.14, 0.12, and 0.44, respectively. The average differences in the a_1, a_2, and d parameter estimates of the 7 DIF items from MCMC were 0.26, 0.14, and 0.38, respectively.

Table 5.6

DIF Detection Results by Lord's Wald Statistic in Multidimensional Two-Parameter Logistic Model (M-2PL): English Usage Data

DIF Item	Type	Lord Wald Statistics for χ^2					
		MML			MCMC		
		χ^2	*p-value*	DIF	χ^2	*p-value*	DIF
24	comparison	9.10	0.0280	Yes	5.50	0.1390	No
25	subordination	12.0	0.0073	Yes	15.60	0.0010	Yes
26	s-v agree	10.3	0.0159	Yes	38.05	0.0000	Yes
27	pronoun agr	12.8	0.0052	Yes	45.21	0.0000	Yes
28	verb tense	13.0	0.0046	Yes	9.02	0.0290	Yes
29	pn clarity	10.0	0.0184	Yes	70.80	0.0000	Yes
30	pronoun agr	7.30	0.0629	No	11.82	0.0080	Yes
31	pn clarity	10.8	0.0128	Yes	55.63	0.0000	Yes

Note. MML = marginal maximum likelihood; MCMC = Markov chain Monte Carlo; DIF = differential item functioning. The critical value of the alpha level of 0.05 for Lord's Wald Statistics is $\chi^2_{df=3}$=7.81.

CHAPTER 6

DISCUSSION AND CONCLUSION

In this chapter, the present study is summarized, and the strengths and limitations of the proposed Lord's Wald test for detecting DIF in the context of multidimensional IRT using the MML and MCMC estimation methods are presented. Similar to Woods et al. (2013), we investigated the performance of the MML estimation method using Wald-1 in Lord's Wald test. Compared to the unidimensional approach used in Woods et al.'s (2013) study in which unidimensional IRT models were assumed to study DIF, the IRT models used in this study were multidimensional IRT models, under which simulation studies and real data illustration were implemented. In addition, similar to Yao and Li (2010), we used an MCMC estimation method for obtaining Wald test statistics in M-3PL of the MIRT framework. Different from Yao and Li's multidimensional DIF test, this study included the comparison of the two estimation methods, MML and MCMC, for conducting Lord's Wald test in the context of M-2PL (Reckase, 1985) and M-3PL (Reckase, 1997). The present study provides the first comparison of the MML and MCMC estimation methods for DIF study in the multidimensional IRT DIF framework. MML and MCMC were implemented in flexMIRT and BMIRT software, respectively. The results from the simulation study and real data study are convincing in their demonstration of the potential for accurately estimating item parameters and show that MML and MCMC can be used in practice for calibrating multidimensional test items. In addition to a summary of the results of the simulation study and real data analysis, implications, limitations of the work, and future research directions are discussed in the following sections.

6.1 Summary and Implications of Results

The results of this study provide an important contribution for the comparison of two estimation methods of Lord's Wald test for DIF detection: Bayesian MCMC using a Metropolis-Hastings algorithm and MML using an SEM algorithm. DIF analysis is critical for ensuring test validity for all subgroups. Therefore, investigating, identifying, and treating DIF is necessary before the DIF effect is contaminated in making inferences for test results. To utilize Lord's Wald test in the context of multidimensional IRT, choosing a more efficient estimation method for DIF detection is imperative. The following summary of the results and relevant implications should provide test practitioners with useful information concerning the relative values of choosing one estimation method over the others for studying DIF in multidimensional tests.

In the simulation studies, the recovery of item parameters under M-2PL and M-3PL were examined using MML and MCMC, and Type I error rates, and power rates of the Wald test were evaluated under different simulation conditions: sample size, DIF type, DIF magnitude, and test length. Bias and RMSE showed very similar results. The item parameter recovery was reasonably good for the MCMC estimation method. MML was consistently inferior to MCMC with the exception of the 46-item test with the four DIF item condition for the intercept parameter under M-3PL.

Sample size was an effective factor in the accuracy of the item parameter estimates in M-2PL. The large sample (R5,000/F5,000) provided more accurate estimates for all item parameters compared to the small sample (R1,000/F1,000). The accuracy of the item parameter estimates increased as the sample size increased for discrimination parameters using MML and MCMC. However, the accuracy of the intercept parameter

estimates tended to be invariant regardless of sample sizes using MML, whereas the R3,000/F3,000 sample was the most accurate among the sample sizes for MCMC.

Test length affected the accuracy of the item parameter estimates when MML and MCMC were used under M-2PL and M-3PL. However, the pattern associated with the test length factor differed for the discrimination and intercept parameters; the bias and RMSEs for discrimination parameters under M-2PL became larger when MML and MCMC were used as the test length increased, whereas those for the intercept parameter became smaller. For M-3PL, the general patterns regarding test length were similar to the M-2PL results with some exceptions in the MCMC estimates.

Regarding the number of DIF items results in M-2PL, the biases and RMSEs of the item parameter estimates from the four-DIF item conditions were slightly lower than those from the eight-DIF item conditions for the a_1 and a_2 parameters using MML and MCMC. For the *d* parameter estimates, the bias and RMSE decreased as the number of DIF items increased when MCMC was used for all test length conditions, whereas the bias and RMSE decreased for the 24-item test and increased for the 46-item test as the number of DIF items increased when MML was used. A similar pattern regarding the number of DIF items from M-2PL was found in M-3PL for the a_1 and a_2 parameters MML was used. For the MCMC estimation, the small number of DIF items (i.e., four DIF items) conditions resulted in less bias and RMSE than the large number of DIF items (i.e., eight DIF items) conditions for the a_1 and a_2 parameters of the 24-item test, but not the 46-item test.

For the guessing parameter in the M-3PL, the guessing parameter was slightly better estimated using MCMC than MML. However, regardless of the estimation

methods, parameter recovery tended to be reasonable across all simulation conditions. In summary, the two estimation methods shared various patterns in discrimination parameter recovery between the M-2PL and M-3PL models. However, one interesting finding in the accuracy of the *d* parameter estimation was that M-3PL was always more accurate than M-2PL regardless of the estimation method.

For the Type I error and power study, 0.05 was used as the nominal alpha level. Thus, the Type I error rates of Lord's Wald test close to 0.05 of the nominal alpha level were reasonable in control. Although two types of power rates (theoretical and empirical) were presented in the simulation results, the empirical power rates were mainly discussed across the simulation conditions, because they took into account uncontrolled Type I error rates. For the power rates, the value of 0.80 was used as the criterion. Therefore, power above 0.80 was sufficient or high, and power below 0.80 was insufficient.

As the results of the simulation study, in general, the Type I error rates of Lord's Wald test when MML was used were close to or below the expected value of 0.05 across all simulation conditions of M-2PL with the exception of the unbalanced sample design (R4,000/F2,000) condition. The Type I error rates of MCMC under M-2PL were highly inflated across all conditions. Regardless of the MIRT models, the unbalanced sample design (R4,000/F2,000) condition reported higher Type I error rates than all the balanced sample design conditions for both estimation methods. This may imply that using balanced sample designs provides better Type I error control for both estimation methods than using unbalanced designs.

Under M-2PL, sample size adversely affected the Type I error rates of both estimation methods. The Type I error rates of the MCMC estimation method decreased as

the sample size increased, whereas the Type I error rates of the MML estimation method increased as the sample size increased. These findings pertain only to the balanced sample design conditions. The Type I error rates for MML under M-3PL were above 0.05 across all conditions in the medium sample condition (R3,000/F3,000) except for the 24-item test with four DIF items condition. Similar to M-2PL, the Type I error rates of MCMC was much higher than 0.05 across all conditions under M-3PL.

The results of the empirical power rates under M-2PL were affected by three primary factors, sample size, test length, and DIF magnitude, for all uniform and nonuniform DIF conditions. In general, larger samples with a balanced sample design (R5,000/F5,000) provided higher power rates. Sample size was the most influential factor on the power rates. The power rates of the two estimation methods increased as the sample size increased. The great improvement in power rates was made when the sample increased from 1,000 to 3,000, with relatively less improvement when the sample increased from 3,000 to 5,000. Thus, the sample size of 3,000 appears to be suitable for obtaining sufficient power rates (0.80) for both estimation methods. When low uniform and nonuniform DIF magnitude conditions were considered, the power rates of MML was above 0.90 for all medium and large samples (e.g., R3,000/F3,000, R4,000/F2,000, and R5,000/F5,000) conditions. The power rates of MCMC with all medium and large samples were above 0.70 in the low nonuniform DIF conditions and above 0.90 in the low uniform DIF conditions, respectively. Regardless of the estimation methods, power increased, as the DIF magnitude increased from low to medium, as expected. When the DIF magnitude was medium, the power rates with both estimation methods were above

0.80 with the small sample condition for all conditions with the exception of the 24-item test condition with eight DIF items using MCMC.

Lower power rates from MML were found as the test length increased, whereas higher power rates from MCMC were shown as the test length increased with small sample of R1,000/F1,000 in the low uniform DIF conditions. Similar to the uniform conditions, lower power rates from MML were found as the test length increased with the four DIF item nonuniform condition but not in the eight DIF item nonuniform condition. For MCMC, higher power rates appeared as the test length increased with small sample of R1,000/F1,000 in the low nonuniform DIF. Thus, test length should be considered carefully for a DIF analysis conducted under either MML or MCMC and with uniform or nonuniform conditions. In general, power was substantially higher in the larger sample conditions, the larger DIF magnitude conditions, and the test length conditions for both estimation methods of Lord's Wald tests.

For M-2PL, the power rates for MML were higher than those for MCMC in the uniform and nonuniform DIF conditions. Similar to M-2PL, higher power rates for M-3PL were found for MML than for MCMC with the exception of the 46-item test with eight DIF item low uniform condition. In general, across all simulation conditions, the power rates for M-2PL were much higher than those for the M-3PL model for MML and MCMC with the exception of the 24-item test with four DIF items when MML was used in the low DIF condition. For medium DIF magnitude, this pattern was less noticeable. This decreasing tendency from M-2PL to M-3PL became apparent from MCMC in nonuniform DIF conditions. The power rates of the longer test (46 items) were higher

than those of the short (24 items) test when MML and MCMC were used for M-3PL in the low and medium nonuniform DIF conditions.

The power results indicate that when uniform DIF is assumed, either of the two estimation approaches is a viable option. When the uniform DIF was small, MCMC tended to show slightly higher power than MML in the small sample, whereas MML had slightly higher power than MCMC for all other sample size conditions. In addition, when the two estimations were applied to the uniform medium DIF model, the power yield was very similar.

In the nonuniform DIF conditions, particularly the low DIF conditions of MML appeared to produce much higher power than those of MCMC on average. This pattern still appeared in the medium nonuniform DIF conditions; however, the difference between the two estimation methods diminished. Unlike the findings for uniform DIF, the power rates for MCMC in the nonuniform DIF condition were not higher than those for MML especially in the balanced small sample condition (R1,000/F1,000) and the unbalanced condition (R4,000/F2,000). The average power for MML was higher than that for MCMC in the uniform DIF condition across all conditions. Therefore, MML is preferred over MCMC when nonuniform DIF patterns are suspected.

Table 6.1 provides a summary of the results from the parameter recovery, the Type I error rates, and the power rates in the comparison of the MML and MCMC estimation methods from the M-2PL and M-3PL models. Based on the simulation study results, MML and MCMC provide comparable results on average for Type I error rates and power rates in detecting DIF. However, when the power values are considered, MCMC is superior to the MML estimation method for a DIF detection test that is

uniform DIF in a long (46 items) test with eight DIF items with the small sample of R1,000/F1,000 condition from M-2PL. For other conditions including nonuniform DIF conditions from M-2PL, MML is always recommended. These conclusions are limited to the conditions used in this study.

Table 6.1

Summary of Simulation Study Results

Test Type	MML	MCMC
Item Parameter Estimation (M-2PL)	Significant effect of test length and number of DIF items on discrimination parameters. Better precision with fewer DIF items for the short test condition on discrimination parameters. Better precision with fewer DIF items for the long test condition on the intercept parameter.	Significant effect of test length and number of DIF items on discrimination parameters. Better precision with fewer DIF items with short test condition on discrimination parameters. Better precision with fewer DIF items with long test condition on intercept parameter.
Item Parameter Estimation (M-3PL)	Significant effect of test length and number of DIF items on discrimination parameters. Better precision with fewer DIF items with short test conditions on discrimination parameters. Better precision with fewer DIF items with the long test condition on the intercept parameter. Better guessing item parameter estimates under most conditions.	Significant effect of test length and number of DIF items on discrimination parameters. Better precision with fewer DIF items with short test condition on discrimination parameter of the primary dimension. Better precision with more DIF items with short test condition on discrimination parameter of the secondary dimension. Better precision with more DIF items with short test condition on intercept

		parameter. Better guessing item parameter estimates at most conditions with the exception of fewer DIF items with long test condition.
Type I Error Rates (M-2PL)	Type I error rates close to or below the expected nominal value of 0.05 across all simulation conditions. Higher Type I error rates (range from 0.09 to 0.13) for unbalanced sample design. No systematic influence on test length or the number of DIF items factors.	Highly increased Type I error rates (over 0.20) at most simulation conditions. Unacceptably high increased Type I error rates (over 0.40) for small samples with all test lengths. Unacceptably high Type I error rates (over 0.40) for unbalanced sample design under all DIF conditions. Less increased Type I error rates as the sample size increased.
Type I Error Rates (M-3PL)	Slightly higher Type I error rates (0.06 or 0.07) under all DIF conditions. Higher Type I error rates with longer test length.	Highly increased Type I error rates (over 0.30) under all DIF conditions. Less increased Type I error rates as test length increased.
Power Rates (M-2PL)	Higher power values with balanced sample design, larger samples, or medium DIF magnitude.	Higher power values with balanced sample design, larger samples, medium DIF magnitude, or longer test. Higher power than MML with small sample with uniform DIF conditions.

	UNIFORM: Unacceptably low power rates (below 0.40) for small sample, long test with more DIF number in uniform conditions.	**UNIFORM:** Low power rates (below 0.60) for small sample, short test with fewer DIF number in uniform conditions. Higher power values than MML with small sample, long test with uniform DIF conditions.
	NONUNIFORM: Low power rates (below or at 0.70) for small sample with low DIF magnitude in nonuniform conditions. Higher power values with medium DIF magnitude and longer test in nonuniform conditions. Higher power values than those from M-3PL (more apparent in nonuniform conditions).	**NONUNIFORM:** Unacceptably low power rates (below 0.40) for small sample with low DIF magnitude in nonuniform conditions. Higher power values with medium DIF magnitude and longer test in nonuniform conditions. Higher power values than those from M-3PL (more apparent in nonuniform conditions).
Power Rates (M-3PL)	Higher power values with medium DIF magnitude. Lower power values with more DIF items in uniform and nonuniform conditions.	Higher power values with medium DIF magnitude. Higher power values with longer test.
	UNIFORM: Higher power values with short test, fewer DIF items in low uniform conditions.	**UNIFORM:** Higher power values with long test, more DIF items in low uniform conditions.
	NONUNIFORM: Higher power values with long test, fewer DIF items in low nonuniform conditions. Lower power values than those from M-2PL (more apparent in	**NONUNIFORM:** Higher power values with long test in low nonuniform conditions. Lower power values than those from M-2PL (more apparent in

	nonuniform conditions).	nonuniform conditions). Unacceptably low power rates (below 0.40) for the short test with small or large DIF number in nonuniform conditions.

6.2 Discussion and Possible Applications

There are three important findings in this present study. First, the item parameters for the multidimensional IRT models investigated were well estimated under certain conditions with the MML and MCMC estimation methods, even in the context of DIF study. In particular, we found that sample size and test length affected the accuracy of the parameter estimation. Zhang's (2012) results showed that MML yields satisfactory estimates of item parameters for the compensatory M-2PL model when the number of items and the sample size are large enough (30 items and 1,000 for his study). Yao and Boughton (2007) suggested that discrimination and intercept parameters from MCMC well recovered compared to an unweighted least squares method for the M-3PL model. Our results showed that the parameter recovery of MCMC is virtually better than that of MML under the M-2PL and M-3PL models. Previous researchers (Kieftenbeld & Natesan, 2012; Mislevy, 1986; Wollack et al., 2002) commented that MCMC parameter estimation method may improve the accuracy with smaller samples, test lengths, and more complex models such as multidimensional models (Béguin & Glas, 2001; Kieftenbel & Natesan, 2012).

Second, the results of this study suggest that MCMC can be a very useful alternative to MML for conditions such as a uniform condition with a small sample when the MML estimation method has not been introduced. As an example, MCMC was a

good estimation method for experimenting with a small sample with a long test in uniform DIF conditions, the 46-item test with the eight DIF item condition of R1,000/F1,000. The results showed that there were larger differences in the small sample than in the medium or large samples. It has been often found in the literature that MCMC worked well (better) in the small sample size compared to MML. A previous study (Mislevy, 1986) confirmed that MCMC outperformed than MML in the small sample conditions. The different results may be resulted from priors used in MCMC, different Wald tests (improved vs. traditional), and/or different estimation procedures.

Finally, MML may be a reasonable estimation method for Lord's Wald test. Based on the simulation study results, MML provided better Type I error rates than MCMC in detecting DIF for conditions that were composed of either the M-2PL model or the M-3PL model. In addition, for the power rate results, MML provided higher power rates than MCMC especially under nonuniform DIF conditions and unbalanced sample designs.

The study has several few important limitations. First, the findings in this study were obtained only from the Wald tests. It would be important to obtain effect size measures in DIF detection along with statistical DIF tests. This could be necessary because in the small sample case, interesting effects can be omitted in the analysis, whereas large samples can overestimate statistically significant findings where the true effect is minimal (Kirk, 1996; Zumbo, 1999). In light of comparison, further research is needed to see how the effect size measure in DIF such as the *p* metric can be applied and compared with Lord's Wald test DIF detection in the multidimensional framework.

Second, in this study, only two groups were considered in DIF detection. However, considering multiple groups is essential in practical educational settings and has been implemented in DIF detection in numerous studies (e.g., Cohen & Kim, 1993; Ellis & Kimmel, 1992; Jeon, Rabe-Hesketh, & Rijmen, 2013; Kim et al., 1995; Magis & de Boeck, 2011; Penfield, 2001). The generalized Lord's Wald test was demonstrated by Kim et al. (1995) in the unidimensional IRT model, and they showed how Lord's Wald test statistic can be extended to multiple group situations. Therefore, applying the generalized Wald test to multidimensional IRT models would be an interesting future study. DIF detection would benefit from other DIF testing methods such as multiple indicator multiple cause models (MIMIC; e.g., Muthen, 1985, 1989) and the generalized Mantel-Haenszel method (GMH; Somes, 1986; Zwick, Donoghue, & Grima, 1993), which can compare three or more groups concurrently. A comparison of these three DIF testing methods using multiple groups would be worthwhile in the multidimensional IRT context. In particular, applying MIMIC interaction models (Woods & Grimm, 2011) in the context of multidimensional IRT would be very valuable to identify DIF, because MIMIC models have been widely applied in unidimensional DIF studies (e.g., Finch, 2005, 2012; Jin, Myers, Ahn, & Penfield, 2013; Woods, 2009). Developing these three approaches should lead practitioners to identify DIF accurately in many realistic situations where multiple groups are most likely involved in and multidimensional tests are assumed.

Third, distributional and correlational differences of two ability parameters were not considered in the present study. Oshima et al. (1997) described that a distributional difference can arise from the correlation of two abilities and/or the location of two

abilities difference. Although the results from different conditions in Oshima et al.'s (1997) work were not severely different from each other, further research on detecting DIF with the mean θ difference and/or the correlational difference between the two groups is needed. However, as of now, the correlation difference between the two dimensions cannot be estimated in flexMIRT due to the model identification problem in the within-item design used in this study.

Fourth, the guessing parameter was set to the true value, 0.2 in this study. It would be worthwhile to compare the effect of randomly generated guessing parameters with that of the fixed guessing parameter in terms of bias and RMSE. Finch and French (2014) recently demonstrated how group differences in the pseudo-guessing parameter influence on detecting uniform and nonuniform DIF using IRT LR test and the logistic regression method for dichotomous items. When guessing differs between groups, it results in poor estimation of the discrimination and difficulty parameters and Type I error inflation for uniform and nonuniform DIF conditions under the unidimensional 3PL model (Finch & French, 2014). Little attention has been given to the group differences in the guessing parameter under the multidimensional IRT framework. Future work regarding the impact of the guessing parameter on DIF study in the multidimensional framework would be crucial.

This study evaluated the accuracy, stability, and viability of Lord's Wald test for DIF detection in the multidimensional IRT framework. MML performed robustly compared to the MCMC estimation method in detecting DIF, although MCMC estimated item parameters slightly more accurately. Since this was the first comparison of two estimation methods, further study of Lord's Wald test for DIF detection in the

multidimensional framework and its applications in various testing situations remain. Future directions include considering other parametric and nonparametric DIF methods and more factor manipulations of DIF conditions.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113-127.

Ackerman, T., Gierl, M., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22(3),* 37-53.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251-269.

American Institute of CPAs® (2015). The CPA Examination. Retrieved March 1, 2015, from http://www.aicpa.org/BECOMEACPA/CPAEXAM/EXAMINATIONCONTENT

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillside: Lawrence Erlbaum Associates.

Ansley, R. A., & Forsyth, T. N. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9,* 37-48.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling approach. *Applied Psychological Measurement, 22,* 153-169.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika, 66,* 541-561.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R. D. (1993). Different DIFs: Comment on the papers read by Neil Dorans and David Thissen. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 115–122). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26,* 381-409.

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33*, 335-352.

Bolt, D. M., & Lall, V.F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement, 27*, 395-414.

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

Budgell, G. R., Raju, N. S., & Quartetti, D.A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19*, 309-321.

Cai, L. (2008). SEM of another flavor: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*, 309-329.

Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

Cai, L., Thissen, D., & du Toit, S.H.C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software International.

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

Camilli, G., Wang, M.-m., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admissions Test. *Journal of Educational Measurement, 32*, 79-96.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (Rev. ed.). New York: Academic Press.

Cohen, A.S., & Kim, S.-H. (1993). A comparison of Lord's χ^2 and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17*, 39-52.

de la Torre, J., & Lee, Y-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, 50, 355-373.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel–Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*, 33-51.

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

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December, 1977: An application of the standardization approach* (ETS Research Rep. No. RR-83-9). Princeton NJ: Educational Testing Service.

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

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

Douglas, J. A., Roussos, L. A., & Stout W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*, 465-484.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two parameter logistic model. *Applied Psychological Measurement, 13,* 77-90.

Drasgow, F. & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 1, 2nd ed.* (pp. 755-636). Palo Alto, CA: Consulting Psychologists Press.

Educational Testing Service® (2015). The GRE® Revised Test. Retrieved March 1, 2015, from http://www.ets.org/gre

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29,* 278-295.

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis - based models. *Applied Psychological Measurement, 34,* 10-26.

Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement, 35,* 67-82.

Finch, W. H. (2012). The MIMIC model as a tool for differential bundle functioning detection. *Applied Psychological Measurement, 36,* 40-59.

Finch, H., & French, B. F. (2007). Detection of crossing differential item functioning item: A comparison of four methods. *Educational and Psychological Measurement, 67,* 565-582.

Finch. W. H. & French, B. F. (2014). The impact of group pseudo-guessing parameter differences on the detection of uniform and nonuniform DIF. *Psychological Test and Assessment Modeling, 56,* 25-44.

Fraser, C. H. (1987). NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models for latent trait theory [Computer program]. Armidale, New South Wales, Australia: Center for Behavioral Studies, the University of New England.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67,* 373-393.

Gamerman, D. (1997). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. London: Chapman & Hall.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science, 7,* 473-483.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 1-19). Washington, DC: Chapman & Hall.

Glas, C. M. A., & Meijer, R.R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27,* 217-233.

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

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

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale NJ: Erlbaum.

Houts, C. R., & Cai, L. (2013). flexMIRT R user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

Hulin, C., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Hillsdale NJ: Dow Jones-Irwin.

Jeon, M., Rabe-Hesketh, S., & Rijmen, F. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavior Statistics, 38*, 32-60.

Jin, Y., Myers, M. D., Ahn, S., & Penfield, R. D. (2012). A comparison of uniform DIF effect size estimators under the MIMIC and Rasch models. *Applied Psychological Measurement, 73,* 339-358.Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Judd, C. M., & McClelland G. H. (1989). *Data analysis: A model-comparison approach*. San Diego, CA: Harcourt Brace Jovanovich.

Kieftenbeld, V. & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36*, 399-419.

Kim, S.-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement, 25*, 163-176.

Kim, S.-H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test in detection of differential item functioning. *Applied Measurement in Education, 8*, 291–312.

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

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*, 261-276.

Kim, E.-S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*, 212-228.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Langer, M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Unpublished doctoral dissertation). University of North Carolina, Chapel Hill.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*, 164-174.

Lopez, G. E. (2012). *Detection and classification of DIF types using parametric and nonparametric methods: A comparison of the IRT-likelihood ratio test, crossing-SIBTEST, and logistic regression procedures.* (Unpublished doctoral dissertation). University of South Florida.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam: Swets and Zeitlinger.

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

Magis, D. & de Boeck, P. (2011). Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach. *Multivariate Behavioral Research, 46*, 733-755.

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

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

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1991). *The effect of sample size on the functioning of the Mantel-Haenszel statistic* (Rep. No. 211). Amherst: University of Massachusetts, Laboratory of Psychometric and Evaluation Research.

McCauley, C.D., & Mendoza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement, 9*, 389-400.

McDonald, R. P. (1997). Multidimensional normal ogive model. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York: Springer-Verlag.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11*, 161-173.

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-108.

Meng, X., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association, 86*, 899-909.

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

Miller, T. R., Spray, J. A., & Wilson, A. (1992, July). *A comparison of three methods for identifying nonuniform DIF in polytomously scored test items*. Paper presented at the Psychometric Society Meeting, Columbus OH.

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

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

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*, 92-109.

Muthen, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics, 10*, 121-132.

Muthen, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer and H. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Lawrence Erlbaum.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus: Statistical analysis with latent variables user's guide 6.0.* Los Angeles, CA: Muthén & Muthén.

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

Oshima, T. C., & Morris, S.B. (2008). An NCME instructional module on Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice, 27,* 43-50.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34,* 253-272.

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71,* 1023-1046.

Patz, R. J., & Junker, B.W. (1999b).A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146-178.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 14,* 235-259.

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

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve problems. *Psychometrika, 56,* 611-630.

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

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

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational and Psychological Measurement, 53,* 301-314.

Raju, N. S. & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and Analyzing Behavior in Organizations:*

Advances in Measurement and Data Analysis (pp. 156-188). San Francisco, CA: Jossey-Bass.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1992). *An IRT-based internal measure of test bias with applications for differential item functioning*. Paper presented at the Annual Meeting of American Educational Research Association, San Francisco.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. Applied *Psychological Measurement, 9*, 401-412.

Reckase, M. D. (1997). *A linear logistic multidimensional model for dichotomous item response data*. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer.

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

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data*. Paper presented at the annual meeting of the Psychometric Society, Toronto, Ontario, Canada.

Rogers, H. J. & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Journal of Applied Psychology, 17*, 105-116.

Rudner, L. M. (1977). *An evaluation of select approaches for biased item identification.* Unpublished doctoral dissertation, Catholic University of America, Washington DC.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5*, 213-233.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 85-100). New York, NY: Springer.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education, 2*, 255-275.

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

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*, 93-128.

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

Snow, T. K., & Oshima, T. C. (2009). A comparison of unidimensional and three-dimensional differential item functioning analysis using two-dimensional data. *Educational and Psychological Measurement, 69*, 732–747.

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician, 40*, 106– 108.

Spray, J. A. (1989). *Performance of three conditional DIF statistics in detecting differential item functioning on simulated tests* (ACT Research Report Series 89-7). Iowa City, IA: ACT.

Stark, S., Chernyshenko, O.S., & Drasgow, F. (2006). Detecting differential item functioning with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292-1306.

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

Suh, Y., & Bolt, D. M. (2011). A nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement, 48*, 188-205.

Suh, Y., & Cho, S-J. (2014). Chi-square difference tests for detecting functioning in a multidimensional IRT model: A Monte Carlo study. *Applied Psychological Measurement, 38*, 359-375.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 4*, 589-601.

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

Swaminathan, H., Hambleton, R., Sireci, S., Xing, D., & Rizavi, S. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement, 27*, 27–51.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19*, 1651-1683.

The College Board (2015). The SAT®. Retrieved March 1, 2015, from https://sat.collegeboard.org/why-sat/topic/sat/what-the-sat-tests

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory [Computer software]. Chicago: Scientific Software.

Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-111). Hillsdale, NJ: Erlbaum.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 56*, 611-630.

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

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123-135). Hillsdale NJ: Erlbaum.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The

1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8,* 157-186.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B., Rosa, K., Nelson, L. et al. (2001). Augmented scores—''borrowing strength'' to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is larger. *Transactions of the American Mathematical Society, 54,* 426-482.

Wang, W. –C., Wilson, M. R., & Adams, R. J. (1995). *Item response modeling for multidimensional between-items and multidimensional within-items.* Paper presented at the International Objective Measurement Conference, Berkeley, CA.

Wang, W.-C., Wilson, M. R., & Adams, R. J. (1997). Rasch model for multidimensionality between items and within items. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Vol. 4. Theory into practice* (pp. 139-155). Norwood, NJ: Ablex.

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

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12,* 239-252.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y.-S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26,* 339-352.

Wollack, J. A., & Cohen, A. S. (1997, March). *Detection of answer copying with unknown item and ability parameters*. Paper presented at the annual meeting of the American Educational Research Association.

Wood, R.L., Wingersky, M.S., & Lord, F.M. (1976). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (RM 76-6)* [Computer program]. Princeton NJ: Educational Testing Service.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44,* 1-27.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*, 532-547.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*, 339-361.

Yao, L. (2003). BMIRT: Bayesian multivariate item response theory. [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Yao, L. (2010). BMIRTII: Bayesian multivariate item response theory—second version. [Computer software]. Monterey, CA: *www.BMIRT.com*.

Yao, L., & Boughton, K. A. (2005b). Multidimensional parameter recovery from BMIRT and NOHARM. Manuscript submitted for publication.

Yao, L., & Boughton, K. A. (2007). A Multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*. *31*, 83-105.

Yao, L., & Li, F. (2010). A DIF detection procedure in multidimensional item response theory framework and its applications. Paper presented at the 2010, annual meeting of the National Council on Measurement in Education, Colorado, Denver.

Zhang, B. & Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement. 68*, 181-196.

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement*. *36*, 375-398.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007a). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Astivia, O. L. O. & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12(1)*, 136-151.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-198.

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