ACCURACY OF DIFFERENTIAL ITEM FUNCTIONING DETECTION METHODS IN STRUCTURALLY MISSING DATA DUE TO BOOKLET DESIGN

by

DEBRA ANNE SANDILANDS

B.Sc., The University of British Columbia, 1987
M.A., The University of British Columbia, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Measurement, Evaluation and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

June 2014

Abstract

Differential item functioning (DIF) analyses are used to analyze structurally missing data (SMD) due to balanced incomplete block (BIB) booklet designs commonly used in large scale assessments (LSAs). Only one DIF method, the Mantel Haenszel (MH) method, has previously been studied in this context. The purposes of this study were to investigate and compare the power and Type I error rates of an additional DIF method, the IRT-based Lord's Wald test, with the MH method and to extend the research on methods of forming the MH matching variable (MV) by proposing and testing a modification to the MH MV in the SMD context.

A simulation study investigated the effects of sample size, ratio of group sizes, test length, percentage of DIF items, and differences in group abilities on the power and Type I error rates of four DIF methods: the IRT-Lord's and MH using a block-wise, a booklet-wise, and a modified MV. The study design was selected to reflect authentic situations in which DIF might be investigated in LSAs that typically use BIB designs.

The three MH methods maintained better Type I error rates than the IRT-Lord's method which was inflated when the group sample sizes were unequal. None of the four methods had high power to detect DIF at the smallest sample size (1200). In the other sample size conditions the IRT-Lord's method had high power to detect DIF only when group sizes were equal. None of the MH methods had high power when the group mean ability levels differed, nor when the proportion of DIF in the MV was high.

These results indicate that DIF may go undetected in many realistic SMD conditions, potentially undermining the validity of score comparisons across groups. Recommendations to maximize DIF detection in SMD include using the MH method with a block-wise MV,

ensuring a large overall sample size, and over-sampling small policy-relevant groups to result in more balanced group sample sizes. Results also indicate that other sources of validity evidence to support score comparability should be provided since DIF analyses cannot yet be solely relied upon for this purpose.

Preface

This dissertation is original, unpublished work of the author, Debra Anne Sandilands. The author identified and designed the research program, performed all parts of the research, and analyzed the research data. Although the author did not write the entire computer code to perform the simulation, she planned and supervised the writing of the code and was responsible for its accuracy and execution.

Table of Contents

List of Tables

List of Figures

List of Acronyms

BIB	Balanced Incomplete Block booklet design
CML	Conditional maximum likelihood
DIF	Differential item functioning
ICC	Item characteristic curve
IRT	Item response theory
JML	Joint maximum likelihood
LSA	Large scale assessment
MCAR	Missing completely at random
MH	Mantel-Haenszel DIF method
MLE	Maximum likelihood estimation
MML	Marginal maximum likelihood
MNAR	Missing not at random
NAEP	National Assessment of Educational Progress
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
RMSE	Root mean square error
SIBTEST	Simultaneous Item Bias Test
R	R language and environment
TIMSS	Trends in International Mathematics and Science Study
YS	Youden Square booklet design
1PL	One parameter logistic IRT model
2PL	Two parameter logistic IRT model
3PL	Three-parameter logistic IRT model

Acknowledgements

I would like to express sincere gratitude to my dissertation supervisor, Dr. Kadriye Ercikan, for her invaluable guidance. I thank her not only for her strong, unwavering dissertation mentorship but also for including me in many research projects, teaching activities, publications, and conference presentations which have greatly broadened my experience and knowledge.

I am also deeply indebted to my dissertation committee members. I thank Dr. Bruno Zumbo for his wisdom and discussions that have contributed to this dissertation and beyond. I also thank Dr. Zumbo for recommending me for an internship that was a most rewarding experience. Thank you also to Dr. Sterett Mercer for his valuable constructive feedback and for encouraging me to think deeply about topics that I would not otherwise have considered.

I am also very grateful for many friends who have supported me along the way: Juliette, Shawna, Mary, Stephanie, Michelle, Chris, Malena, Sheila, and Santiago. I will also be forever indebted to Graham for his support with the R code and his amazing sense of humour that kept me sane when the coding got tough.

My most amazing family has also been a never-ending source of love, strength and support. Thank you. I couldn't have done this without each of you: Andy, Kate, Steph, Merrin, Laura, Mom, and Cynthia.

I would also like to acknowledge the Social Sciences and Humanities Research Council for generous financial support through a Joseph-Armand Bombardier Canada SSHRC Doctoral Fellowship.

1 Introduction

In the current global era of accountability in education, large scale assessments of educational achievement (LSAs) are increasingly being used to make comparisons of achievement within and across countries or regions (Chung, 2010; DeBoer, 2010; Klinger, DeLuca & Miller, 2008). Within countries, comparisons are often made between the LSA results of policy-relevant groups, such as groups based on gender, socio-economic status, first or second language, or special needs status for the purpose of informing educational policy, curriculum, program development and evaluation. However, if these comparisons are based on scores that are not comparable across the groups, validity may be reduced and inferences regarding differences in group performances may be inaccurate, leading to ill-informed or inappropriate policy-related decisions. Thus it is critical to gather evidence to support the comparability of scores across groups.

Differential item functioning (DIF) analyses can provide evidence of the degree to which scores are comparable across groups. DIF occurs when a test item has different psychometric characteristics for members of different groups despite there being no difference in their overall ability on what is being measured. DIF items are unexpectedly easier for one group relative to the other after the groups have been matched on their overall ability, with the result that the ensuing scores lack direct comparability across groups. By identifying DIF items, DIF analyses provide critical validity evidence about score comparability.

It is essential that DIF analyses produce accurate results in order to provide meaningful evidence of the comparability of scores across groups. The accuracy of DIF methods can be assessed in terms of Type I error rate and statistical power. In the context of

DIF, Type I error refers to the incorrect identification of non-DIF items as DIF items and power refers to the correct identification of items that do contain DIF. There are many factors that have been shown to affect the power and Type I error rate of various DIF methods. One factor whose impact on DIF detection accuracy has not yet been comprehensively studied is the presence of sparse data due to LSA booklet design.

1.1 LSA Booklet Design and Structurally Missing Data

The accuracy of DIF analyses may be affected by the presence of sparse data resulting from complex item sampling designs used by many LSAs. The most commonly used type of complex item sampling design is the balanced incomplete block booklet (BIB) spiraling design which was investigated in this dissertation. Complex booklet designs are necessitated by the large number of items required in LSAs to adequately assess broad curricular areas or content domains. It is not practical or feasible to present a large number of items to every examinee due to the amount of testing time that would be required and examinee fatigue that would result. Therefore in complex booklet designs the item pool is divided into groups of items referred to as blocks or clusters and each examinee is presented one booklet containing a limited number of blocks or clusters from one or more content areas. The booklets are randomly assigned or "spiraled" through the examinee population. This is referred to as matrix sampling. Complex booklet designs such as BIB designs and matrix sampling allow the large number of items deemed necessary to adequately cover the cognitive domains to be administered to a sufficient number of examinees so that stable estimates of the items' psychometric characteristics and examinee proficiency can be determined (Rutkowski, Gonzalez, Joncas, & von Davier, 2010).

One challenge arising from the use of complex booklet designs such as BIB designs is that they contribute to sparseness of data at the item level because there are large portions of empty cells (missing data) for each examinee due to items that were not administered to that examinee. This type of data is referred to as structurally missing data, data that is missing by design, sparse data or incomplete data (Goodman, Willse, Allen, & Klaric, 2011; Kaplan, 1995). These terms will be used throughout this dissertation to refer to the sparse data resulting from BIB designs.

Booklet design and structurally missing data have implications for subsequent statistical analyses (Frey, Hartig, & Rupp, 2009) and ultimately for the validity of inferences drawn from an assessment. Missing data presents a problem in all types of statistical analyses because statistical methods were developed to analyze rectangular data sets containing complete data (Little & Rubin, 1989) and many software programs make the assumption that all variables are measured for all cases. Problems associated with conducting statistical analyses in the presence of missing data include decreased statistical power (Baraldi & Enders, 2010) which decreases the ability to detect meaningful relationships in the data. Another problem is that ignoring or inappropriately handling missing data can give rise to misleading analysis results which can lead to erroneous conclusions about the data. Although there are techniques for handling missing data in statistical analyses they are not suitable for use with structurally missing data due to the large amount of data that is missing in these designs.

In addition to these problems, booklet design can have particular implications for DIF analyses. For example, in BIB designs each item may be administered in more than one booklet. Therefore, if the data is analyzed by individual booklets each item will have more

than one DIF statistic associated with it. On the other hand, if the data is analyzed by blocks (which have fewer items than the entire booklet) the DIF analysis will be based on fewer items which may reduce reliability of the criterion on which the groups are matched and thereby reduce the meaningfulness of the DIF analysis results.

The following sections present an overview of DIF and DIF analyses which will provide a context through which the past research in this area and the research problem addressed in this dissertation can be better understood.

1.2 Overview of DIF and DIF Analyses

Differences in measurement by test items can be identified as differences in psychometric properties (such as an item's level of difficulty or discrimination) for examinees of equal ability from different groups (Hambleton, Swaminathan, & Rogers, 1991). When DIF occurs, members of one group have a different conditional probability of correctly responding to or endorsing an item than members of the comparison group despite there being no difference in their overall ability on the characteristic the test is intended to measure. In other words, when DIF occurs there is an unexpected performance difference on an item between examinees of the same ability after conditioning on ability.

In DIF terminology the studied groups are referred to as focal and reference groups. The reference group is typically a majority group against whom the focal group is compared. The focal group may be believed to have potential educational or societal disadvantage whereas the reference group may be believed to have a relative advantage. Focal groups that have historically been studied in DIF analyses include minority cultural or ethnic groups, language groups, and groups based on examinee characteristics such as gender, or socioeconomic status.

In general, DIF methods are concerned with statistically testing a null hypothesis that no DIF occurs – that the conditional probability of correctly responding to or endorsing an item does not depend on group membership – or providing a measure of the manner and extent to which the conditional probabilities differ (Penfield & Camilli, 2007). DIF methods based on classical test theory use contingency tables (e.g., Mantel-Haenszel, MH (Holland & Thayer, 1988)) or regression models (e.g., logistic regression methods (Swaminathan & Rogers, 1990)). DIF methods based on item response theory (IRT) include unidimensional IRT methods (Thissen, Steinberg, & Wainer,1993) and methods that view DIF as occurring due to multidimensionality or a secondary nuisance dimension unrelated to the measured construct (e.g., Simultaneous Item Bias Test, SIBTEST (Shealy & Stout, 1993)).

DIF analysis methods can be categorized as observed score or latent score approaches. In observed score DIF approaches the groups are matched on the ability of interest (often the total score on the test) before investigating whether there is a group effect. The observed score approach interpretation of a finding of DIF is that after statistically controlling for (or "conditioning on") the differences in item responses due to the ability being measured the groups still differ. In contrast, latent variable DIF approaches do not match the groups by conditioning on the total score. Rather, a latent ability score and item psychometric characteristics are estimated from the data through an algorithm based on the examinees' item response patterns. During DIF analysis the latent ability score is integrated out and not used as a conditioning or matching variable. The latent score approach interpretation of a finding of DIF is that a DIF item has different psychometric characteristics for the two groups. Item response theory methods are latent score approaches whereas the contingency table, regression equation and SIBTEST (Shealy & Stout, 1993) methods are

observed score approaches. It is important to understand how structurally missing data affects both observed score and latent score DIF methods since both are routinely used in practice to assess DIF in LSAs.

1.3 Limited Past Research on DIF Detection in Structurally Missing Data

The implications of booklet design for subsequent statistical analyses have not been thoroughly dealt with in the literature (Frey et al. 2009; Kubinger et al., 2011). In particular, although many studies have investigated the identification of DIF using complete data sets or data sets containing missing data not due to assessment design, very few studies have examined the performance of DIF analysis methods when data is missing due to booklet design. A thorough review of relevant literature revealed only two published studies (Allen & Donoghue, 1996; Goodman et al., 2011).

The first study (Allen & Donoghue, 1996) focused on the impact of different approaches to forming the matching criterion on the performance of MH for BIB design data. This study recommended a pooled booklet approach in which information from all booklets in which an item appears is included in the creation of the matching variable. The other study (Goodman et al., 2011) investigated the Type I error rate, power and accuracy of the MH method to detect DIF in three sparse booklet designs (one of which was a BIB design) using the pooled booklet matching variable approach recommended by Allen and Donoghue. In contrast to Allen and Donoghue, Goodman et al. found that using the pooled booklet approach led to an increase in the Type I error rate and a drastic decrease in the power of MH to detect DIF in BIB data, particularly at the lowest sample size and longest test length. This was attributed to the way in which the matching criterion was formed when using the pooled booklet approach.

There are significant gaps in the past research that remain to be investigated regarding the performance of DIF methods in identifying DIF in structurally missing data. Only one method, the MH method, has been investigated in this context, and only to a very limited extent. Further, the two relevant past studies present conflicting results regarding the effectiveness of the pooled booklet approach to forming the MH matching criterion, particularly in low sample sizes and long tests. Further investigations of the MH method and the pooled booklet approach in the context of structurally missing data are needed. In addition, since no studies have investigated the performance of other DIF detection methods in structurally missing data such studies are also vitally needed. Both of the studies that investigated identification of DIF in structurally missing data emphasized the importance of conducting further studies using the MH method as well as investigating IRT-based DIF detection accuracy for structurally missing data, neither of which have been done to date.

1.4 Purpose and Significance of the Study

Given the paucity of research in this area, the main purpose of this dissertation was to extend the previous research on the performance of the MH method to detect DIF in structurally missing data and to produce new research on the performance of another DIF method in this context and compare it to the performance of MH.

Specifically, the research goals were to investigate and compare the MH DIF method and a second DIF method with respect to power and Type I error rates, and to explore how the power and Type I error rates of these methods are affected when analyzing structurally missing data from a BIB design by other factors that have been shown to affect DIF detection in complete data (such as sample size, ratio of focal to reference group sample sizes, test length, percentage of DIF items, and true ability differences between the groups). To build

on past research and provide information about the performance of latent variable DIF methods in structurally missing data, this study investigated the MH and Lord's chi-squared method (Lord, 1977, 1980) which is also referred to as Lord's Wald method (Langer, 2008). Lord's Wald method is a unidimensional IRT method that uses a latent variable approach. In addition, a third goal of this dissertation was to extend the previous research about the MH method by comparing methods of forming the matching variable and proposing and testing a modification to the pooled booklet approach for use in structurally missing data.

As LSAs are increasingly being used to draw comparisons across groups and make policy and curricular decisions, it is becoming more important to provide evidence to support such comparisons. Because DIF analyses are one way of providing this type of evidence, it seems reasonable to assume that DIF analyses will be used more frequently in this context. It is therefore increasingly more important to understand the performance of commonly-used DIF detection methods in structurally missing data and to investigate modifications through which their power and Type I error rate may be improved.

The results of this study may be useful to guide LSA developers in designing tests from which the resulting data will be appropriate for analysis using these commonly-used DIF methods. For example, working within the constraints of a specific test design, test developers may wish to know whether a particular combination of test length and sample size would result in data that is suitable for analyzing DIF for small yet important policy-relevant groups. They will also benefit from knowing which DIF detection method performs best under those conditions.

1.5 Structure of the Dissertation

The remainder of this dissertation is organized in four chapters. Chapter 2 presents a review of relevant literature, beginning with a review of BIB booklet designs commonly used in LSAs and the data structure that results from these designs. Chapter 2 also provides more details about the DIF detection methods to be used in this study and the potential impact of structurally missing data on them. This is followed by a review of factors that affect these DIF detection methods and a review of previous research on structurally missing data and DIF detection. Chapter 2 also presents details of the proposed modification to the pooled booklet approach for the MH procedure. Chapter 3 presents details of the simulation study methods used to address the research goals, and Chapter 4 presents study results. Finally, Chapter 5 provides a discussion of the results and the contributions of this research to the literature.

2 Literature Review

This chapter is organized in six sections. In the first section I provide more thorough details about LSA booklet designs than was provided in the introductory chapter. I also describe and provide examples of two generic BIB designs and one specific BIB design used by the Programme for International Student Assessment (PISA), an international LSA. In the first section I also explain the data structure and the mechanism of missing data in BIB designs in terms of Rubin's (1976) missing data classification scheme. The second section describes the DIF analysis methods that will be used in this study, and discusses the advantages and disadvantages of each. In the third section I provide a literature review of research on DIF detection in structurally missing data. The fourth section explains the proposed modification to the pooled booklet approach for the MH DIF method. The factors that affect DIF detection accuracy that were investigated in this study are introduced in the fifth section. Finally, the last section restates the research purpose in the context of the findings from the literature review. This serves to refocus the reader on the purpose of the study prior to being introduced to the study methodology in the following chapter.

2.1 LSA Booklet Designs and Data Structure

The goal of LSAs is to obtain reliable information about student achievement in one or more content domains, predominantly at the group level. Hundreds of questions may be required in order to achieve this. Since it is not practical or feasible to administer so many questions to every examinee, the examinees are randomly assigned different booklets which contain only a portion of the total number of items.

The assignment of items to booklets in BIB spiraling design is not random but is guided by constraints. There are many constraints on booklet design, including the number

of items to be used, the number of content domains to be assessed, item psychometric attributes and formats, testing format and testing time, whether the test is to be linked to other assessments, and whether item confidentiality is required (van der Linden, Veldkamp, & Carlson, 2004). These constraints vary with each assessment situation, and they make the design of booklets a complex and challenging undertaking specific to each assessment.

One way of approaching booklet design is to view it as a type of experimental design in which items or groups of items are systematically assigned to booklets. As in experimental design, one goal of test booklet design is to control for extraneous sources of variance. Three types of variance are of particular concern in LSA booklet design. They are variance due to booklet effects, variance due to position effects, and variance due to carryover effects. Variance due to booklet effects arises from differences between the booklets used. To ensure that the scores resulting from different booklets are comparable, the psychometric properties of the booklets should be comparable. Position effects refer to the fact that the placement of clusters of items within booklets or of items within clusters can affect the psychometric properties of the cluster or item. For example, an item placed at the end of the test may appear to be more difficult than the same item placed near the beginning of the test simply because examinees ran out of time or became fatigued before reaching the item at the end of the test. Carryover effects occur when the context in which an item is presented affects examinee responses to the item. As an example, responses to an item may be positively or negatively affected by the items that precede it. Most LSAs use blocking to attempt to control for these types of variance.

Blocking to control for variability due to different booklets can be achieved by dividing the total number of items into groups and assigning the groups to booklets. The

groups of items are referred to as either “blocks” or “clusters”. Throughout the remainder of this dissertation “block” will be used although it is meant to be interchangeable with “cluster”. The resulting booklet design is referred to as an incomplete block design if the number of blocks per booklet is smaller than the total number of blocks that exist. Most large-scale assessments use some form of incomplete block design (Kubinger et al., 2011). Two particular incomplete block designs are commonly used in LSAs: balanced incomplete block designs and Youden square designs which are a special case of balanced incomplete block designs. These two booklet designs are described next.

2.1.1 Balanced Incomplete Block Design

Balanced incomplete block (BIB) designs have been described as having three structural constraints which are: (1) the number of blocks assigned to each booklet is between certain bounds; (2) the number of booklets to which each block is assigned is between certain bounds; and (3) combinations of blocks are assigned to a minimum number of booklets if statistical relations between items in different blocks are required (van der Linden et al., 2004). Frey, Hartig, and Rupp (2009) make these constraints more explicit by stating that BIB designs satisfy four conditions involving five design parameters. The design parameters are: the number of blocks; the number of booklets; the number of times blocks appear in booklets; the number of blocks per booklet; and the frequency with which pairs of blocks appear in booklets. The four conditions that are satisfied in BIB designs are: (1) every block occurs no more than once in a booklet; (2) all the blocks appear the same number of times across all the booklets; (3) the booklets are of identical length, i.e. containing the same number of blocks; and (4) every pair of blocks occurs together in

booklets with equal frequency. An example of possible placement of blocks in booklets in a BIB design (from Frey et al., 2009) is shown in Table 1:

Table 1: ***Sample BIB Design***

Booklet	Block		
1	1	2	4
2	2	3	5
3	3	4	6
4	4	5	7
5	1	5	6
6	2	6	7
7	1	3	7

The placement of blocks into booklets and the position in the booklet into which the block was placed are shown in the shaded area. For example, students who were administered booklet 1 received blocks 1, 2 and 4 in that order in their booklet. It can be seen that this sample BIB design has 7 booklets, 3 positions in each booklet, 7 blocks, and 3 occurrences of each block. Further, there is 1 occurrence of each pair of blocks, for example the blocks 1 and 2 only appear together in booklet 1. Please note that, although it does occur in this particular example, there is no requirement that the number of blocks equals the number of booklets or that the number of positions in each block and occurrences of blocks are equal in a BIB design. For example, another possible BIB design has 12 booklets, 3 positions per booklet, 9 blocks and 4 occurrences of each block. As long as the booklets are administered to an equal number of students the blocks will also be administered to an equal number of students, thus booklet effects are controlled. Further, there can be statistical linking across booklets because of the shared blocks across booklets. This BIB design does not control for

position effects: note that block 1 only appears in the first position in booklets, and block 7 only appears in the last position in booklets.

2.1.2 Youden Square Design

A Youden square (YS) design is a special case of a BIB design that can be used to control for position effects. YS designs have the same 4 conditions as those required for BIB designs, but YS designs incorporate an additional blocking factor to control for position effects by requiring that each block appear in each position within a booklet. An example of a YS design (also from Frey et al., 2009) is shown in Table 2:

Table 2: ***Sample YS Design***

Booklet	Block		
1	1	2	4
2	2	3	5
3	3	4	6
4	4	5	7
5	5	6	1
6	6	7	2
7	7	1	3

Similarly to the BIB design example shown in Table 1, the YS design has 7 blocks, 7 booklets, 3 occurrences of each block, 3 positions within each booklet, and 1 occurrence of each block pair. In addition to the requirements for a BIB design, YS designs require that every block occurs in every position the same number of times which controls for position effects in addition to booklet effects. Notice now that block 1 appears in position 1 (booklet 1), in position 2 (booklet 7), and in position 3 (booklet 5). Despite the advantage of controlling for position effects, YS designs have disadvantages of being very restrictive in that only a few combinations of design parameters exist, and of being difficult to construct.

While the BIB designs attempt to control for booklet effects and YS designs attempt to control for both booklet effects and position effects, neither design can control for carryover effects. Carryover effects may be minimized through the use of other designs (for example, complete permutation designs or repeated treatment designs). However, these designs are typically not used for LSA booklets because the constraints required in order to utilize them make them infeasible for most LSAs. For example, they require too many items or too many booklets to be practical in LSA applications. For this reason carryover effects are usually not controlled in LSAs.

2.1.3 Example Booklet Design: PISA 2006

In order to further exemplify LSA booklet design the design used by PISA is presented in Table 3. PISA is an international LSA administered in three year cycles to 15-year-old students nearing the end of compulsory education. Its purpose is to determine the extent to which the examinees have attained the required knowledge, skills, and attitudes to succeed and fully participate in society after compulsory education. Table 3 shows the allocation of blocks to booklets in the PISA 2006 study (OECD, 2009). (Although the PISA documentation refers to groups of items as clusters, I refer to them as blocks for consistency with the remainder of this dissertation.) The PISA main study consisted of 179 items which were placed in 13 booklets composed of 4 blocks each as shown next.

Table 3: ***PISA 2006 Test Booklet Design***

Booklet	Block			
1	S1	S2	S4	S7
2	S2	S3	M3	R1
3	S3	S4	M4	M1
4	S4	M3	S5	M2
5	S5	S6	S7	S3
6	S6	R2	R1	S4
7	S7	R1	M2	M4
8	M1	M2	S2	S6
9	M2	S1	S3	R2
10	M3	M4	S6	S1
11	M4	S5	R2	S2
12	R1	M1	S1	S5
13	R2	S7	M1	M3

S1….S7 = science blocks 1 to 7
M1…M4 = mathematics blocks 1 to 4
R1, R2 = reading blocks 1 and 2

Table 3 shows that PISA 2006 used a mixed design meaning that booklets contain blocks from 3 different content domains: science (a total of 7 blocks labeled S1 through S7), mathematics (a total of 4 blocks labeled M1 through M4), and reading (a total of 2 blocks, labeled R1 and R2). Further, the domains are not equally represented in the booklets. For example, Booklet 1 has 4 science blocks and no mathematics or reading blocks whereas Booklet 2 has 2 science blocks, 1 mathematics block and 1 reading block. PISA 2006 utilized a YS design: each block occurs no more than once per booklet; all blocks appear the same number of times across the booklets (four times); the booklets all contain 4 blocks; every pair of blocks occurs together in booklets with equal frequency (once); and each block appears in each of the four positions in the booklets. However, there remains a possibility that booklet effects would occur because of the different locations of the domains within the booklets.

2.1.4 Data Structure of BIB Designs

The structure of data that results from BIB designs is illustrated in Table 4. Table 4 shows sample data for the YS design example shown in Table 2 and described in the previous subsection. For the purpose of providing a simplified example, each block contains only two items and each booklet was administered to only two students. Examinees and the booklets they responded to are shown in the two left hand columns. Blocks and the items they contain are shown in the top two rows. Examinee scores on each item are shown in the body of the table, coded 1 or 0 for correct and incorrect responses respectively, and the total observed score for each examinee is shown on the right hand column of the table. The shaded blank spaces in the body of Table 4 represent the items that were not presented to examinees, i.e. the structurally missing portions of the data.

Table 4: ***Sample Data for YS Design***

		Block #	1		2		3		4		5		6		7		
Examinee #	Booklet #	Item #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total Observed Score
1	1		1	1	1	0			1	1							5
2	1		1	1	1	1			1	1							6
3	2				1	0	1	1			1	0					4
4	2				1	0	1	1			0	0					3
5	3						1	1	0	1			0	1			4
6	3						0	1	1	1			1	1			5
7	4								1	0	0	1			0	0	2
8	4								1	0	1	0			0	1	3
9	5		1	0							1	1	0	1			4
10	5		1	1							1	0	1	1			5
11	6				1	1							1	1	1	1	6
12	6				1	1							1	0	1	0	4
13	7		0	0			0	1							0	1	2
14	7		1	1			1	0							1	1	5

To clarify the contents of the table, note that Block 1 contains items 1 and 2, Block 2 contains items 3 and 4, and so on. Examinees 1 and 2 wrote booklet 1 which contained 6 items in total: items 1 and 2 (in block 1), items 3 and 4 (in block 2) and items 7 and 8 (in block 4). These two examinees were not administered items 5, 6, or 9 through 14. Examinees 1 and 2 had observed scores of 5 and 6 respectively. In this example for simplicity there is no missing data other than that due to booklet design therefore the pattern of missing data that results only from booklet design is apparent. Please note that in this particular example the number of blocks is equal to the number of booklets; however that is not a requirement for BIB or YS designs.

2.1.5 Mechanism of Missing Data in BIB Design

Data that is missing as a result of booklet design is classified as missing completely at random (MCAR) as defined by Rubin (1976). In MCAR data the probability that a value is missing does not depend on any variable in the data including the variable on which the value is missing. The observed data are said to be a simple random sample of the complete data. The MCAR missing data mechanism is contrasted with data that is missing at random (MAR) in which the probability of a missing response is associated with a measurable variable, and data that is missing not at random (MNAR) in which the probability of a missing response is associated with the missing variable itself.

In many real scenarios the assumption of MCAR is not tenable; however missingness is said to be due to a random mechanism if the researcher has control over the occurrence of the missing data. In the case of BIB data, where the different booklets are randomly assigned to examinees it is assumed that examinees that are administered one booklet are not systematically different from those that are administered other booklets. Therefore when data are missing by design the MCAR assumption is expected to hold and the missing data mechanism is said to be ignorable (Graham, Hofer, & MacKinnon, 1996; Mislevy & Wu, 1988; Schafer & Graham, 2002). Here, ignorable means it is possible to estimate parameters of interest without knowing the parameters of the missing data distribution or modeling the mechanism for missingness in the estimation process. Thus, it is theoretically justified to ignore the process that causes missing data due to booklet design when making inferences about the distribution of the data. However, as noted by Zwick (1987) and Eggen and Verhelst (2011) (discussed in a following section) the results of statistical analyses may still be affected by missing data due to booklet design even though the data is MCAR. For

example, position effects and the fact that the different booklets are not necessarily designed to be statistically equivalent may affect statistical analysis results.

To summarize, most LSAs use booklets with a BIB design or a YS design which is a special case of a BIB design. In these designs each booklet contains only a subset of the total number of items and blocks of items in the assessment. Both designs involve blocking to control some of the extraneous sources of variation that may have an impact on item and person parameters. Both allow for statistical linking between items because of the shared items between booklets, and both result in large amounts of structurally missing data which is classified as MCAR.

2.2 DIF Analysis Methods Investigated in this Study

The three most frequently used methods to provide statistical evidence about the equivalence of items for different groups are the MH procedure and its extensions; logistic regression (LR) procedures; and unidimensional IRT procedures (Hambleton, 2005). Two of these three methods, MH and the unidimensional IRT DIF Lord's Wald method (Lord, 1977, 1980) were investigated in this study. MH was investigated because it is the only method which has been investigated in past research on the accuracy of DIF detection in structurally missing data thereby allowing a comparison between the results of this research and past research. A unidimensional IRT DIF procedure was used because unidimensional IRT models for binary items are the most commonly used IRT models for DIF detection. In addition, no previous study has investigated the performance of a latent variable approach such as Lord's Wald method on DIF detection in structurally missing data.

The statistics of various DIF detection methods can detect different types of DIF. DIF can be uniform, that is, constant across the ability continuum such that one group is

advantaged equally over all levels of ability. Alternatively, DIF can vary across the ability continuum (referred to as non-uniform DIF). When non-uniform DIF occurs the difference between the groups in conditional probability of a correct response varies along the ability continuum in magnitude or in direction. This can result in one group being advantaged to different degrees at different locations, or in the case of non-uniform crossing DIF one group could be advantaged at one end of the ability continuum while the other group is advantaged at the other end of the continuum (Swaminathan & Rogers, 1990). In addition, different DIF methods are able to detect DIF in dichotomous item responses (binary scored formats) and polytomous item responses (ordinal scoring formats). Past research on DIF in structurally missing data has used the MH observed score method to investigate uniform DIF in dichotomous items. Since the purpose of this study was to build on previous research, this study investigated the performance of MH and Lord's Wald method (Lord, 1977, 1980) to detect uniform DIF in dichotomous items to enable comparison of results between this study and previous studies.

2.2.1 Mantel-Haenszel DIF Method

The MH approach (Holland & Thayer, 1988) is the most widely known of the observed score DIF methods. It is based on the analysis of contingency tables wherein groups are compared on their relative likelihood of success on an item after being matched on ability and the expectation that examinees of equal ability should have an equal probability of success on an item regardless of their group membership. The matching criterion is a variable determined by the researcher and is usually the total test score. The matching criterion is divided into groups referred to as slices or levels; often there are as many levels as there are possible score values. A 2 x 2 table is formed for every level of the matching

criterion, as shown in Table 5 (Holland & Thayer, 1988) and the groups are compared at every level as described next.

Table 5: ***Sample Data for the j^{th} Level of the MH Matching Variable***

	Score on Studied Item		
Group	1	0	Total
Reference	A*j*	B*j*	N*rj*
Focal	C*j*	D*j*	N*fj*
Total	M1*j*	M0*j*	T*j*

In this table, A*j* is the number of reference group members at the *j*th level of the matching criterion who scored correctly on the item, and B*j* is those who scored incorrectly. Similarly, C*j* and D*j* represent the number of focal group members at *j*th level of the matching criterion who scored correctly and incorrectly, respectively. N*rj* and N*fj* are the numbers of reference and focal group members at *j*th level of the matching criterion. M1*j* and M0*j* are the total numbers of examinees who scored correctly and incorrectly respectively, while T*j* is the total number of examinees at *j*th level of the matching criterion.

At each level of the matching criterion the null hypothesis of no DIF is tested against the alternative hypothesis of DIF as follows:

$$H_1 = \frac{P_{rj}}{Q_{rj}} = \alpha \frac{P_{fj}}{Q_{fj}} \qquad P_{rj} = \frac{A_j}{N_{rj}} \quad Q_{rj} = \frac{B_j}{N_{rj}} \quad P_{fj} = \frac{C_j}{N_{fj}} \quad Q_{fj} = \frac{D_j}{N_{fj}}$$

where $\alpha \neq 1$ and α can be estimated as:

$$\hat{\alpha}_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j}$$ Equation 1

$\hat{\alpha}_{MH}$ is the odds ratio, that is the odds that a reference group examinee will have a correct response to an item divided by the odds that a matched focal group examinee will have a correct response. It is assumed that the odds ratio is constant across all levels of the matching criterion. The odds ratio can be difficult to interpret, therefore a logistic transformation is computed which makes the scale symmetric around zero. To make this result comparable to the commonly-used Educational Testing Service delta scale, the logistic transformation is multiplied by -2.35 which results in the measure known as Δ_{MH} (also referred to as MH D-DIF):

$$\Delta_{MH} = -2.35\ log_e\ (\hat{\alpha}_{MH})$$ Equation 2

It is possible to determine the statistical significance of Δ_{MH} through a z-test in which the MH log odds ratio is divided by its standard error.

A one degree of freedom chi-squared test of significance is calculated as:

$$MH_{\chi^2} = \frac{\left(|\sum_j A_j - \sum_j E(A_j)| - \frac{1}{2}\right)^2}{\sum_j var(A_j)}$$ Equation 3

where:

$var(A_j) = \frac{N_{rj} N_{fj} M_{1j} M_{0j}}{T_j{}^2 (T_j - 1)}$.

and $E(A_j) = N_{rj} M_{1j} / T_j$.

2.2.1.1 Effect Size Measure for the MH DIF Statistic

An effect size measure can be used in addition to the MH DIF chi-squared statistic to distinguish statistically significant but trivial DIF from non-trivial DIF. Effect size measures complement MH DIF statistics and provide guidance on the practical significance of DIF items. Using both a statistical significance test and an effect size measure to identify DIF items represents a blended heuristic decision rule for identifying DIF items (Gómez-Benito, Hidalgo, & Zumbo, 2013; Zumbo, 2008).

A blended decision rule is often used in MH analyses. For example, the Educational Testing Service uses a three-level classification to distinguish between A, B and C level DIF, described by Zwick and Ercikan (1989) in the following manner:

> "A" items are those for which MH D-DIF is not significantly different from 0 (a = .05) or has an absolute value less than 1. These items are considered to be free of DIF. "B" items are those for which MH D-DIF is significantly different from 0 (a = .05) and has either (a) an absolute value at least 1 but less than 1.5 or (b) an absolute value at least 1 but not significantly greater than 1 (a = .05). These items may be used, but if there is a choice among otherwise equivalent items, it is considered desirable to select for inclusion in a test those with the smallest absolute value of MH D-DIF. "C" items are those for which the absolute value of MH D-DIF is at least 1.5 and is significantly greater than 1 (a = .05). These items are to be selected only if it is essential to meet test specifications. (p. 58-59)

2.2.1.2 Advantages and Limitations of the MH Method

The benefits of using the MH approach are that it has been shown to be effective with small sample sizes, that it is efficient in terms of statistical power, that it is computationally straightforward, and that the effect size measure is easily interpretable (Rogers, 2005). There are, however, some drawbacks to using the MH approach and situations in which it may not be the most appropriate method. First, the MH test statistic assumes equal item discrimination for both groups and is therefore only able to detect uniform DIF. However, if

it can be assumed that only uniform DIF is present, then MH has the most power of the observed score methods for testing the null hypothesis (Mellenbergh, 2005). A second limitation of the MH method is that correct and incorrect responses from both reference and focal groups are required at every level of the matching criterion. This can be problematic in small sample sizes or if there is a substantial difference in ability between the groups because there may be incomplete 2 x 2 cells at the highest and lowest score levels. In addition, MH is appropriate for dichotomous item formats only, although extensions for polytomous item formats have been developed (Zwick, Donoghue, & Grima, 1993). Finally, classical test theory methods such as MH are not appropriate for making comparisons across groups because they are population dependent; therefore MH may not be adequate to provide information about measurement equivalence more generally (Budgell, Raju, & Quartetti, 1995).

2.2.2 Unidimensional Item Response Theory and DIF Methods

In this section I first describe item response theory (IRT) and IRT mathematical models and their advantages for use in LSAs. Following that I introduce IRT DIF methods and Lord's Wald method (Lord, 1977, 1980) that will be used in this study, and discuss IRT model assumptions as well as advantages and limitations of IRT DIF analyses.

There are two basic postulates of IRT. The first is that examinee performance on an item is predicted or explained by underlying examinee factors referred to as traits, latent traits, or abilities. Second, there is a monotonically-increasing function (called an item response function, item characteristic function or item characteristic curve, ICC) that describes the relationship between item performance and the traits underlying the

performance (Hambleton, Swaminathan, & Rogers, 1991). The ICC is an S-shaped trace of the proportion of individuals at the same ability level who answer a given item correctly.

A number of IRT mathematical models describe the ICC. All models contain one or more parameters describing the item and one or more parameters describing the examinee. Typically, one, two or three item parameters may be estimated using IRT methods. They are the difficulty parameter, the discrimination parameter, and the guessing or pseudo-chance parameter. The item difficulty parameter b_i, is defined as the point on an ability scale where the probability of an examinee correctly responding to an item is 0.5. The greater the value of b_i the more difficult the item is. The item discrimination parameter a_i, is proportional to the ICC slope at the point b_i, and provides an indication of an item's capacity to distinguish between examinees of differing ability levels. Items with steep slopes are better at discriminating between examinee ability levels than items with less steep slopes. The guessing or pseudo-change parameter $c_i$represents the probability of examinees with very low ability correctly responding to an item. The c_i parameter provides a lower asymptote of the ICC. The IRT ability parameter theta, θ, represents the unobserved continuous latent trait of an examinee estimated from the data.

The three-parameter logistic (3PL) IRT model is a widely-used general unidimensional model for dichotomously scored tests. It provides an expression for the probability of a correct response to an item as follows:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}} \qquad i = 1, 2, \ldots, n. \qquad \text{Equation 4}$$

where $P_i(\theta)$ is the probability that a randomly chosen examinee with ability θ answers item i correctly, n is the number of items in the test, and a_i, b_i, and c_i are as defined in the preceding paragraph. The 2-parameter logistic (2PL) model is equivalent to the 3PL model

except it does not include the pseudo-guessing parameter c_i. The 1PL model (a Rasch-type model) has neither the pseudo-guessing parameter c_i, nor the discrimination parameter a_i. In other words, in the 1PL model the discrimination and pseudo-guessing parameters are assumed to be constant across items.

In IRT models, the ability and item parameters are estimated from the data, often through the use of maximum likelihood estimation (MLE). MLE maximizes the probability of an observed response pattern with respect to the item and ability parameters. Three commonly used MLE procedures are conditional maximum likelihood (CML), marginal maximum likelihood (MML) and joint maximum likelihood (JML) (Sijtsma & Junker, 2006). The choice of an IRT method may be influenced by the properties of and assumptions required for available estimation procedures and their suitability for the particular measurement situation (DeMars, 2002; Glas & Geerlings, 2009). For example, if data contains scores for two groups that have different ability levels then this should be taken into account when using MML estimation (DeMars, 2002). On the other hand, CML estimation makes no assumption about the ability parameter distribution but can only be used with a 1PL model whereas MML estimation can be used with 1, 2 or 3PL models. An alternative approach to MLE is to obtain Bayesian estimates of the parameters using prior distributions. However, this method is not frequently used in practice because it has the shortcoming of being difficult and time-consuming to use, and requires extra care to ensure valid approximations (Johnson, 2007).

IRT models have practical advantages for use in LSA because they can be used to analyze data from incomplete designs such as BIB designs (Glas & Geerlings, 2009). IRT models allow for horizontal equating of scores across booklets designed to be equally

difficult and where examinees responding to one booklet are assumed to be comparable to those responding to other booklets. For example, it is possible to calibrate multiple booklets simultaneously by estimating item parameters on the examinees who were given the item, and estimating examinee parameters from the items the examinee was administered (DeMars, 2002; Kim & Cohen, 1998). In the case of the BIB booklet design, it is possible to estimate item parameters on a common scale due to the linking or overlap of items between the booklets administered (Lord, 1980).

The identification of DIF items is one application of IRT. The theoretical underpinning of IRT DIF methods is that item parameters should be invariant across groups; they should be the same regardless of the group tested (Lord, 1980). If the item parameters are estimated separately for two groups (and the data fit the IRT model as explained below) then the resulting ICCs should be the same. So, within the IRT framework an item is said to display DIF if its ICC differs across groups, or equivalently, if different parameter values describe the item in each group (Embretson & Reise, 2000; Sireci, Patsula, & Hambleton, 2005). Unlike the MH and LR procedures, IRT DIF approaches do not involve matching groups by conditioning on a variable such as their total scores. Rather, IRT DIF approaches are unconditional analyses; the ability distribution is integrated out in the sense that the area between the ICCs is computed across the entire distribution of the latent variable θ (Zumbo, 2007).

IRT methods for detecting DIF therefore evaluate: whether a common set of item parameters describe the functioning of the item across groups (such as Lord's Wald method; Lord, 1977, 1980); whether the ICCs differ across groups (such as Raju's test of the area between two ICCs; Raju, 1988, 1990); or whether there is an improvement in fit between the

data and the mathematical model underlying the IRT method with or without separate group parameter estimates (Thissen, Steinberg, & Wainer, 1988, 1993). Because Lord's Wald method was used in this study it is described next.

2.2.2.1 IRT-Based Lord's Wald Method

Lord (1977, 1980) observed that an unbiased test is one in which the items have the same ICC for every group. Since the ICC is defined by the item parameters a null hypothesis for DIF would stipulate that the item parameters are invariant across groups. Lord proposed a chi-square test to test the joint difference between the difficulty and discrimination parameters across groups, with the null hypothesis that for a given item, *i*, both $b_{iR} = b_{iF}$ and $a_{iR} = a_{iF}$ (where b_{iR} and b_{iF} represent the difficulty parameter for item *i* in the reference and focal group respectively and a_{iR} and a_{iF} are the respective discrimination parameters in the two groups). The chi-square test is:

$$\chi_i^2 = v_i' \sum_i^{-1} v_i \qquad \text{Equation 5}$$

where v_i is the vector of differences in maximum likelihood estimates of the parameters between the two groups, $\{\hat{a}_{iR} - \hat{a}_{iF},\ \hat{b}_{iR} - \hat{b}_{iF}\}$, and $\sum_i^{-1}$ is the inverse of the asymptotic variance-covariance matrix for the differences $\hat{a}_{iR} - \hat{a}_{iF}\ and\ \hat{b}_{iR} - \hat{b}_{iF}$. The significance test for the joint differences in *a* and *b* parameters, χ_i^2, has a chi-square distribution with 2 degrees of freedom.

Lord's Wald method has recently been improved by Langer (2008). Langer's enhancements include the use of MML estimation (Lord's original test made use of JML estimation), as well as a supplemented EM algorithm (Meng & Rubin 1991, as cited in Langer, 2008) which provides more accurate standard errors, and concurrent estimation of the focal group population parameters.

Lord's Wald method (Lord, 1977, 1980), as enhanced by Langer (2008) has been selected for use in this study over other IRT DIF methods for the following reasons. First, Langer found that this method performed well in terms of power and Type I error rate in detecting DIF in data generated under the 3PL model, which is the model to be used in this study. Second, Lord's Wald method can be extended for use with more than 2 groups. This is an advantage since it is often desirable to compare more than two groups in the types of assessments that contain structurally missing data such as PISA and National Assessment of Educational Progress (NAEP). Last, Lord's Wald test requires only one model fitting (unlike IRT DIF methods that involve the improvement in fit between two models) which is computationally simpler and faster.

2.2.2.2 IRT Assumptions, Advantages and Limitations

There are assumptions about the data to which the IRT model is applied that must be satisfied in order to use IRT-based DIF methods. The assumption of unidimensionality refers to the requirement that the test measure only a single ability, that examinee ability is described by a single latent variable. In practice, this assumption cannot be strictly met and the requirement is relaxed to one of a "dominant" component or factor underlying test performance labelled as "essential unidimensionality", and it has been shown that unidimensional IRT models are robust to moderate violations of unidimensionality (Drasgow & Parsons, 1983). The assumption of local independence states that after taking examinee ability into account the responses to items are statistically independent.

Further, the advantages of IRT models are only gained when there is fit between the mathematical model and the test data. Model fit can be assessed for every item by examining differences between values predicted by the model and observed values and

computing item fit statistics which assess the degree to which the item fits the model. In addition, person fit statistics can be calculated to determine whether individual responses fit the model (Glas & Geerlings, 2009). Last, IRT DIF methods and in particular the 2PL and 3PL methods require large numbers of items and examinees to provide accurate parameter estimates with small standard errors.

Presuming that these assumptions and requirements hold, IRT DIF methods have the advantages of estimating item parameters irrespective of the particular group for whom the data was collected whereas the classical test theory observed score methods of detecting DIF are sample dependent. Another advantage of IRT DIF methods is that they can take differences in item discrimination and pseudo-guessing into account whereas other methods may not. This is an advantage of the IRT DIF methods because differences in item discrimination and guessing have been shown to exist (Angoff, 1993).

As noted by Holland and Thayer (1988) while specific IRT approaches to DIF may be statistically optimal to other approaches in terms of power and efficiency, this is only true when the IRT models actually hold in the data, when the required IRT assumptions and conditions are fulfilled, and (in the case of the 2PL and 3PL methods) when there are an adequate number of items and examinees. For these reasons, and their ease of use, the classical test theory-based approaches such as MH are often preferred in practice (Mellenbergh, 2005).

2.3 Research on Booklet Design and DIF Detection

Although some studies have investigated DIF detection in the presence of non-structurally missing data (for example, H. Finch, 2011; W. H. Finch, 2011; Robitzsch & Rupp, 2009) they will not be reviewed here. The goal of those studies was to investigate

missing data treatments that could be applied to data prior to DIF analysis (for example, listwise deletion or replacing missing data with imputed values) whereas the intent of this study is not to remove or replace missing data but to analyze the data in its original structurally-missing form. In addition, the methods of dealing with missing data in these studies are not suitable for use with incomplete data designs. For example, listwise deletion would result in the deletion of all cases in the data because no examinee is administered all the items. In addition, multiple imputation is only appropriate for up to approximately 30% missing data and can only be used when data is missing on a small number of items (Finch, 2008; Leite & Beretvas, 2010). Both of these limitations make multiple imputation methods unsuitable for analyzing DIF with structurally missing data.

There is a limited amount of research on the performance of DIF methods for structurally missing data from booklet designs. Despite a thorough search of the relevant literature no studies were found that directly investigated the performance of IRT DIF methods for analyzing data from BIB designs. However, other studies that investigated the impact of structurally missing data on other related analyses such as factor analysis and structural equation modeling methods were found. As well, studies on the impact of structurally missing data on MLE and on the estimation of ability and item parameters using IRT were found. This body of literature is reviewed next, followed by a review of the research that has investigated the performance of the MH DIF method on structurally missing data.

2.3.1 IRT DIF Analyses and Booklet Design – Closest Related Studies

Research has investigated the impact of structurally missing data on factor analysis and structural equation modeling methods. Given the established equivalence of the 2PL

IRT model and factor analysis for dichotomous items (Takane & de Leeuw, 1987) and the comparability between structural equation models and 2PL IRT (Willse, Goodman, Allen, & Klaric, 2008), it is possible that the results of these studies may also apply to IRT models although this has not yet been established in the literature.

2.3.1.1 Booklet Design, Factor Analysis and Structural Equation Modeling

Zwick (1987) conducted a simulation study to investigate the impact of BIB design on factor analysis dimensionality assessment. She found that BIB design had little impact on the recovery of dimensionality when compared to a full data set, but that the chi-square goodness of fit statistic was unexpectedly lower for the BIB data than for the complete data. Subsequently, Kaplan (1995) investigated the impact of BIB design on goodness of fit tests for factor analysis of dichotomous items. He found that the chi-square goodness of fit test was sensitive to violations of the distribution assumptions as a result of pairwise available case covariance matrices even when the missing data were MCAR, and concluded that caution should be used in interpreting goodness of fit tests for factor analyses of data from BIB designs.

Structural equation modeling was effectively used in a simulation study to investigate subgroup differences with structurally missing data (Willse et al., 2008). In this research MLE was used to analyze the data without the need to impute missing values, and group differences were examined with a multiple indicators multiple causes model. The estimation of model parameters in the sparse data conditions was comparable to the complete data design. Although there were a limited number of factors considered in this study, the type of sparse data design made only a modest difference to the structural equation modelling analysis.

2.3.1.2 Booklet Design, Maximum Likelihood Estimation and Parameter Estimation

Researchers have investigated the impact of structurally missing data on MLE and on the estimation of ability and item parameters using IRT. As noted earlier, most IRT methods use MLE to estimate item and person parameters. Rubin (1976) showed that for analyzing MCAR data under both direct maximum likelihood and Bayesian estimation procedures, the mechanism that caused the missing data can be ignored and the estimation can proceed without modeling the missing data. In addition, Little and Rubin (1989) note that when data are MCAR it is appropriate to use MLE on the incomplete data to estimate the parameters of the data.

Mislevy and Wu (1988; 1996) investigated the impact of different types of missing data (including structurally missing data) on the estimation of examinee ability parameters when item parameters were known. They also investigated the analogous situation of estimating item parameters when the ability parameters were eliminated by marginalization. Their results indicated that in both cases, the process that caused the missing data is ignorable under direct likelihood and Bayesian inference. The correct value of the MLE was obtained under direct likelihood inference, and sampling distribution inferences based on the MLE were found to be appropriate.

Other studies have been conducted on the impact of incomplete data designs on the estimation of item parameters using IRT. In particular, Eggen and Verhelst (2011) investigated item parameter estimation using both MML and CML and the Rasch model in incomplete testing designs. Although incomplete data can be analyzed with IRT models using several standard IRT software programs (Glas & Geerlings, 2009; Huisman & Molenaar, 2001), Eggen and Verhelst (2011) noted that software programs analyze missing

data by making the assumption that the ignorability principle holds in the data. They investigated whether it is justifiable to make this assumption about ignorability for incomplete data designs for CML and MML methods of estimation. They found that in random incomplete designs the ignorability conditions are met and MML can be applied using the marginal distribution of the observations. However, the ignorability condition cannot be guaranteed for CML estimation so the design mechanism should explicitly enter the analysis. Although Eggen and Verhelst's study was conducted using a 1PL model and dichotomously scored items the results were said to apply to other types of scoring and to IRT models with more than 1 parameter.

DeMars (2002) compared the estimation of item parameters in structurally missing data using JML and MML estimation methods for groups of differing abilities. She found that JML estimation provided accurate estimation of item parameters when group abilities differed. However, MML estimation resulted in either an over- or under-estimation of item parameters. However, it is important to note that the study was conducted in a vertical equating context wherein test forms of different difficulties were administered non-randomly to groups of examinees of different abilities. It is not known whether the results of DeMars' study would apply for structurally missing data due to BIB design where test forms of similar difficulty are administered randomly to examinee groups assumed to have similar ability levels. Nonetheless, it does provide an example of the potential differences in item parameter estimation results due to different estimation methods.

While not directly focused on IRT DIF methods the preceding research may provide insight into the potential impact of structurally missing data on IRT DIF methods because IRT DIF methods utilize MLE and characterize DIF as differences between item parameters

for the groups. If the estimation of item parameters is affected by structurally missing data then it is possible (perhaps even likely) that IRT DIF estimation would also be affected under similar conditions.

2.3.2 MH DIF Analysis and Booklet Design

Structurally missing data can present several challenges for MH DIF analyses. One potential problem is that the blocks of missing data complicate the selection of a matching criterion. The selection of a suitable matching criterion is paramount to obtaining meaningful results from DIF analyses. For the observed score DIF methods such as MH, the usual matching criterion is the sum of scores on all items. However, there are situations in which creating a matching variable based on individuals' total scores in designs such as BIB with structurally missing data may not provide an equivalent matching criterion or measure of proficiency across examinees. For example, the sum of observed scores for examinees who wrote different booklets will be based on different items. Referring to Table 4 again, it can be seen that examinee 1 who wrote booklet 1 and examinee 14 who wrote booklet 7 have the same observed score. However booklet 7 does not contain the same 6 items as booklet 1, therefore the total observed scores of these two examinees may not represent an equivalent measure of proficiency. This is particularly the case if the difficulty levels of the items in the two booklets are not comparable which results in different psychometric characteristics of the booklets.

Another potential source of non-comparability of matching scores across booklets is that booklets may contain different numbers of items and have different maximum total scores. This is the case for the PISA 2006 data described earlier. Two examinees may have equal observed scores that represent different fractions of the total possible score if the total

possible scores are not equal across booklets. In this case if matching occurs on the sum of observed scores the matching variable may not represent similar levels of proficiency.

A further difficulty with the MH DIF method in complex sampling designs such as BIB is the sparseness of data in some of the cells of the 2 x 2 tables when conditioning takes place (Zwick & Ercikan, 1989). Compared to a full data design in which all items are administered to all examinees, there will be fewer examinees at each level of a matching variable. This can have an impact on the functioning of the MH statistics because a minimum number of examinees are required at each level of the matching variable, and score levels that appear in one group but not the other are dropped from the MH calculation. In addition, as noted earlier, correct and incorrect responses from both reference and focal groups are required at every level of the matching criterion. These situations are more likely to occur if groups have different ability levels (for example, few or no examinees in one group may be found at high or low ends of the ability scale depending on group ability) and in small sample sizes (Clauser, Mazor, & Hambleton, 1994).

Allen and Donoghue (1996) investigated a number of ways of creating the matching criterion scores for MH analyses of data from complex item sampling designs. Specifically, they investigated two traditional methods of forming the matching variable ("block-level" matching and "booklet-level" matching) and two alternative methods of forming the matching variable ("pooled booklet matching" and an "extra-information approach"). They compared the MH DIF odds ratio, the transformed log odds and its standard error, and the MH chi-squared statistics for the 4 methods against an analysis of complete data.

Block-level matching involves summing the item scores in one block and comparing all examinees that were administered that block. Booklet-level matching involves creating a

total score for each booklet separately and comparing all examinees that were administered that booklet. Each of these types of matching variables has advantages and disadvantages. Block-level matching has the advantage of producing MH statistics for a greater number of examinees than booklet-level matching because more examinees write blocks than booklets. In the example in Table 4, three times more examinees would be included in block-level matching than in booklet-level matching. Therefore block-level matching may be preferable to booklet-level matching because the MH DIF statistics will have lower sampling variability. However, blocks contain fewer items than booklets so block-level matching is based on fewer items than booklet-level matching. Therefore block-level matching may be less reliable which can adversely affect the MH statistic. If a booklet-level matching variable is used reliability can be increased due to the larger number of items in a booklet than in a block. Another disadvantage to using booklet-level matching is that each item may have multiple measures of DIF associated with it because items can be located in more than one booklet. In the example shown in Table 4, item 1 appeared in three booklets (1, 5 and 7) therefore if using booklet-level matching there will be three different MH statistics associated with that item.

Seeking potentially improved methods to block-level and booklet-level matching for MH analyses of complex item sampling such as BIB designs, Allen and Donoghue (1996) also investigated two alternative approaches to forming the matching variable. The first was the pooled booklet approach in which the MH statistic is based on a concatenation of all 2 x 2 tables in all the booklets an item appears in, i.e. creating 2 x 2 x (k_1+ k_2 +k_3…+ k_n) tables where there are *n* booklets and k represents the number of matching score levels in each booklet the item appears in. While this approach has the advantages of producing a single

MH statistic for each item and using information from all booklets an item appears in, it makes an assumption that the odds ratio is constant across all of the 2 x 2 tables, across all levels of the matching variable and all booklets containing the studied item. The second alternative approach considered by Allen and Donoghue, the extra-information approach, makes the same assumption. In this approach the number-right score for a block is separated from the number-right score for the other blocks in a booklet. The matching variable for each booklet is formed by crossing each level of total score on one block with each level of total score on the other blocks and combining the results. A disadvantage to this approach is that it can lead to very small cell sizes due to the crossing of levels.

Allen and Donoghue (1996) concluded that the results of the pooled booklet approach were superior to the other approaches studied, and were very close to those for the complete data analysis. Therefore they recommended that the pooled booklet approach be used when conducting MH analyses of items from a BIB design. They also suggested that other DIF methods, such as IRT-based procedures, be investigated for use with sparse data designs.

Subsequently Goodman et al. (2011) used Allen and Donoghue's (1996) pooled booklet approach in a simulation study to investigate the Type I error rate, power, and accuracy of MH DIF statistics to detect uniform DIF in three sparse data designs (a BIB design, a common block design, and a non-overlapping matrix design) compared with a complete data design. For the two booklet designs in which items can appear in multiple test booklets (BIB and common block design designs), Goodman et al. used the pooled booklet approach as recommended by Allen and Donoghue. They found that MH was not adversely affected in any of the sparse data designs (as compared to the complete design) when the sample size was large, but power decreased and Type I error rates increased at the lowest

sample size (N=1200), particularly at the longest test length. Of the three incomplete designs, the BIB design had the lowest power at the smallest sample size.

The decrease in power was attributed to the reduced sample size per item in the sparse designs compared to the complete design. Two additional factors that could reduce power in the incomplete designs were also noted. As the test length increased the number of levels of the matching variable increased which resulted in even more sparsely-populated 2 x 2 cells in all three sparse designs. In addition, for the two sparse designs in which items appeared in more than one booklet the pooled booklet approach was used which increased the number of levels of the matching variable by a factor equal to the number of booklets an item appeared in. For example, rather than matching on 2 x 2 x k_1 levels for one booklet, pooled booklet matching took place on 2 x 2 x (k_1+ k_2 +k_3…+ k_n) levels for *n* booklets. The increase in the number of levels resulted in a drastic effect on MH power to detect DIF. Goodman et al. concluded that other methods of forming the matching variable and other DIF methods than the MH should be investigated with sparse data designs.

2.4 Modification to the Pooled Booklet Approach

The results of the Goodman et al. (2011) study motivate further investigation of DIF methods and of the matching criterion for observed score DIF methods in sparse data designs. In this study I modified the pooled booklet approach recommended by Allen and Donoghue (1996) by reducing the number of levels of the matching criterion. Similarly to Allen and Donoghue I pooled information from all booklets an item appeared in. However, rather than matching on 2 x 2 x (k_1+ k_2 +k_3…+ k_n) levels for *n* booklets in which an item appeared I matched on 2 x 2 x k_x levels where k_x represents the average number of matching score levels for the booklets in which an item appeared and the matching criterion for each

examinee was his or her percent correct score on his or her booklet expressed as a proportion of k_x-1 (so that scores could be placed into k_x levels).

Two examples are provided to clarify the modification. In the first example the maximum total scores are the same across all booklets an item appears in and in the second example the maximum possible scores are different across booklets.

First consider the situation in which an item appears in three booklets, each with a maximum score of 25. The number of matching levels per booklet, k, is 26 (representing score levels from 0 to 25). Using Allen and Donoghue's (1996) pooled booklet approach, the number of matching levels for this item would be 2 x 2 x (26+ 26 +26) = 312 and the matching criterion would be the examinees' observed scores. Using the modified pooled booklet approach, k_x would also be 26, the number of matching levels would be 2 x 2 x 26 = 104, and the matching criterion would be the examinees' percent correct scores multiplied by k_x -1 = 25 which in this case is the same as matching on observed scores.

Next consider the situation in which an item appears in three booklets, with different maximum scores of 30, 35 and 40 respectively. The number of matching levels per booklet, k, are 31, 36 and 41. Using Allen and Donoghue's (1996) pooled booklet approach, the number of matching levels for this item would be 2 x 2 x (31+ 36 +41) = 432 and the matching criterion would be the examinees' observed scores. Using the modified pooled booklet approach, k_x would be 36, the number of matching levels would be 2 x 2 x 36 = 144, and the matching criterion would be the examinees' percent correct scores multiplied by 35.

The modified pooled booklet approach has the same advantages as the pooled booklet approach: it produces only one MH statistic for each item regardless of the number of booklets the item appears in, and it uses information from all booklets an item appears in.

The modified approach also has two further potential advantages compared to the pooled booklet approach. First, the modified version greatly reduces the number of levels of the matching criterion which may lead to increased power and decreased Type I error for MH DIF detection. In addition, combining information from all booklets at each level of the contingency table leads to fewer cells with sparse data. Second, because of the conversion of observed scores to percent scores the modified approach can be used to combine item information from booklets that contain different numbers of items or have different maximum score values.

It is important to note that while the modified pooled booklet MH matching criterion may have some advantages, it does not address the first challenge exemplified in Section 2.3.2. That is, the matching criterion may not represent an equivalent measure of proficiency because booklets may not necessarily be psychometrically equivalent. Differences in measurement accuracy may exist across booklets which may lead to differences in reliabilities of the matching criterion.

2.5 Factors that Affect the Performance of DIF Methods

The list of factors that have been shown to affect DIF detection in complete data is extensive. These factors include: sample size; ratio of focal and reference group sample sizes; similarity or differences in group ability and ability distributions; test length; score reliability; number of DIF items; unidimensionality of the test; amount and type of missing data; type of DIF; DIF item discrimination and difficulty levels; effect size of DIF items; and composition and reliability of the matching variable. It is not possible to investigate all factors in one study; therefore I selected five factors: sample size and ratio of focal and reference group sample sizes, test length, percentage of DIF, and true ability differences

between the groups. These five factors were selected because they may help to clarify and extend the results of the previous studies on DIF in structurally missing data.

2.5.1 Total Sample Size

The Allen and Donoghue (1996) and Goodman et al. (2011) studies had somewhat different results. Whereas Allen and Donoghue concluded that the pooled booklet approach performed well in BIB data in all studied conditions, Goodman et al. found that it did not perform well in the smallest sample size. It seems reasonable that the difference in results between these studies may be attributable in part to the different sample sizes investigated in the two studies, especially when combined with the increased number of levels of matching criterion when using the pooled booklet matching approach. Allen and Donoghue used one sample size with unequal reference and focal group examinees (N_R=5100 and N_F =1050) whereas Goodman et al. used sample sizes of 1200, 6000 and 12000 and within each sample size investigated both equal and unequal reference and focal group sizes. Allen and Donoghue did not investigate the performance of the pooled booklet approach in sample sizes as low as 1200, which is the sample size at which Goodman et al. saw the greatest reduction in power for MH to detect DIF in the BIB data.

Further, it is not known from these two studies at what sample size the power begins to decrease. Goodman et al. found MH had adequate power at the 6000 sample size and poor power at the 1200 sample size. Therefore this study investigated sample sizes that include the range of 1200 to 6000 to begin to discern at what sample size the MH method begins to lose power and Type I errors begin to increase when using the pooled booklet approach. By investigating both MH and IRT DIF methods over a range of sample sizes in this study the

goal was to be able to gain insight into which sample size conditions may be optimal for each of the two methods.

2.5.2 Ratio of Group Sample Sizes

In DIF research, it is common to encounter equal reference and focal group sample sizes as well as focal group sizes that are much smaller than reference group sizes. For example, it would normally be expected to have approximately equal sample group sizes when analyzing tests for DIF between males and females. However, there are times when the reference and focal group sizes may be considerably different, such as when examining DIF between English and French-speaking students in the predominantly English-speaking Canadian province of British Columbia. Herrera and Gómez (2008) found that the power of the MH method is significantly affected by the size of the focal and reference groups and by the interaction of group sizes. In addition, unbalanced samples may have unequal variances which has been shown to have an impact on MH DIF analyses (Monahan & Ankenmann, 2005), particularly when the ability distributions of the groups are different, as discussed below. It is important to know how DIF methods perform in both equal and unequal group sizes that reflect real testing situations.

2.5.3 Test Length

Test length has been commonly used as a factor in simulation studies of DIF. In general, longer tests contribute to overall reliability of the test. The use of the observed score for the matching criterion when reliability is low can increase Type I errors. Short and unreliable tests will inflate Type I error rates with the MH method. For example, when the total number of items in a test is less than 20 the MH method has been found to have increased Type I error rates (Donoghue, Holland, & Thayer, 1993; Millsap & Everson,

1993). In contrast, IRT DIF methods, particularly those that use MML estimation, have been found to function well in short tests although they perform better in longer tests. IRT methods that use MML can be employed in tests with as few as 10 to 20 items whereas the MH method may not be suitable for so few items (Bock, 1993).

2.5.4 Percentage of DIF Items

It is likely that in a real testing situation more than one item in a test will contain DIF. There are many examples of analyses of LSA data that have found over 30% of items containing DIF (see, for example, Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; Sireci & Allalouf, 2003). It is important to investigate the effect that the presence of DIF in the remaining items on a test has on the identification of DIF in another item on the test. Studies of the MH procedure have found that MH maintains reasonable power and Type I error rates in cases where up to approximately 10% of the items contain DIF (Miller & Oshima, 1992; Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993) However, when there are 15% or 30% of items that contain DIF the MH procedure has been shown to have insufficient power (Fidalgo, Mellenbergh, & Muñiz, 2000). IRT DIF procedures have also been found to have increased Type I error rates as the percentage of DIF items increases in a test (see, for example, Finch, 2005; Wang, 2004; and Wang & Yeh, 2003).

2.5.5 Group Ability Differences

Zwick (1990) showed that if data follow the Rasch (1PL) model the observed score used in MH produces a good matching for ability when the studied item is included in the matching variable. However, for data that do not follow the 1PL model or when test scores are not highly reliable the MH Type I error rate will be inflated if there is a difference in group ability levels. If data follow the 2PL or 3PL models, the observed scores will not

provide an adequate matching for groups with different ability even if the studied item is included in the matching variable. This is particularly the case in short tests with larger sample sizes, although there is improvement as test length increases (DeMars, 2009).

Given that the 2PL and 3PL IRT models are likely to provide a better fit than the 1PL model for LSA data (Mazzeo & von Davier, n.d.), it is important to investigate the use of DIF methods other than MH (or other observed score methods) with LSA data under conditions in which the group ability levels are different. In addition, Allen and Donoghue (1996) had generated their item parameters based on a 3PL model, and noted that the difference between the method of item generation and the assumptions of the MH procedure may have contributed to a reduction in power for MH in their complete data analysis. This provides further motivation to investigate DIF in BIB data with an IRT-based DIF method.

2.6 Restatement of Research Purpose in Light of the Literature Review

The issues related to identifying DIF in BIB booklet design data are complex. The methods available for identifying DIF are not ideally suited to analyzing data with large missing portions. The only method that has been investigated to date is the MH procedure, which is recommended to be used with a pooled booklet approach. However, this approach resulted in a lack of power for MH to detect DIF at a sample size of 1200, which would often be considered to be a large sample size in DIF applications. Even in the LSA context, sample sizes of less than 1200 are common in DIF analyses. Further, it is reasonable to assume that LSA data would follow a 2PL or 3PL model and that comparison groups may have different ability levels. In this case, MH has been shown to have inflated Type I error rates.

There is a need to investigate other methods of detecting DIF in data arising from BIB designs. One other class of DIF methods, IRT DIF methods, may be appropriate but to date there is no published research that investigates their use in this application. However, research on estimating IRT parameters and the use of maximum likelihood in incomplete data suggests that IRT parameters can be accurately estimated in the presence of BIB data if a marginal maximum likelihood estimation procedure is used.

Therefore this study investigated the use of the MH and the IRT-based Lord's Wald DIF methods on BIB booklet design data. I investigated the MH method using three different approaches: booklet-level matching (referred to in this dissertation as "MH-Booklet"), block-level matching (referred to in this dissertation as "MH-Block"), and modified pooled-booklet matching (referred to in this dissertation as "MH-Modified"). The MH-Modified approach involved making a relatively simple modification to the pooled booklet approach to forming the matching criterion to investigate whether the power and Type I error rates of MH may be improved by reducing the number of levels of the matching criterion. The MH and IRT-based Lord's Wald DIF methods were chosen to enable the study of one observed score method and one latent score method of detecting DIF in BIB data. They also allow a re-examination of one method (the MH) and a new examination of another method (the IRT-based Lord's Wald DIF method, referred to throughout the remainder of this dissertation as "IRT-Lord's") applied to structurally missing data. This study also investigated the effects of sample size, test length, ratio of focal to reference group sample sizes, percentage of DIF items, and true ability differences between the groups on these methods when analyzing data from a BIB design.

In order to carry out these investigations I conducted a series of simulation studies. Simulation studies have the advantage of allowing the researcher to control the data and appraise the effect of manipulated factors in situations in which the truth is known, for example, where it is known which items were simulated to have DIF and which were not. Many simulation studies have investigated the Type I error rates and statistical power of a variety of DIF methods. However, to date, very little research has been conducted to investigate the use of DIF detection methods in situations of incomplete data due to BIB design, and none have investigated the use of an IRT DIF method in these situations.

3 Method

Simulated data was used to investigate the power and Type I error rate of four DIF analysis methods in this study: three MH approaches (MH-Block, MH-Booklet, and MH-Modified), and the IRT-Lord's DIF procedure (Lord, 1977, 1980) which were introduced in Chapter 2. This study also investigated the effects of five factors on the power and Type I error rate of these DIF methods for analyzing structurally missing data: total sample size, test length, ratio of focal to reference group sample sizes, percentage of DIF items, and true ability differences between the groups.

The study design and levels of each factor investigated in the study were selected to reflect real situations in which DIF might be investigated in LSAs that typically use BIB designs such as PISA, NAEP or Trends in International Mathematics and Science Study (TIMSS). For example, the levels of factors are similar to what might be encountered when comparing policy relevant groups within each U.S. state in the NAEP assessment or within each of the Canadian provinces that take part in PISA. Further, to the extent possible the levels of each factor are aligned with both of the previous two studies on which this study builds to allow comparison with the results of those studies.

The following flow chart (Figure 1) provides a broad overview of the study and is intended to assist the reader in following the description of the study design and methods.

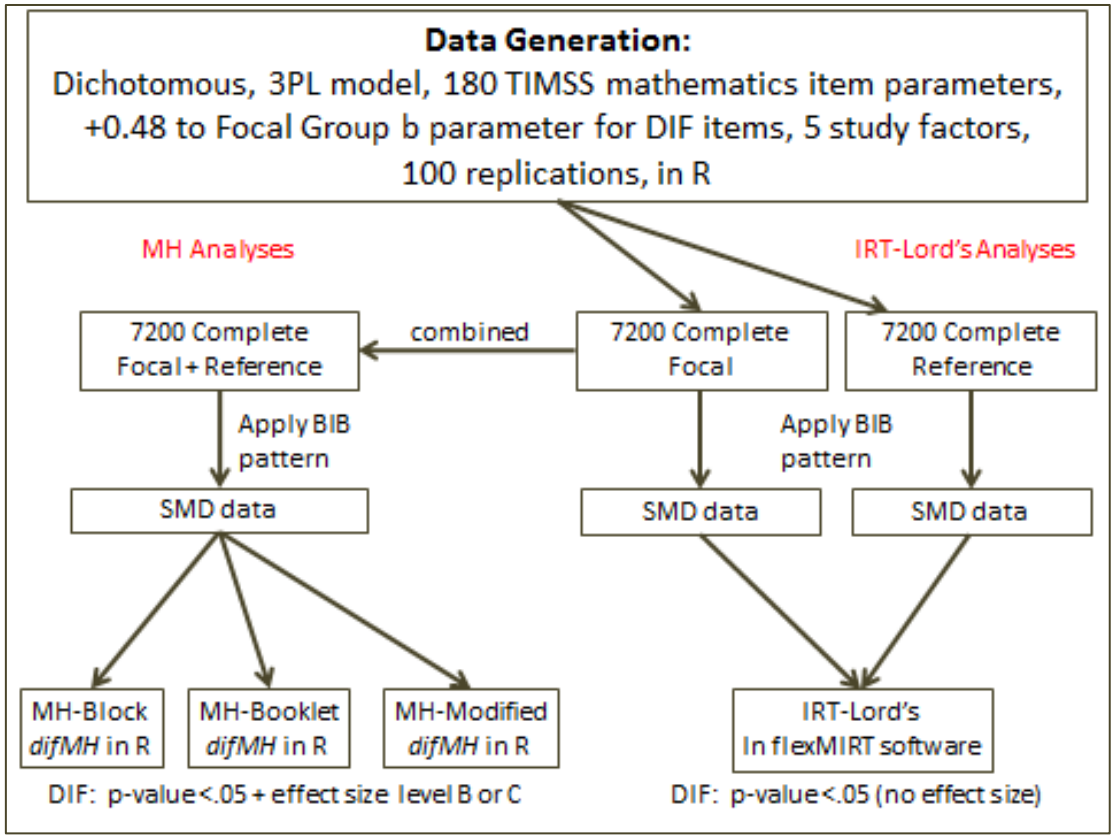

Figure 1: Flow Chart of Study Design

3.1 Booklet Design Simulation

The BIB design simulated in this study is shown in Figure 2. This design consists of a total of 9 blocks distributed in 12 booklets. Each booklet contains 3 blocks, indicated by grey shading.

	Blocks of items in each booklet								
Booklet	1	2	3	4	5	6	7	8	9
1	■						■	■	
2		■	■					■	
3	■		■		■				
4	■	■		■					
5		■			■				■
6			■	■		■			
7				■	■		■		
8				■				■	■
9	■					■			■
10		■				■	■		
11					■	■		■	
12			■				■		■

Figure 2: Simulation Booklet Design

For example, Booklet 1 consisted of blocks 1, 7 and 8, while Booklet 12 consisted of blocks 3, 7 and 9. Each block was contained in 4 booklets, i.e. block 1 was in Booklets 1, 3, 4 and 9. This design was selected because it is the BIB design used by Goodman et al. (2011) (J. T. Goodman, personal communication, June 4, 2012) and it more closely resembles the designs used operationally by LSAs than does the Allen and Donoghue (1996) design which simulated data in seven blocks distributed in three booklets, but had one block common to all booklets. No omitted or not reached items were simulated in this study; therefore the only missing data was structurally missing data.

3.2 Simulation Factors

The effects of five factors on the power and Type I error rate of the MH-Block, MH-Booklet, MH-Modified and IRT-Lord's DIF methods were considered in this study. These factors are: group ability means, with two levels; ratio of group sizes, with two levels; the

percentage of DIF items in the test, with three levels; the total sample size in the assessment, with three levels; and the number of items in each block, with 2 levels.

The two levels of group ability means in this study reflect equal and unequal reference and focal group abilities. In the equal group mean ability condition both groups were simulated to have a mean ability of 0 and standard deviation of 1. In the unequal condition the focal group mean ability was one standard deviation below the reference group mean ability, that is, the reference group had a mean theta of 0 and standard deviation of 1 and the focal group had a mean theta of -1.0 and standard deviation of 1. These levels were selected to represent the situation in which the studied groups may be expected to have similar means (for example, when the reference and focal groups are males and females) and the situation in which the focal group may be expected to have a lower mean (such as when the focal group consists of students with limited proficiency in the assessment language or students with disabilities). Furthermore, mean focal group ability differences are commonly simulated up to 1 standard deviation below the reference group mean (for example, Li, Brooks, & Johanson, 2012).

Similarly, two levels of group size ratios were investigated in this study: equal group sizes in which each group represented 50% of the total sample (the 50/50 condition), and unequal group sizes in which the reference group represented 80% of the total sample size (the 80/20 condition). These relative group sizes were selected to represent typical groups investigated in DIF analyses such as males and females (the 50/50 condition) and students with disabilities or those with limited language ability (the 80/20 condition).

The amount of DIF in the assessment was simulated at three levels: 0%, 10%, or 30% of the items in each block were simulated as DIF items. These levels were selected

because values of between 0% and 30% are commonly studied in DIF simulation studies to represent small to moderate percentages of DIF (for example, Fidalgo, Ferreres, & Muñiz, 2004; Fidalgo, Mellenbergh, & Muñiz, 2000; Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1996; Wiberg, 2009). Further, Goodman et al. (2011) and Allen and Donoghue (1996) varied the levels of DIF between 0% and 30%.

The total sample sizes in the assessment were simulated at 1200, 4800, and 8400. These sample sizes were selected for four reasons. First, they are within the range of the sample sizes investigated in the Goodman et al. (2011) and Allen and Donoghue (1996) studies. Goodman et al.'s results differed most from Allen and Donoghue's results at the 1200 sample size therefore it is important to include that sample size in this study. Second, these sample sizes are aligned with the U.S. state NAEP and Canadian province PISA sample sizes. As examples, the 2007 grade 8 NAEP reading assessment sample size ranged from approximately 1700 to 7100 per state (Vanneman, Hamilton, Baldwin Anderson, & Rahman, 2009), and the Canadian provinces sample sizes in PISA 2006 ranged from approximately 1500 to 4000. Third, in the BIB design used in this study these total sample sizes reflect the range of *per booklet* and *per item* sample sizes (shown in Table 6) of assessments such as NAEP and PISA. As examples, in most countries that participated in PISA 2009 each booklet was administered to between 300 and 700 students, and in PISA 2006 the number of students in each Canadian province who were administered each item ranged from approximately 600 to 1200. Last, the range of sample sizes allowed investigation of differences between the methods at sample sizes where the methods have previously been shown to function differently (for example whereas MH performs well in small sample sizes IRT DIF methods generally require larger sample sizes (Clauser & Mazor, 1998)).

Table 6: ***Simulation Sample Sizes***

Total N	N per booklet	N per item
1200	100	400
4800	400	1600
8400	700	2800

This simulation study investigated either 10 or 20 items per block which reflect a total test length of 30 or 60 items respectively since each booklet contained 3 blocks. These block lengths were again selected to be similar to the two previous studies which used 5, 10 or 20 items per block (Goodman et al., 2011) and 10, 20 or 30 items per block (Allen & Donoghue, 1996). In addition, these block lengths are similar to those used by LSAs such as: NAEP whose block lengths vary from 10 to 20 items with the exception of writing blocks which consist of only 1 item (E. Germino Hausken, personal communication, August 29, 2012); TIMSS whose block lengths range from 10 to 18 items (Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009); and PISA whose average block lengths range from 12 to 17 items (Hopstock, Pelczar, & Xie, 2011).

The overall simulation study design is summarized in Table 7, which shows there are a total of 72 conditions in the study.

Table 7: ***Study Design***

Factor	Levels	Number of Levels
Group ability means	Equal [N(0,1)] or Unequal [Ref = N(0,1) and Foc = N(-1.0,1)]	2
Ratio of group sizes	Equal (50/50) or Unequal (80/20)	2
Percentage of DIF items	0%, 10% or 30%	3
Total sample size	1200, 4800 or 8400	3
Number of items per block	10 or 20	2
TOTAL NUMBER OF STUDY CONDITIONS:	2 X 2 X 3 X 3 X 2 = 72	

A total of 72 conditions were simulated. In each condition, 100 replications were performed for each of the 4 DIF methods, producing a total of 28,800 simulation analyses. Although the recommended minimum number of replications is 25 for simulation studies in IRT-based research (Harwell, Stone, Hsu, & Kirisci, 1996), it is common for DIF simulation studies to have between 50 and 100 replications therefore this study used 100 replications.

3.3 Data Generation

The R language and environment (R Development Core Team, 2012) were utilized to generate the data for this study.

Dichotomously scored data were generated for each study condition according to the 3PL model shown in Equation 4. A different seed was used to generate the data for each replication in each study condition, i.e. 7200 unique seeds were used to generate the data for the study. In order to be representative of real LSA parameters, item parameters for 180

items were randomly selected from the 3PL multiple choice mathematics items published for the TIMSS 2007 assessment (Olson, Martin, & Mullis, 2008). The randomly selected item parameters are shown in Appendix 1. These 180 item parameters were divided into 9 blocks of 20 items for the 20-item per block study condition (i.e. Block 1 contained items 1 to 20 and Block 2 contained items 21 to 40, etc.). For the 10-item per block study condition, the first 10 items of each block in the 20-item per block condition were used (i.e. Block 1 contained items 1 to 10, Block 2 contained items 21 to 30, Block 3 contained items 41-50, etc.). For clarification, the items in the 10-item per block condition are marked with an asterisk in Appendix 1.

In the conditions with DIF, uniform DIF was introduced to 10% or 30% of the items by increasing the value of the difficulty (b) parameter for 10% or 30% of the items for the focal group only. Specifically, DIF was introduced by adding a value of 0.48 to the difficulty parameter for the focal group (i.e. $b_{Focal} = b_{Reference} + 0.48$) for the first 10% or 30% of items in a block. That is, in the 10% DIF and 10-item per block conditions the first item in each block was simulated to contain DIF by adding 0.48 to the b parameter for the focal group. In the 10% DIF and 20-item per block conditions the first two items in each block were simulated to contain DIF. In the 30% DIF and 10-item per block conditions the first 3 items in each block were simulated to contain DIF and in the 30% DIF and 20-item per block conditions the first 6 items in each block were simulated to contain DIF. An increase in the b parameter of 0.48 was selected because it represents a moderate degree of DIF or Level B DIF (Swaminathan & Rogers, 1990) and because it is similar to the values used by Allen and Donoghue (1996) and Goodman et al. (2011) which were 0.5 and 0.4 respectively.

Ability values were randomly selected from normal distributions with means and standard deviations in accordance with each study condition. That is, in the equal ability condition both groups' abilities were randomly selected from a normal distribution with a mean of 0 and standard deviation of 1, i.e., $N(0,1)$. In the unequal ability condition the reference group's abilities were selected from an $N(0,1)$ distribution and the focal group's abilities were selected from a normal distribution with a mean of -1 and standard deviation of 1, i.e. $N(-1,1)$.

Dichotomously scored responses to items under each study condition were obtained as follows: (1) the probability of a correct response for each simulee to each item was obtained using Equation 4 with the ability and item parameters described in the previous paragraphs; (2) this probability was compared to a random sample generated from a Bernoulli distribution with probability equal to the simulee's probability of a correct response; (3) if the simulee's probability of a correct response was equal to or greater than the random sample generated from the Bernoulli distribution the simulee was assigned a score of 1 for the item (i.e. was scored as if correct); and (4) if the simulee's probability of a correct response was less than the random sample generated from the Bernoulli distribution the simulee was assigned a score of 0 for the item (i.e. incorrect). This process was repeated for 100 replications of each of the 72 study conditions. This resulted in 7200 "complete data sets" (containing values of 1 for correct responses and 0 for incorrect responses) that did not contain any blocks of missing data. Subsequently, scores for blocks of items were replaced with missing data notation ("NA") in accordance with the BIB study design shown in Figure 2, resulting in 7200 "BIB data sets".

3.4 DIF Analyses

In order to explain the four DIF analysis methods used I provide an example of how each analysis was conducted using one of the 72 study conditions as an example (a condition in which the group ability means are equal, the ratio of group sizes is equal, 10% of the items are simulated to contain DIF, the total sample size is 4800, and there are 20 items per block). The simulation booklet design that represents this condition is shown in Figure 3:

	Blocks (items) [DIF items] in each booklet								
Booklet (Examinees*)	1 (1-20) [1,2]	2 (21-40) [21,22]	3 (41-60) [41,42]	4 (61-80) [61,62]	5 (81-100) [81,82]	6 (101-120) [101,102]	7 (121-140) [121,122]	8 (141-160) [141,142]	9 (161-180) [161,162]
1 (1-400)									
2 (401-800)									
3 (801-1200)									
4 (1201-1600)									
5 (1601-2000)									
6 (2001-2400)									
7 (2401-2800)									
8 (2801-3200)									
9 (3201-3600)									
10 (3601-4000)									
11 (4001-4400)									
12 (4401-4800)									

Figure 3: Sample Simulation Booklet Design for Conditions in which Total Sample Size is 4800, 20 Items per Block, and 10% of Items are DIF

*in this condition reference and focal group ability levels are equal and reference and focal group sample sizes are equal

Figure 3 is similar to the simulation booklet design shown in Figure 2 except that more information is provided to reflect the specific study condition used in the following examples. Similarly to Figure 2, the blocks that were contained in each booklet are indicated by grey shading. However, in Figure 3 the specific items contained in each block are shown in the 2nd row in parentheses and the items that were simulated to contain DIF in each block are shown in square brackets. Also, the examinees who were administered each booklet are shown in the first column. The data for this example was simulated such that each booklet contained an equal number of reference and focal group members and the reference and focal group ability levels were simulated to be equal (theta = 0).

3.4.1 MH Analyses

The R language and environment (R Development Core Team, 2012) was used to conduct all aspects of the MH-Block, MH-Booklet and MH-Modified analyses. Specifically, the *difMH* method in the *difR* package (Magis, Beland, & Raiche, 2012) was used in its original form for the MH-Block and MH-Booklet analyses and was adapted for the MH-Modified analyses.

In the three MH analyses, items were identified as DIF if they exhibited Level B or Level C DIF using the criteria outlined in Zwick and Ercikan (1989). That is, items were identified as DIF if they had $|\Delta_{MH}| \geq 1.0$ and a significant MH chi-square statistic at $p < .05$. It is important to note that this represents a blended heuristic decision rule for identifying DIF items (Gómez-Benito et al., 2013; Zumbo, 2008) as described in Section 2.2.1.1. This method of identifying DIF items was chosen because it is often used in practice, and more specifically because it is the method used by Goodman et al. (2011) which allowed comparison of results between this study and Goodman et al.

To explain the three types of MH analyses conducted, it is helpful to consider the differences between them in terms of 4 specific factors: which examinees' responses were included in the analysis, which item responses were analyzed, which item responses contributed to the score used as the matching variable in the analysis, and the number of levels of the matching criterion.

3.4.1.1 MH-Block Analyses

In block level matching the MH analysis was conducted "block-wise" or one block at a time. The MH statistic was computed based on the responses of examinees that were administered the items in the particular block being analyzed. All of the items in that block were included in the MH analysis. The sum of scores on the items in that block was computed for each of those examinees and used as the MH matching criterion. The matching criterion was divided into as many levels as there were possible score values for the block. Using the condition exemplified in Figure 3 and considering block 1 as an example, responses to items 1-20 of the 1600 examinees who were administered booklets that contained block 1 (examinees numbered 1-400, 801-1200, 1201-1600 and 3201-3600) were included in the DIF analysis. Examinees were matched based on their sum score on block 1 items only (that is, the matching criterion was the total score on items 1-20), and the number of levels of the matching criterion was 84, representing 2 x 2 tables formed over 21 possible score levels (since the possible scores for block 1 range from 0 to 20).

3.4.1.2 MH-Booklet Analyses

In the MH-Booklet approach the MH analysis was conducted "booklet-wise" or one booklet at a time. The MH statistic was computed based on the responses of only those examinees that were administered the items in a particular booklet and all of the items in the

booklet were included in the DIF analysis. The sum of scores on all of the items in that booklet was computed for each of the examinees and used as the MH matching criterion. The matching criterion was divided into as many levels as there were possible score values for the booklet. Referring to Figure 3 and considering booklet 2 as an example, responses to items 21-40, 41-60 and 141-160 of the 400 examinees who were administered booklet 2 (examinees numbered 401-800) were included in the DIF analysis. Examinees were matched based on their sum score on booklet 2 items (that is, the matching criterion was the total score on all items in booklet 2), and the number of levels of the matching criterion was 244, representing 2 x 2 tables formed over 61 possible score levels (since the possible scores for booklet 2 range from 0 to 60).

3.4.1.3 MH-Modified Analyses

The modified pooled booklet approach can be conceptualized as being "booklet-wise" in terms of which item scores comprised the matching criterion and "block-wise" in terms of which examinees and which items were included in the analyses. The MH statistic was computed based on the responses of all examinees that were administered the items in a particular block and all of the items in that block were included in the DIF analysis. However, the matching criterion was composed of the percent correct scores on all of the items that were administered to each examinee (that is, their percent correct score based on all items in the booklet that they were administered). The matching criterion was divided into the average number of matching score levels for all booklets in which the block appeared. Again, referring to Figure 3 and considering the items in block 2 as an example, responses to items 21-40 of the 1600 examinees who were administered booklets 2, 4, 5 and 10 (examinees numbered 401-800, 1201-1600, 1601-2000 and 3601-4000) were included in

the DIF analysis. Examinees were matched based on their percent correct score on all items that they were administered (that is, the matching criterion was their percent correct score on all items in their booklet), and the number of levels of the matching criterion was 244, representing 2 x 2 tables formed over 61 possible score levels (since each booklet contained 60 items the average number of levels was 61). It is worth noting that the significant differences between the MH-Block method and the MH-Modified method are the composition of and number of levels of the matching criterion.

3.4.2 IRT-Lord's DIF Analyses

flexMIRT (Cai, 2012) was batched through R and used to conduct the IRT-Lord's DIF analysis. flexMIRT is capable of handling multiple groups and multilevel data such as LSA data (Houts & Cai, 2012), including structurally missing data due to BIB design (M. Edwards, personal communication, May 30, 2012). I conducted a "DIF sweep" analysis using an assumed group invariant model in flexMIRT. Because nonrandom groups such as those typically investigated in DIF analyses may have different population means it is necessary to have a set of anchor items to link the groups' scores to test group differences. A DIF sweep represents an "all-other" procedure in which all items except the item under study are used as anchor items (this is also referred to as the "Wald-2" test, Langer, 2008). In a DIF sweep all items are initially constrained to be equal across groups to obtain conditional population distribution estimates and then each item is freed one at a time and tested for DIF utilizing the Wald test as enhanced by Langer (2008).

flexMIRT allows placing constraints and prior distributions on item parameters. Prior distributions of N(0, 1.5) for the *b* parameter, N(1, 1.5) for the *a* parameter and N(-1.39, 0.5) for the *c* parameter were used in this study. The values for the prior of the

means for the *a* and *b* parameters were selected because these values are typical mean values for large-scale assessments composed of multiple choice items. The prior values for the standard deviations for the *a* and *b* parameters were selected because they are considerably larger than those typically observed in large-scale assessments and thus would not overly guide the parameter estimation process. The prior values for the mean and standard deviation of the *c* parameter were selected because they represent realistic values for the lower asymptote in a 3PL model (Houts & Cai, 2012).

flexMIRT was used to conduct an analysis of a test of DIF in the *b* parameter only (referred to as "*b*-DIF"). This was done in order to be similar to the MH analyses of uniform DIF conducted in this study since *b*-DIF represents uniform DIF. Further an analysis of *b*-DIF was congruent with the method of DIF simulation wherein DIF was simulated in the *b* parameter only.

The IRT-Lord's analyses were conducted by passing all data for all examinees in one replication of a given condition to flexMIRT. Using Figure 3 as an example, data for items 1-161 for all 2400 reference group examinees and data for items 1-161 for all 2400 focal group examinees were submitted separately to flexMIRT as required by the flexMIRT software. Scores for items that were not administered to examinees were indicated as missing. For example, for examinees that were administered booklet 1, scores for all items in blocks 2, 3, 4, 5, 6 and 9 were indicated as missing. The analyses were conducted as explained above. Items were identified as DIF if they had a significant *b*-DIF chi-square statistic at $p < .05$. An effect size measure was not used because no empirically validated effect size measure for the IRT-Lord's DIF method currently exists (Kim & Oshima, 2013).

3.5 flexMIRT Item Parameter Recovery

The results of the IRT-Lord's DIF analyses in this study were dependent on the ability to accurately recover the true difficulty (*b*) parameter from the simulated data. Therefore, prior to conducting the IRT-Lord's DIF analyses a small separate study was carried out to check the flexMIRT accuracy of recovering difficulty parameters from the generated data. One hundred complete data sets with no structurally missing data were generated for a fully crossed 2 (number of items: 90 or 180) x 5 (sample size: 240, 600, 1200, 4800 or 8400) study using the same item parameters and data generation process as the main study. These sample sizes were selected to cover the range of per group and combined group sample sizes used in the main study (for example, in the 1200 sample size and 80/20 sample size ratio condition there would be 240 examinees in the focal group). For all datasets, a single group ability mean and standard deviation of 0 and 1 respectively were used to generate the data. The item parameters for each data set were estimated using a single-group 3PL model calibration in flexMIRT. The parameter recovery study analysis used the same constraints and prior distributions described in Section 3.4 for use in the main study. For each of the ten parameter recovery study conditions the item difficulty parameters estimated by flexMIRT were compared to the known item difficulty parameters by calculating the bias and root mean square error (RMSE) for each item in each study condition, then averaging over items and over the 100 replications for each study condition:

$$Bias_i = \frac{1}{n}\sum_1^n\left(\hat{b}_{ij} - b_j\right) \qquad \text{Equation 6}$$

$$RMSE_i = \sqrt{\frac{1}{n}\sum_1^n\left(\hat{b}_{ij} - b_j\right)^2} \qquad \text{Equation 7}$$

where n represents the number of replications (100 in this study), $\hat{b}_{ij}$ represents the estimated item parameter for item j in replication i, and b_j represents the true item parameter for item j. This resulted in either 90 or 180 bias and RMSE values per study condition depending on whether there were 90 or 180 items in the condition. These 90 or 180 values were averaged to produce an overall summary bias statistic and an overall summary RMSE statistic for each parameter recovery study condition.

3.6 Outcome Measures: Type I Error Rate and Power

For each of the four DIF methods the Type I error (the incorrect identification of an item as a DIF item or "false detection") rate and power (the correct identification of a known DIF item or "true detection") rate were calculated for each of the study conditions in the following manner. As noted earlier, for the MH analyses a blended Type I error decision rule based on statistical significance and effect size was used, and items were identified as DIF if they exhibited Level B or Level C DIF using the criteria outlined by Zwick and Ercikan (1989). For the IRT-Lord's DIF analyses items were identified as DIF if they had a significant *b*-DIF chi-square statistic at $p < .05$.

Although it is also possible to consider bias of the MH and IRT-Lord's DIF estimators, this study focussed on significance tests as it was the intention of the study to present a thorough examination of this specific element of DIF detection practice. While future research may examine the effects on DIF effect magnitudes, such investigations fall outside the scope of the current study.

For all analysis methods the Type I error and power rates were calculated *per item* and averaged across replications to produce one Type I error and power statistic per condition; however, it was necessary to adjust the manner in which this was carried out in

order to account for the different ways in which the analyses were conducted. Again, the condition exemplified in Figure 3 is used to motivate the descriptions of the analyses of Type I error and power rates for the four DIF methods.

3.6.1 MH-Block and MH-Modified Analyses

Type I error and power analyses for MH-Block and MH-Modified proceeded in the same manner since these DIF analyses were conducted similarly except for the composition and number of levels of the matching variable (that is, although the matching variable was different the DIF statistics were based on "block-wise" analyses of items). For each DIF item, power was computed as the proportion of times the item was correctly identified as having DIF in the 100 replications for each study condition. The per item power statistics were averaged across DIF items in each block and the blocks were averaged to result in one power statistic for each study condition. Similarly, for each non-DIF item, Type I error rate was computed as the proportion of times the item was incorrectly identified as having Level B or C DIF in the 100 replications for each study condition, the per item Type I error statistics were averaged across all non-DIF items in each block and the blocks were averaged to result in one Type I error statistic for each study condition. Using the study condition exemplified in Figure 3 as an example, power was calculated in the following manner. In block 1, items 1 and 2 were known to be DIF items due to the study design. The proportion of times item 1 was identified as DIF was calculated across the 100 replications of this study condition, the proportion of times item 2 was identified as DIF was calculated across the 100 replications of this study condition, and these two proportions were averaged to produce a power result for block 1 of this study condition. Similar power results were obtained for each of the remaining 8 blocks (i.e. items 21 and 22 in block 2, items 41 and 42 in block 3,

etc.) and the results for all 9 blocks were averaged to arrive at an overall power statistic for this study condition. Type I error rate was calculated similarly. The proportions of times the known non-DIF items in block 1 (items 3-20) were incorrectly identified as DIF items across the 100 replications were calculated and averaged to produce a Type I error result for block 1. Similar Type I error results were calculated for each of the remaining 8 blocks (i.e. using items 23-40 for block 2, items 43-60 for block 3, etc.) and the results for all 9 blocks were averaged to arrive at an overall Type I error rate for this study condition.

3.6.2 MH-Booklet Analyses

For the MH-Booklet analyses, power was computed as the proportion of times each known DIF item in a booklet was correctly identified as DIF in the 100 replications for each study condition. The per item power statistics were averaged across DIF items in each booklet and the power statistics for the 12 booklets were averaged to result in one power statistic for each study condition. Similar analyses of non-DIF items in each booklet were conducted to compute a Type I error rate for each non-DIF item across the 100 replications for each study condition. These were averaged to produce a Type I error rate for the booklet and the Type I error rates for the 12 booklets were averaged to result in one Type I error rate per study condition. Using Figure 3 as an example, power was calculated in the following manner. In booklet 1, items 1, 2, 121, 122, 141 and 142 were known to be DIF items due to the study design. The proportion of times each of these items was identified as a DIF item was calculated across the 100 replications of this study condition, and these proportions were averaged to produce a power result for booklet 1 of this study condition. Similar power results were obtained for each of the remaining 11 booklets (i.e. items 21, 22, 41, 42, 141 and 142 for booklet 2, etc.) and the results for all 12 booklets were averaged to arrive at an

overall power statistic for this study condition. Type I error rate was calculated similarly. The proportions of times the known non-DIF items in booklet 1 (items 3-20, 123-140 and 143-160) were incorrectly identified as DIF items across the 100 replications were calculated and averaged to produce a Type I error result for booklet 1. Similar Type I error results were calculated for each of the remaining 11 booklets and the results for all 12 booklets were averaged to arrive at an overall Type I error rate for this study condition.

3.6.3 IRT-Lord's Analyses

The computations for Type I error and power rates for the Lord's Wald analyses were straightforward because all items were included in every analysis. For power analysis, the proportion of times each known DIF item was correctly identified as DIF across the 100 replications of a study condition was calculated and averaged across DIF items to result in one power statistic for each study condition. Similarly, the proportion of times each known non-DIF item was incorrectly identified as a DIF item across the 100 replications was calculated and averaged across items to result in one Type I error statistic for each study condition. Following the example of Figure 3, power was calculated in the following manner. The proportion of times each of the DIF items (items 1, 2, 21, 22, 41, 42, 61, 62, 81, 82, 101, 102, 121, 122, 141, 142, 161 and 162) was correctly identified as a DIF item was calculated across the 100 replications of this study condition and averaged across DIF items to produce one power statistic for this study condition. Similarly, the proportion of times each of the non-DIF items (all remaining items) was incorrectly identified as a DIF item was calculated across the 100 replications of this study condition and averaged to produce one Type I error statistic for this study condition.

3.6.4 Inflation in Type I Error Rate

The Type I error rates for each DIF analysis method for each study condition were examined to determine whether they were within acceptable limits or inflated. Inflation in Type I error rate was documented according to the Bradley (1978, as cited in Li & Zumbo, 2009) approach. For a Type I error rate of .05 Bradley categorized empirical Type I error rates between .045 and .055 as "fairly stringent", those between .040 and .060 as "moderate", and those between .025 and .075 as "very liberal". In this study Bradley's "moderate" criterion was used as an upper cut-off to distinguish Type I error inflation. That is, conditions in which the Type I error rate exceeded .060 were categorized as inflated in this study.

In each study condition where the Type I error rate exceeded the moderate criterion of .060 the power result was not interpreted because an inflated Type I error renders the associated power spurious in that study condition.

3.7 Verification of Simulation Procedures

Verification of the simulation procedures is a critical step in a simulation study; however every simulation is unique and no single set of verification procedures will apply in every simulation (Sargent, 2005). This study followed steps suggested by Bratley, Fox and Schrage (1987) to ensure the simulation was occurring as anticipated and the output values were reasonable. A full list of the specific verification procedures used and their purposes are provided in Appendix 2.

R code was written to check that all data files created had the expected dimensions and to create warnings if files were found with unexpected dimensions. For example, code was written to ensure that the number of rows in the BIB data for each study condition was

equal to the sample size corresponding to the study condition and that the number of columns was equal to the number of items corresponding to the study condition. R code was also written to check that data values were consistent with study conditions. For example, code was written to verify that the *b* parameters for items without DIF were equal to the original generating *b* parameters, and that the *b* parameters for items with DIF were equal to the original *b* parameter values plus 0.48 if and only if the item was indexed as a DIF item.

Manual verification was used to ensure that the same results obtained in this simulation study would also be obtained using hand calculations or other software analyses. A series of visual checks were conducted on randomly selected data files to verify their anticipated structure. Visual checks were also conducted on randomly selected flexMIRT output files to review the output for reasonableness (such as reasonable estimations of group theta values according to the study conditions) and to ensure there were no error messages or warnings related to failure to converge or to estimate item parameters. Excel was used to calculate and verify values for the matching variable in the MH-Modified analyses in randomly selected data files. MH DIF analyses were conducted on randomly selected data files using the DIFAS software (Penfield, 2005) as a means of verifying the R codes used for the three MH methods used in this study. Finally, calculations of power and Type I error rates were calculated in Excel on randomly selected output files for each DIF method and compared to the power and Type I error rates obtained through the simulation process.

4 Results

This simulation study investigated the Type I error rate and power of three approaches to the MH DIF analysis method and the IRT-based Lord's DIF procedure (Lord, 1977, 1980) to identify DIF in structurally missing data. The three MH approaches used were: MH-Block, MH-Booklet and MH-Modified. This study also investigated the effects of five factors (group ability means, ratio of group sizes, percentage of DIF items, total sample size and number of items per block) on the Type I error rate and power of the studied DIF methods for analyzing DIF in structurally missing data.

In this chapter the results of all aspects of the study are presented, beginning with the results of the simulation verification procedures, descriptive statistics for the item parameters used to generate the data, and the results of the preliminary item parameter recovery study which was conducted to ensure that the flexMIRT software could accurately recover the difficulty (*b*) parameter from the simulated data. Following this, the Type I error and power results of the main simulation study are presented. Finally, the Type I error and power of the MH-Modified method are compared to Type I error and power results of the MH pooled booklet method reported by Goodman et al. (2011).

4.1 Verification of Simulation Procedures

The results produced from the R code for checking data dimensions and data values confirmed that all data files had the correct dimensions and data values. In addition, visual checks of randomly selected data file structures revealed that data files were constructed as anticipated; that is, files had the correct number of items and examinees according to the study conditions and had missing data only where anticipated according to the simulation booklet design shown in Figure 2. Visual checks of randomly selected flexMIRT output files

confirmed the output of the IRT-Lord's analysis was reasonable according to the study conditions and that there were no error messages or warnings related to failure to converge or to estimate item parameters. Excel calculations of randomly selected data files for the MH-Modified analyses verified the values for the matching variables were correctly calculated. MH DIF analyses of randomly selected data files using the DIFAS software (Penfield, 2005) verified the results obtained in the simulation. Finally, calculations of power and Type I error rates on randomly selected output files using Excel verified those obtained through the simulation process.

4.2 Descriptive Statistics of Item Parameters used to Generate the Data

The item parameters used in generating data in this simulation study are described in Table 8 and Figure 4 below. The extent to which these parameters are representative of item parameters in typical LSAs may influence the degree to which the results of this study might generalize to other LSA situations.

The *a* parameter distribution was moderately positively skewed and one item (item number 69) was found to be an outlier with a value of 2.37. The *b* parameter distribution was slightly negatively skewed with no outlier values. The *c* parameter distribution was approximately symmetric with no outlier values. The means and standard deviations for the *a*, *b* and *c* parameters are similar to those typically found in LSAs therefore with the possible exception of item 69 the item parameters are representative of those typically found in LSAs.

Table 8: ***Descriptive Statistics of 180 Item Parameters***

	Parameter		
	a	*b*	*c*
Mean	1.08	-0.01	0.15
Standard Deviation	0.31	0.64	0.05
Median	1.07	0.08	0.15
Minimum	0.44	-1.81	0.03
Maximum	2.37	1.55	0.26
Range	1.93	3.36	0.23
Skewness	0.56	-0.40	-0.15
Kurtosis	0.85	-0.16	-0.11
Standard error of the mean	0.02	0.05	0.00

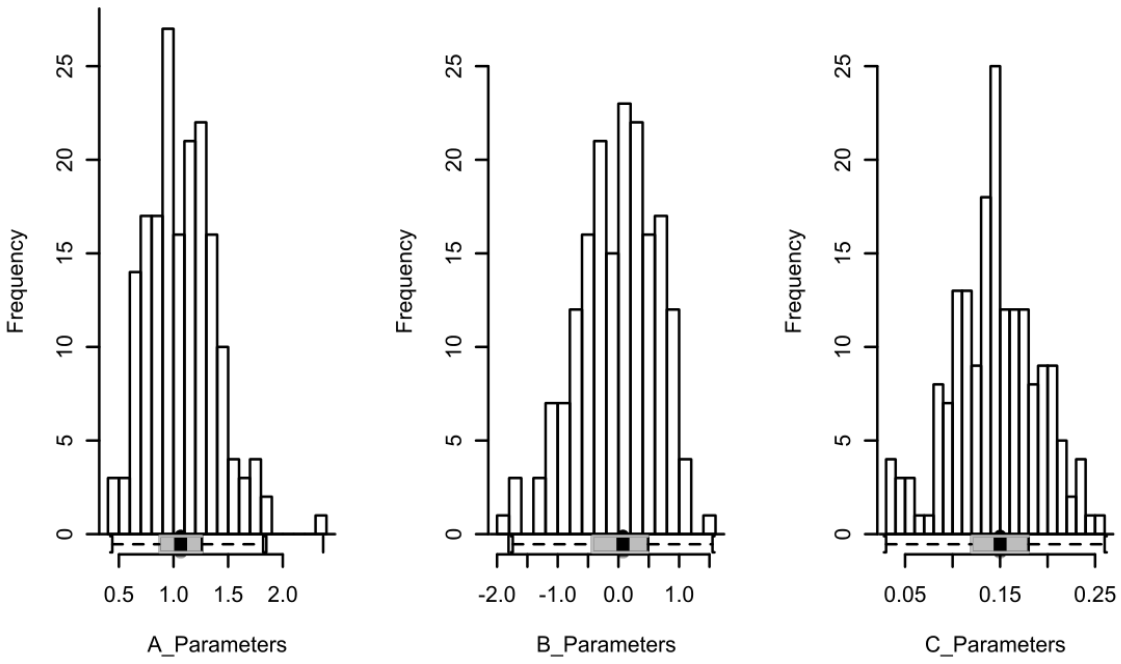

Figure 4: Histograms with Boxplots of Item Parameters Used to Generate the Data

4.3 flexMIRT Item Parameter Recovery

Bias and RMSE analyses were conducted to investigate the accuracy of flexMIRT to recover the known item difficulty (*b*) parameters in each of ten item parameter recovery conditions as described in Section 3.5. Results are shown in Table 9. For the 180-item test size (representing the 20 items per block main study condition) average bias values were very small and ranged from 0.007 to 0.036. For the 90-item total test size (representing the 10 items per block main study condition) average bias values were small to medium, ranging from 0.037 to 0.079. In all sample size conditions bias was greater for the 90-item test size than for the 180-item test size. In addition, in all conditions the bias was positive indicating a consistent slight overestimation of the difficulty parameter.

A general trend of decreasing RMSE with increasing sample size is seen in both the 90-item and the 180-item conditions. Across all sample sizes the RMSE value is greater in the 90-item conditions than in the 180-item conditions. The RMSE values for the 90-item conditions ranged from 0.124 to 0.216 and the RMSE values for the 180-item conditions ranged from 0.053 to 0.102. These values are similar to those found in other studies of *b* parameter recovery with a 3PL IRT model (see, for example, Yoes, 1995) and provide evidence that flexMIRT accurately estimated the *b* parameters in a complete data set in sample size conditions similar to those investigated in the main study of this dissertation.

Table 9: ***Parameter Recovery Study: Bias and RMSE Results***

		90 Items					180 Items				
		N=240	N=600	N=1200	N=4800	N=8400	N=240	N=600	N=1200	N=4800	N=8400
Bias	Mean	0.074	0.079	0.075	0.048	0.037	0.007	0.025	0.036	0.030	0.023
	Standard deviation	0.096	0.058	0.048	0.036	0.031	0.111	0.056	0.039	0.030	0.025
	Median	0.084	0.067	0.061	0.038	0.029	0.011	0.025	0.029	0.018	0.013
	Minimum	-0.239	-0.043	-0.001	-0.003	-0.004	-0.404	-0.170	-0.044	-0.004	-0.004
	Maximum	0.299	0.235	0.195	0.148	0.157	0.283	0.195	0.190	0.136	0.133
	Standard error of the mean	0.010	0.006	0.005	0.004	0.003	0.008	0.004	0.003	0.002	0.002
RMSE	Mean	0.216	0.204	0.151	0.130	0.124	0.102	0.080	0.066	0.064	0.053
	Standard deviation	0.088	0.103	0.052	0.045	0.047	0.038	0.041	0.035	0.037	0.031
	Median	0.206	0.176	0.135	0.118	0.112	0.091	0.066	0.054	0.054	0.040
	Minimum	0.110	0.093	0.075	0.064	0.057	0.049	0.029	0.027	0.020	0.020
	Maximum	0.699	0.951	0.328	0.290	0.245	0.256	0.225	0.194	0.219	0.193
	Standard error of the mean	0.009	0.008	0.006	0.003	0.005	0.003	0.004	0.003	0.004	0.002

4.4 Main Study Results

Type I error rates were calculated for each of the four DIF methods for each of the 72 study conditions and power rates were calculated for each of the four DIF methods for the 48 study conditions in which DIF was simulated. As noted previously, both Type I error and power were calculated per item and averaged across items and replications to produce one Type I error or power statistic per DIF method per condition. Inflation in Type I error was assessed using the Bradley (1978) moderate criteria: empirical Type I error rates that exceeded .060 were categorized as inflated. Power was interpreted in only those study conditions in which Type I error was not inflated, that is, in conditions in which the Type I error did not exceed Bradley's moderate criterion.

4.4.1 Failure to Obtain Simulation DIF Results in Some MH-Booklet Analyses

Initially there were 19 study conditions for the MH-Booklet analyses in which the Type I error and power had not been calculated during the simulation and "NA" was returned as the analysis result. There are three factors that may have contributed to this outcome. As noted in Sections 2.2.1.2 and 2.3.2, one limitation of the MH method is correct and incorrect responses from both reference and focal groups are required at every level of the matching criterion. When this condition is not met, the $\widehat{\alpha}_{MH}$ odds ratio (Equation 1) or the MH_{χ^2} ratio (Equation 3) may have improper solutions or may become undefined as a result of having a zero denominator. This is particularly likely to happen in small sample sizes. In addition, complex sampling designs such as BIB designs result in more sparse data in some cells of the 2 x 2 table as compared to full data designs. Last, as illustrated in the examples provided in Sections 3.4.1.1 through 3.4.1.3, of the three MH methods used in this study MH-Booklet is

the method in which the MH analysis is based on the smallest number of examinees and therefore is the most likely of the three methods to have incomplete 2 x 2 cells.

Examination of the *difMH* results for MH-Booklet analyses of these 19 study conditions revealed that there were some items in some replications for which the $\hat{\alpha}_{MH}$ odds ratio or the MH_{χ^2} ratio had improper solutions or became undefined and the MH DIF result was returned as "NA". This led to Type I error and power results not being calculated and being returned as "NA". The study conditions in which the Type I error and power for the MH-Booklet method were not calculated during the simulation and the number of times this occurred within each condition are shown in Table 10.

Table 10: ***Study Conditions in which the MH-Booklet Type I Error and Power were not Calculated during Simulation***

Condition	Focal Group Sample Size Ratio	Focal Group Theta	Proportion of DIF Items	Block Size	Sample Size	Total # of NA occurrences in this condition
1	0.2	-1	0	10	1200	3
2	0.2	-1	0	20	1200	44
3	0.2	-1	0.1	10	1200	43
4	0.2	-1	0.1	20	1200	1
5	0.2	-1	0.3	10	1200	2
6	0.2	-1	0.3	20	1200	63
7	0.2	0	0	10	1200	8
8	0.2	0	0	20	1200	58
9	0.2	0	0.1	20	1200	85
10	0.2	0	0.3	10	1200	4
11	0.2	0	0.3	20	1200	56
12	0.5	-1	0	20	1200	5
13	0.5	-1	0.1	10	1200	1
14	0.5	-1	0.1	20	1200	1
15	0.5	-1	0.3	20	1200	4
16	0.5	0	0	10	1200	2
17	0.5	0	0	20	1200	5
18	0.5	0	0.1	20	1200	5
19	0.5	0	0.3	20	1200	8

As shown in Table 10, the MH-Booklet Type I error and power were not calculated in the smallest sample size (1200) only. This situation occurred with much greater frequency in conditions in which the group sample sizes were unequal (367 of 398 occurrences, or 92% of occurrences) than in conditions in which the group sample sizes were equal (31 of 398 occurrences, or 8% of occurrences). This could be anticipated because in these conditions the total sample size was 1200, the total sample size per booklet was 100, the reference group sample size per booklet was 80, and the focal group sample size per booklet was only 20. Given such small sample sizes, particularly for the focal group, it was less likely that the

requirement for correct and incorrect responses from both the reference and focal group at every level of the matching criterion would have been met and therefore the MH DIF result would not be returned.

There were a total of 28,800 replications for the MH-Booklet analyses in the 1200 sample size conditions (12 booklets analyzed x 24 study conditions involving sample size of 1200 x 100 replications per condition = 28,800). In 398 of those replications (1.4% of the total replications involving the sample size of 1200) the Type I error and/or power were not calculated for at least one item using the MH-Booklet method. Table 10 reveals that the frequency with which this occurred did not appear to vary systematically with the focal group theta, the proportion of DIF items or the number of items in each block.

As a result of the failure to calculate the Type I error and power for these 19 study conditions for the MH-Booklet analyses in the simulation study, I calculated the Type I error and power rates for these study conditions in Excel. Specifically, for every condition in which a DIF result had been returned as "NA", I averaged each item's Type I error or power rate over only those replications in which its MH DIF status was properly returned. For example, for a simulated non-DIF item, if the item's MH DIF status was not returned in 2 out of 100 replications in a study condition that item's Type I error rate was calculated as the average number of times the item was incorrectly identified as a DIF item in the 98 replications in which the DIF status was returned. As another example for a simulated DIF item, if the item's MH DIF status was not returned in 1 out of 100 replications in a study condition that item's power rate was calculated as the average number of times the item was correctly identified as a DIF item in the 99 replications in which the DIF status was returned.

4.4.2 Type I Error Results

The Type I error results are presented in seven sections. The first section presents an overview of the Type I error rates for each of the DIF methods across the five study factors. The second section provides an overview of the Type I error rates of all methods across all study conditions. In the subsequent four sections the Type I error results are presented in greater detail for each of the four DIF methods in turn. In the last section the Type I error rates are summarized when averaged across all levels of two of the study factors: DIF ratio and focal theta. These two study factors are unlikely to be known by test developers at the time of test development or by DIF analysts prior to conducting DIF analyses. Therefore averaging the results across these two factors is intended to guide DIF analysts and test developers by allowing them to make decisions based on the types of information they are likely to have at hand (i.e., the total sample sizes, ratio of group sample sizes, and block size).

4.4.2.1 Overall Type I Error Results

Table 11 presents the overall Type I error rates by study factor for each of the four DIF analysis methods to provide information about whether and how the Type I error rate varies across each study factor holding the other study factors constant. Inflated Type I error rates (those that exceed Bradley's (1978) moderate criterion of .060) are denoted with an asterisk in Table 11.

Table 11: ***Overall Type I Error Rates by DIF Method and Study Factor***

		DIF Method							
		MH-Block		MH-Booklet		MH-Modified		IRT-Lord's	
Study Factor	Number of Conditions	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Total Sample Size									
1200	24	.051	.019	.020	.004	.045	.015	.026	.025
4800	24	.030	.028	.047	.017	.019	.021	.178*	.151
8400	24	.011	.015	.057	.030	.006	.010	.238*	.206
Focal Group Mean Theta									
-1	36	.031	.023	.039	.020	.019	.018	.157*	.189
0	36	.031	.030	.044	.030	.027	.027	.137*	.154
Focal Group Sample Size Ratio									
0.2	36	.033	.025	.039	.023	.025	.021	.237*	.190
0.5	36	.029	.029	.044	.029	.022	.025	.057	.083
Proportion of DIF items									
0	24	.018	.017	.030	.009	.015	.016	.188*	.217
0.1	24	.020	.018	.033	.011	.016	.017	.146*	.170
0.3	24	.055	.025	.062*	.035	.040	.025	.108*	.107
Number of Items per Block									
10	36	.035	.028	.043	.027	.024	.024	.151*	.170
20	36	.027	.025	.040	.025	.022	.022	.144*	.176

*Type I error is inflated (exceeds Bradley's (1978) moderate criterion of .060.

MH-Block. The overall Type I error rate of the MH-Block method ranges from .011 to .055. The Type I error rate decreases as sample size increases, and increases as the proportion of DIF items increases. The MH-Block method maintains a similar Type I error rate across both levels of focal group ability, across both sample size ratios, and across both 10 and 20 items per block.

MH-Booklet. The overall MH-Booklet Type I error rate ranges from .020 to .062. The MH-Booklet method Type I error rate is inflated when 30% of the items are simulated to have DIF. The Type I error rate increases as sample size increases and as the proportion of DIF items increases. It maintains a similar Type I error rate across both levels of focal group ability, across both sample size ratios, and across both 10 and 20 items per block.

MH-Modified. The overall MH-Modified method Type I error rate ranges from .006 to .045. The Type I error rate tends to decrease as sample size increases, and to increase as the proportion of DIF items increases. It maintains a similar Type I error rate across both levels of focal group ability, across both sample size ratios, and across both 10 and 20 items per block.

IRT-Lord's. As can be seen in Table 11, the IRT-Lord's method is inflated across many of the study conditions, ranging from .026 to .238. The Type I error rate increases as sample size increases, and is inflated at sample sizes of 4800 and 8400. The IRT-Lord's method is severely inflated when the focal and reference groups have equal ability and when the focal group ability is 1 standard deviation below the reference group although to a slightly lesser extent when the focal and reference group abilities are equal than when they are unequal. IRT-Lord's method maintains moderate Type I error rates when group sizes are equal but has severely inflated Type I error rates when the group sample sizes are unequal.

The IRT-Lord's method Type I error rate decreases as the proportion of DIF items increases but is inflated across all three proportions of DIF items. The IRT-Lord's method is inflated to approximately the same extent in both the 10 items per block and the 20 items per block conditions.

4.4.2.2 Overview of Type I Error Rates of All Methods across All Study Conditions

Figure 5 presents graphs of the mean Type I error rates of all DIF methods investigated across all study conditions to elucidate the general trends in the Type I error results across study factors. In the figure the lower two rows of graphs represent the conditions in which the focal and reference group sample sizes are equal ("Focal.Ratio: 0.5") and the upper two rows represent conditions in which the focal and reference group sample sizes are unequal ("Focal.Ratio: 0.2"). Nested within these rows are the conditions in which the focal group mean ability level is equal to the reference group mean ability level ("Focal.Theta: 0") and one standard deviation below the reference group mean ability level ("Focal.Theta: -1"). The left three columns represent conditions with 10 items per block ("Block.Size: 10") and the right three columns represent conditions with 20 items per block ("Block.Size: 20"). Nested within block size conditions are the proportions of DIF items in the matching variable ("DIF.Ratio: 0", "DIF.Ratio: 0.1", and "DIF.Ratio: 0.3"). The y-axis of each graph represents mean Type I error for the study condition. (Please note that the y-axis had to extend to a value of 0.6 in order to accommodate the IRT-Lord's Type I error rates. This has the effect of making it difficult to discern differences between the MH methods all of which are in the lower range of the y-axis. However, the intent of this section is to provide an overview of the effects of the factors; more specific details about each method are provided in the next section.) The x-axis represents the total sample size. The

horizontal dotted line represents Type I error value of .060, the upper limit of Bradley's (1978) moderate criterion above which Type I error is considered to be inflated in this study.

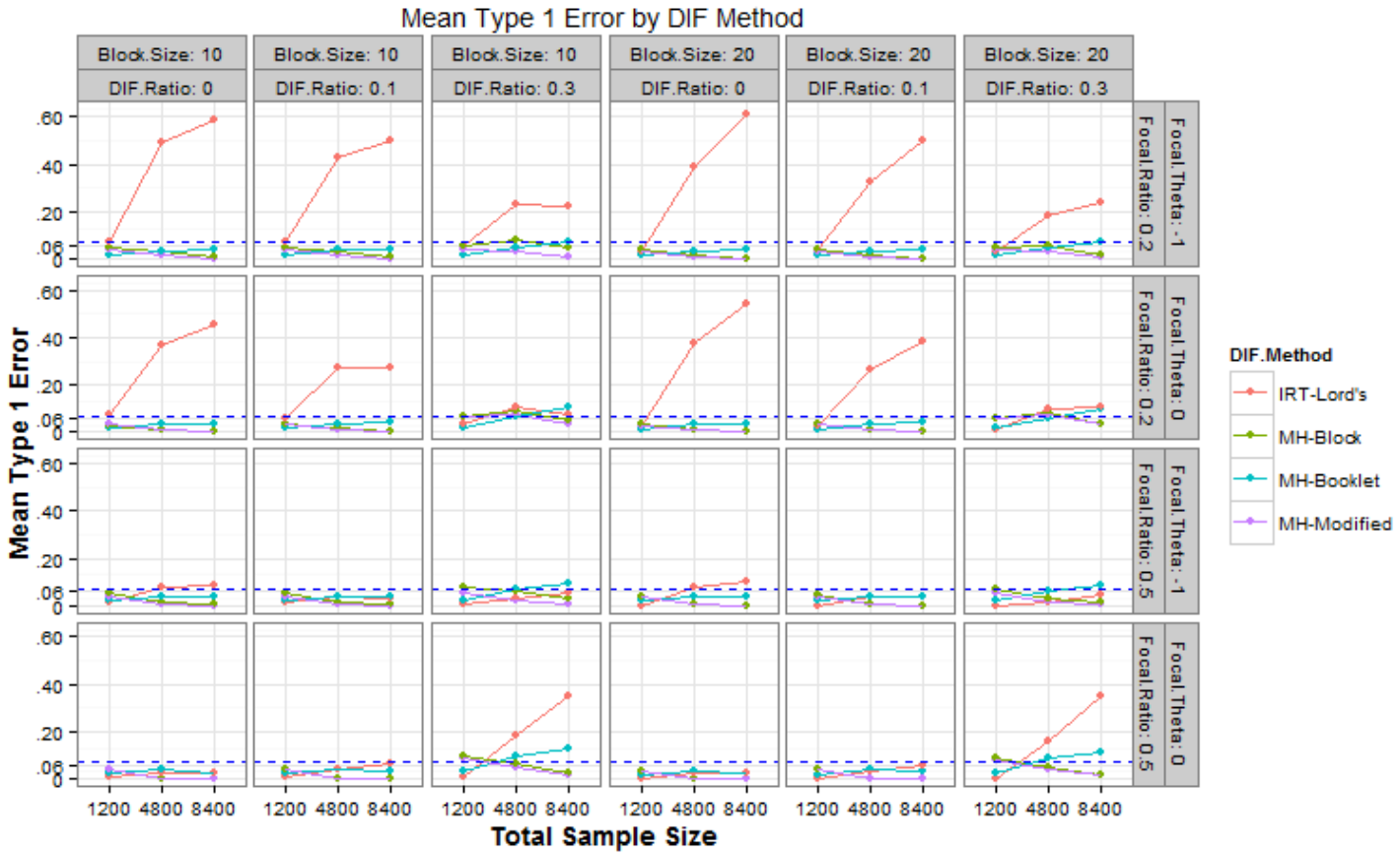

Figure 5: Type I Error Rates for all Methods Across all Study Conditions.

(NOTES – 1. the dotted line represents the upper limit of Bradley's moderate criterion above which Type I error rate is inflated. 2. DIF.Ratio refers to proportion of DIF items. 3. Focal.Theta refers to focal group mean ability. 4. Focal.Ratio refers to the focal group sample size ratio. 5. Block.Size refers to the number of items in each block.

Total Sample Size. In general, the IRT-Lord's method and the MH-Booklet method Type I error rates tend to increase as sample size increases. The MH-Block and MH-Modified method Type I error rates tend to decrease as sample size increases. The impact of total sample size on the IRT-Lord's method is particularly noticeable with highly inflated Type I error rates in the larger sample sizes when the group sample sizes are unequal.

Proportion of DIF items. Overall, for the three MH methods, as the proportion of DIF items in the matching variable increases from 0 to 0.1 there are very modest or no changes in Type I error rates across the remaining study conditions. However, as the proportion of DIF items in the matching variable increases from 0.1 to 0.3 each of the MH methods may become inflated but this appears to vary with total sample size. For example, the graph in the bottom row, third column from the left shows that the MH-Booklet method becomes inflated at the 8400 sample size whereas the MH-Block and MH-Modified methods are inflated at the smallest sample size. The effect of increasing proportions of DIF items in the matching variable on the IRT-Lord's method appears to vary with the ratio of group sample sizes. In the top two rows in which the groups are unequal in size and in the third row where the groups are equal in size but unequal in ability level, the IRT-Lord's method Type I error rate decreases with increasing proportions of DIF items. On the other hand, in the bottom row that represents equal group sizes and equal group ability the IRT-Lord's method Type I error rate increases with increasing proportions of DIF items.

Number of items per block. The general trend shows that the number of items in each block has little or no impact on the Type I error rates of any of the four methods studied indicated by the similarities between the left-most three columns and the right three columns.

Focal group mean theta. There are few or modest changes in Type I error rates of the three MH methods as a result of changing the focal group mean ability level. There appears to be a larger impact of focal group mean ability on the IRT-Lord's method when the group sample sizes are unequal than when they are equal.

Ratio of group sample sizes. The ratio of group sample sizes has a dramatic effect on the IRT-Lord's method as evidenced by the much higher Type I error rates in the top two rows of Figure 5 compared to the bottom two rows. This is particularly noticeable at the larger two total sample sizes. When the proportion of DIF items in the matching variable is high (0.3) there are small differences between each of the 3 MH methods which appear to vary according to the total sample size.

The following sections provide more detailed information about the Type I error rates for each of the DIF methods across all 72 study conditions.

4.4.2.3 MH-Block Method Type I Error Rates

Figure 6 presents graphs of the MH-Block method Type I error rates in each of the 72 study conditions. The layout of this figure is similar to Figure 5, with the exception that the plot lines represent equal (focal group theta = 0) and unequal (focal group theta = -1) group ability levels.

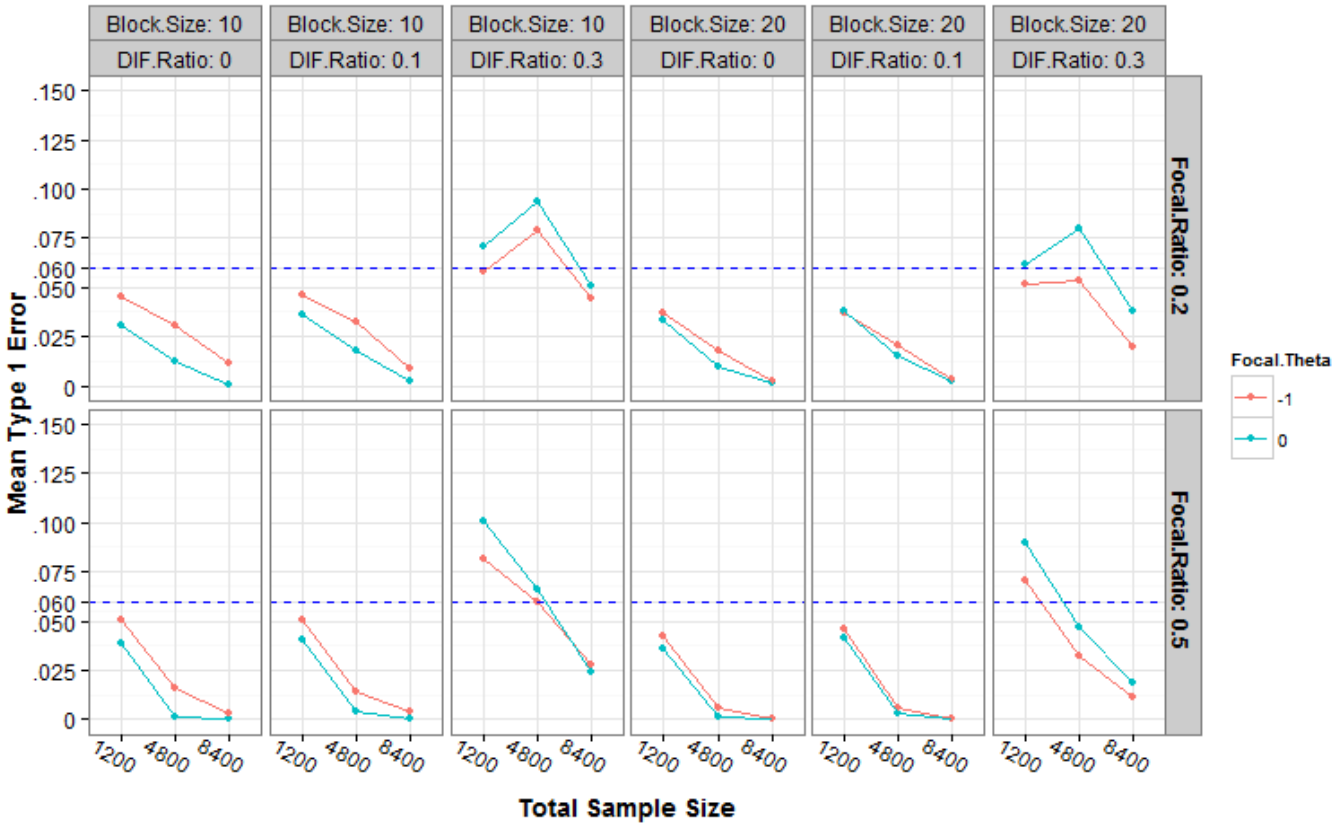

Figure 6: Type I Error Rates for MH-Block Method.

(NOTES – 1. the dotted line represents the upper limit of Bradley's moderate criterion above which Type I error rate is inflated. 2. DIF.Ratio refers to proportion of DIF items. 3. Focal.Theta refers to focal group mean ability. 4. Focal.Ratio refers to the focal group sample size ratio. 5. Block.Size refers to the number of items in each block.

These graphs provide additional information to the overall trends reported in the previous sections by revealing that although the Type I error rates of the MH-Block method indicated in Table 11 are very good when averaged across conditions, there are some study conditions in which the MH-Block Type I error rate is inflated. Trends across study factors are presented next, followed by the conditions in which the Type I error rate is inflated.

Total sample size. In general, the MH-Block method Type I error tends to decrease as total sample size increases from 1200 to 8400 with the largest decrease occurring between 1200 and 4800, although there is an anomaly to this pattern when the focal group sample size ratio is 0.2 and the proportion of DIF items is large. In these cases the Type I error rate is higher when the total sample size is 4800 than when it is either 1200 or 8400.

Proportion of DIF items. The MH-Block Type I error rate increases when the proportion of DIF items shifts from 0.10 to 0.30.

Number of items per block. The MH-Block Type I error rate appears to be slightly lower when there are 20 items in a block than when there are 10 items per block.

Focal group theta. When there are 20 items in a block and the proportion of DIF items is 0 or 0.1 the Type I error rate is the same for conditions in which the focal and reference group ability levels are equal and those in which the focal group ability level is lower than the reference group ability level. When there are 10 items in a block and the proportion of DIF items is 0 or 0.1 the Type I error rate is higher for conditions in which the focal group ability level is lower than the reference group ability level. However, when 30% of the items are DIF the Type I error rate for the MH-Block method is higher when the focal and reference groups have the same ability value than when they have equal ability levels.

Ratio of group sample sizes. When the proportion of DIF items is low (0 or 0.1) and the sample size is either 1200 or 8400, the Type I error rate is very similar when the groups are equal in size to when they are not. When the proportion of DIF items is low and the sample size is 4800 the Type I error rate is higher when the focal group sample size is smaller than when the groups are equal in size. When the proportion of DIF items is high and the sample size is either 4800 or 8400 the Type I error rate is higher when the focal group sample size is smaller than when the groups are equal in size. When the proportion of DIF items is high and the sample size is 1200 the Type I error rate is lower when the focal group sample size is smaller than when the groups are equal in size.

Inflation in Type I error rate. There are 10 of the 72 study conditions in which the MH-Block Type I error is inflated. All 10 occur when the proportion of DIF items is 0.3. The third graph of the top row of Figure 6 shows that MH-Block Type I error is inflated in the conditions in which the focal group sample size is smaller than the reference group sample size, there are 10 items per block, 30% of the items are DIF and the total sample size is either 1200 or 4800. This occurs both when the focal group theta value is equal to the reference group and when it is not. The sixth graph of the top row of Figure 6 shows that MH-Block Type I error is inflated in the condition in which the focal group sample size ratio is smaller than the reference group sample size, there are 20 items per block, 30% of the items are DIF and the total sample size is 1200 or 4800 when the focal group theta value is equal to the reference group theta value. Similarly, the third and sixth graphs of the bottom row show that MH-Block Type I error is inflated when the focal and reference group sizes are equal, 30% of the items are DIF and the total sample size is 1200 both when the focal and

reference groups have equal theta values and when they do not in the 10 item per block and in the 20 item per block conditions.

To summarize, the MH-Block method may demonstrate Type I error inflation in the 1200 or 4800 sample size conditions when there are a large proportion of DIF items.

4.4.2.4 MH-Booklet Method Type I Error Rates

Figure 7 presents graphs of the MH-Booklet method Type I error rates in each of the 72 study conditions. The layout of Figure 7 is the same as that of Figure 6 therefore the layout is not described again.

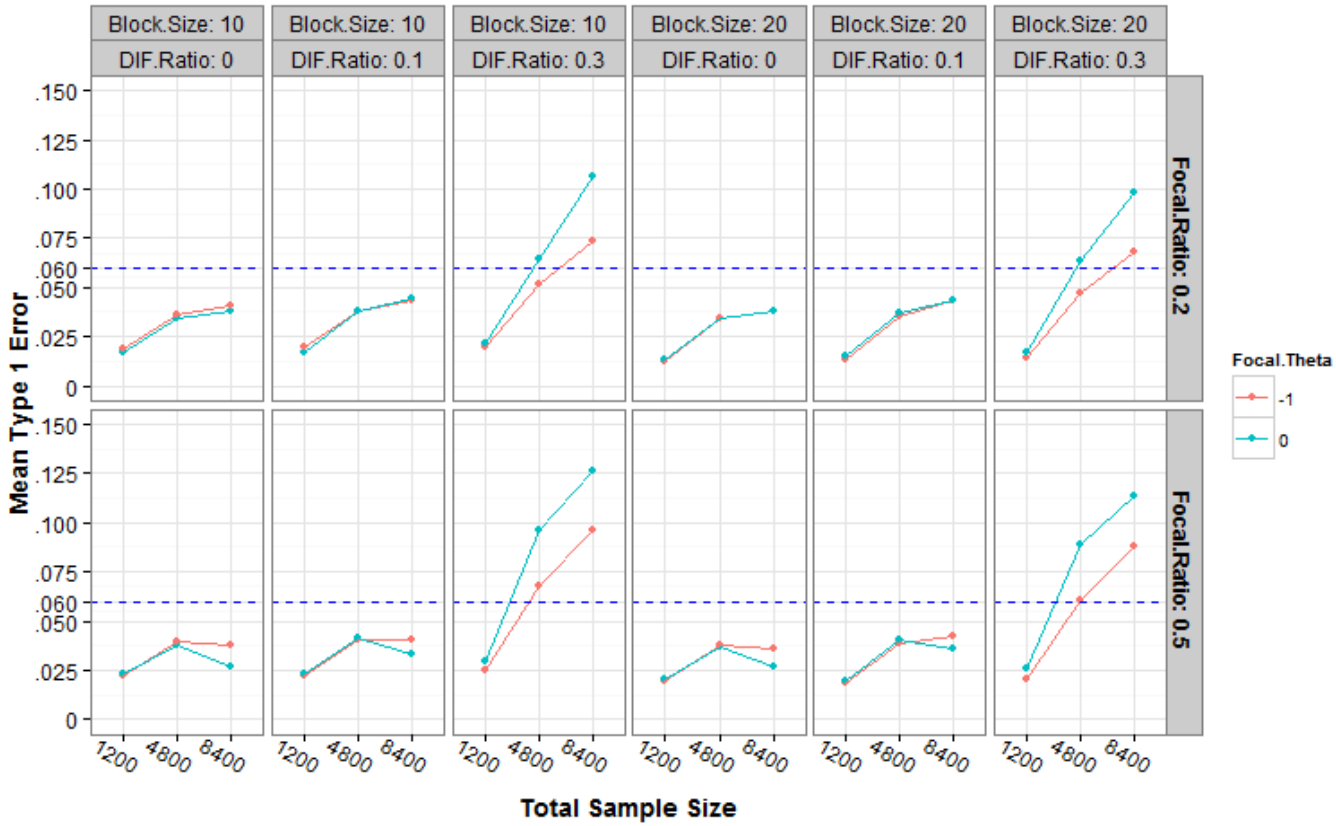

Figure 7: Type I Error Rates for MH-Booklet Method.

(NOTES – 1. the dotted line represents the upper limit of Bradley's moderate criterion above which Type I error rate is inflated. 2. DIF.Ratio refers to proportion of DIF items. 3. Focal.Theta refers to focal group mean ability. 4. Focal.Ratio refers to the focal group sample size ratio. 5. Block.Size refers to the number of items in each block.

Again, these graphs confirm the general trends reported in the previous sections but also reveal specific study conditions in which the MH-Booklet Type I error rate is inflated.

Total sample size. The MH-Booklet Type I error tends to increase as total sample size increases from 1200 to 4800. The Type I error change between sample size 4800 and 8400 varies according to the focal ratio, the focal theta and the proportion of DIF items. When the focal and reference group thetas are equal, the group sample size ratios are equal, and the proportion of DIF items is either 0 or 0.1, the Type I error rate decreases between 4800 and 8400. When the focal and reference group thetas are not equal and the proportion of DIF items is either 0 or 0.1 the Type I error rate stays approximately the same between 4800 and 8400 regardless of whether the group sample size ratios are equal or unequal. When the proportion of DIF items is 0.3 there is a much greater increase in Type I error as sample size increases than when the proportion of DIF items is lower.

Proportion of DIF items. The MH-Booklet Type I error rate increases when the proportion of DIF items in the matching variable shifts from 0.10 to 0.30 and the sample size is 4800 or 8400. This increase is greater when the focal group theta is equal to the reference group theta and when the focal group sample size is the same as the reference group sample size.

Number of items per block. The MH-Booklet method Type I error rate appears to be unaffected by the number of items in a block when the proportion of DIF items is low (0 or 0.1). When the proportion of DIF items is high and the sample sizes are 4800 or 8400 the Type I error rate is slightly lower for the 20 item per block conditions.

Focal group theta. When the proportion of DIF items is 0 or 0.1 the MH-Booklet Type I error rate is the same across both levels of focal group theta except when the sample

size is 8400 and the ratio of group sample sizes is equal. In these cases the MH-Booklet Type I error rate is slightly higher when the focal group theta is lower than the reference group theta than when the group thetas are equal. When the proportion of DIF items is high (.3) and the sample size is 1200, the MH-Booklet Type I error rate is the same across both levels of focal group theta however when the sample size is 4800 or 8400 the Type I error rate is higher when the reference and focal groups have equal ability levels than when the focal group has a lower ability level.

Ratio of group sample sizes. The Type I error rates of the MH-Booklet method are very similar when the sample size ratios are equal and when they are not when the proportion of DIF items is low and the total sample size is either 1200 or 4800 . However, when the proportion of DIF items is high and the sample size is either 4800 or 8400 the Type I error rate is higher when the focal and reference group sample size ratios are equal than when the focal group sample size is smaller than the reference group sample size.

Inflation in Type I error rate. There are 14 of the 72 study conditions in which the MH-Booklet Type I error is inflated. Similarly to the MH-Block method, all occur when the proportion of DIF items is 0.3. The third and sixth graphs of the top row of Figure 7 show that MH-Booklet Type I error is inflated in the conditions in which the focal group sample is smaller than the reference group, there are either 10 or 20 items per block, 30% of the items contain DIF, the total sample size is either 4800 or 8400, and the focal group theta is equal to the reference group theta. Inflation also occurs when the total sample size is 8400 and the group ability levels are different. The third and sixth graphs of the bottom row show that MH-Booklet Type I error is inflated when the focal and reference group sizes are equal, there

are either 10 or 20 items per block, 30% of the items are DIF and the total sample size is 4800 or 8400 regardless of group ability levels.

To summarize, the MH-Booklet method may demonstrate Type I error inflation in the larger sample sizes (4800 or 8400) when there are a large proportion of DIF items.

4.4.2.5 MH-Modified Method Type I Error Rates

Figure 8 presents graphs of the MH-Modified method Type I error rates in each of the 72 study conditions. The layout of Figure 8 is the same as that of Figure 6 so is not described again.

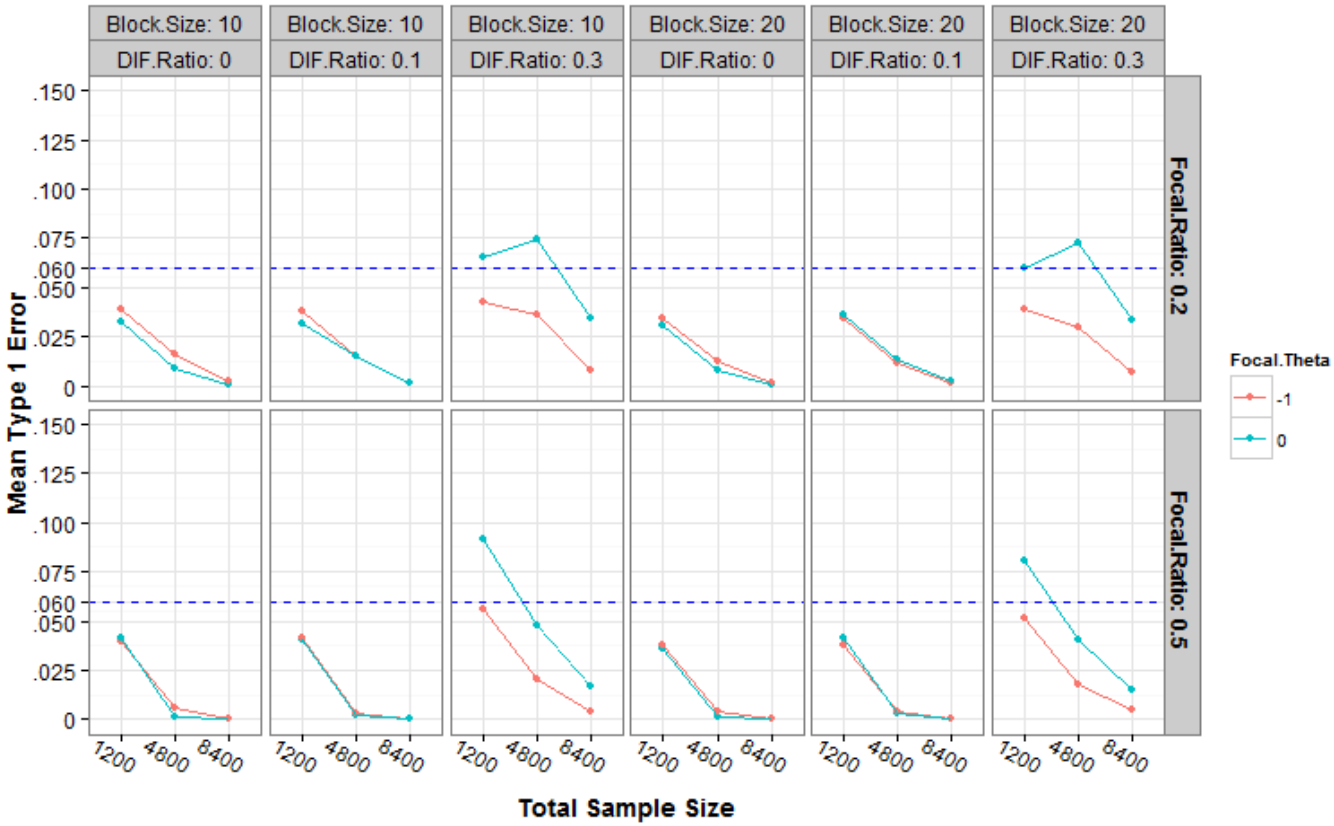

Figure 8: Type I Error Rates for MH-Modified Method.

(NOTES – 1. the dotted line represents the upper limit of Bradley's moderate criterion above which Type I error rate is inflated. 2. DIF.Ratio refers to proportion of DIF items. 3. Focal.Theta refers to focal group mean ability. 4. Focal.Ratio refers to the focal group sample size ratio. 5. Block.Size refers to the number of items in each block.

Again, these graphs confirm the general trends reported in the previous sections. They also reveal two study conditions in which the MH-Modified Type I error rate is inflated.

Total sample size. In general, the MH-Modified method Type I error tends to decrease as sample size increases. When the proportion of DIF items is low (0 or 0.1) the Type I error rate decreases rapidly as the sample size increases from 1200 to 4800. When the focal and reference group sample sizes are equal there is very little further change in Type I error rate as the sample size increases from 4800 to 8400 and when the focal group sample size is smaller than the reference group the Type I error rate continues to decrease as sample size increases from 4800 to 8400. When the proportion of DIF items is high (.3), the focal group sample size is smaller than the reference group, and the focal and reference group ability levels are equal the Type I error rate increases as the sample size increases from 1200 to 4800 and then decreases as the sample size increases to 8400. When the proportion of DIF items is high and the reference and focal group sample sizes are equal, the Type I error rate decreases continuously as the sample size increases from 1200 to 8400.

Proportion of DIF items. The MH-Modified method Type I error rate tends to increase when the proportion of DIF items shifts from 0.10 to 0.30.

Number of items per block. The MH-Modified method Type I error rate appears to be unaffected by the number of items in a block when the proportion of DIF items is low (0 or 0.1). When the proportion of DIF items is high, the reference and focal group samples are equal in size, the reference and focal group ability levels are equal and the total sample sizes are 1200 or 4800 the Type I error rate is slightly lower for the 20 item per block conditions.

Focal group theta. The MH-Modified Type I error rate appears to be similar across both levels of focal group theta when the proportion of DIF items is low (0 or 0.1). However, when the proportion of DIF items is high the Type I error rate is higher when the group ability levels are the same than when the focal group ability is lower than the reference group ability.

Ratio of group sample sizes. The Type I error rates of the MH-Modified method are very similar when the sample size ratios are equal and when they are not when the proportion of DIF items is low and the total sample size is either 1200 or 8400 . However, when the proportion of DIF items is high and the sample size is 1200 the Type I error rate is higher when the focal and reference group sample size ratios are equal than when the focal group sample size is smaller than the reference group sample size. When the proportion of DIF items is high and the sample size is 4800 or 8400 the reverse pattern is noted, with the Type I error rate being lower when the focal and reference group sample size ratios are equal than when the focal group sample size is smaller than the reference group sample size.

Inflation in Type I error rate. There are 5 of the 72 study conditions in which the MH-Modified Type I error rate is inflated. In the upper row of graphs it can be seen that inflation occurs when the proportion of DIF items is 0.3, the sample size is 1200 or 4800 (when there are 10 items per block) or 4800 (when there are 20 items per block), and the focal and reference groups have equal ability levels. In the lower row of graphs it can be seen that the Type I error is inflated in similar conditions as the top row, except only when the total sample size is 1200.

4.4.2.6 IRT-Lord's Method Type I Error Rates

The Type I error rates for the IRT-Lord's method across all study conditions are shown in Figure 9. It is important to note that the y-axis of Figure 9 covers a wider range of mean Type I error values than Figures 6-8 in order to accommodate higher Type I error values for the IRT-Lord's method than were observed for the other three methods. Again, the dotted line represents the upper limit of Bradley's (1978) moderate criterion above which Type I error rate is inflated.

Total sample size. In general, the IRT-Lord's method Type I error rate increases as sample size increases. However, anomalies are noted when the focal group sample size is smaller than the reference group sample size, there are 10 items per block, the proportion of DIF items is 0.1 or 0.3, and the sample size is 4800 or 8400 (the second and third graphs of the upper row of Figure 9).

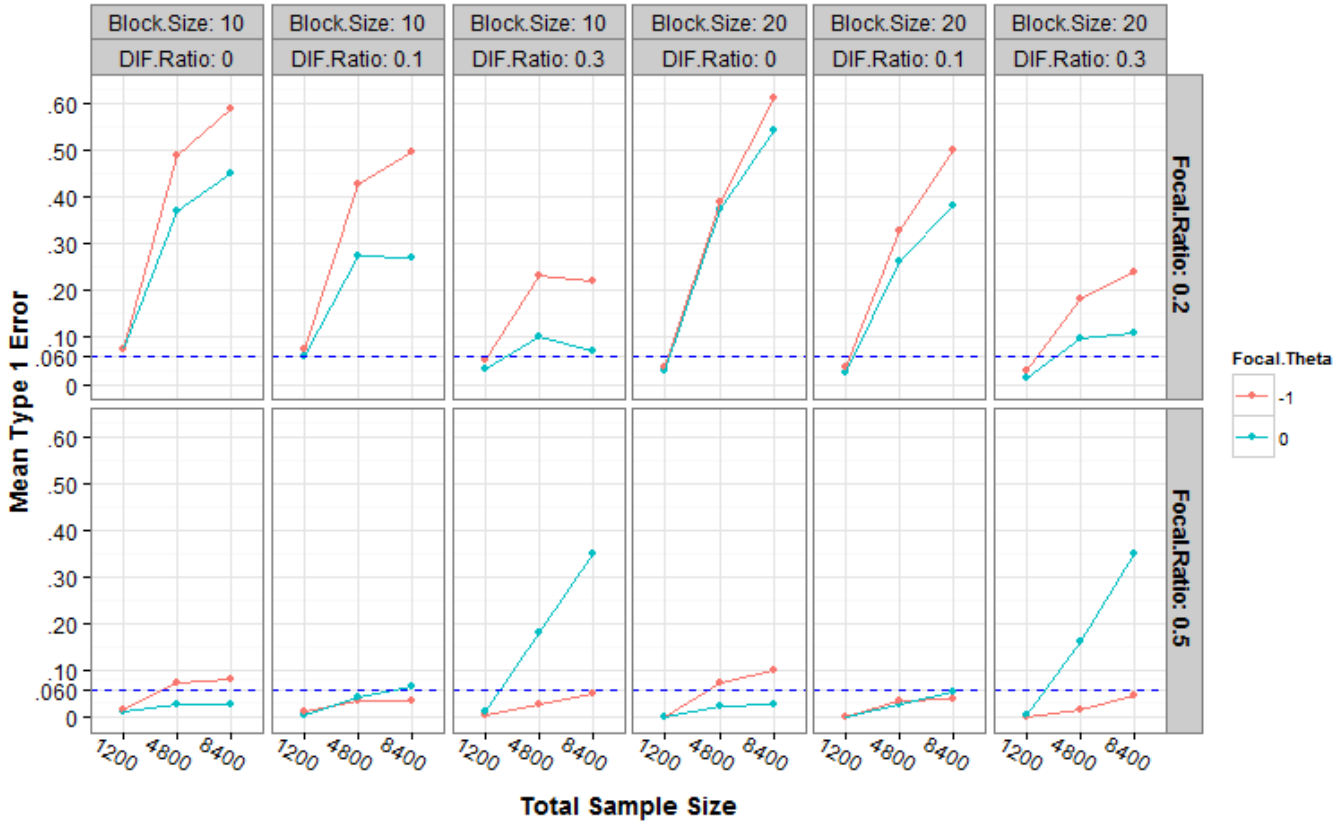

Figure 9: Type I Error Rates for IRT-Lord's Method.

(NOTES – 1. the dotted line represents the upper limit of Bradley's moderate criterion above which Type I error rate is inflated. 2. DIF.Ratio refers to proportion of DIF items. 3. Focal.Theta refers to focal group mean ability. 4. Focal.Ratio refers to the focal group sample size ratio. 5. Block.Size refers to the number of items in each block.

Proportion of DIF items. Referring first to the bottom row of graphs in Figure 9 in which the focal and reference group sample sizes are equal, it can be seen that The IRT-Lord's Type I error tends to increase as the proportion of DIF items increases (as would be expected) when the group ability levels are equal. This increase is particularly noticeable when 30% of the items are DIF and the sample size is 4800 or 8400. Yet the IRT-Lord's Type I error tends to decrease modestly as the proportion of DIF items increases when the focal group ability level is 1 standard deviation below the reference group ability level. The top row of Figure 9 in which the focal group sample size is smaller than the reference group size shows that the IRT-Lord's method Type I error rate decreases as the proportion of DIF items increases when the sample size is either 4800 or 8400 regardless of whether there are differences in the group ability levels or not. The decreases in Type I error as the proportion of DIF items increases are unexpected results.

Number of items per block. When the focal and reference group sample sizes are equal the number of items per block does not appear to have an impact on the Type I error rate of the IRT-Lord's method. When the focal group sample size is smaller than the reference group, the focal group ability is lower than the reference group, the proportion of DIF items is 0 or 0.1, and the sample size is 1200 or 4800, the Type I error rate appears to be lower when there are 20 items in a block than when there are 10 items in a block. When the focal group sample size is smaller than the reference group, the focal group ability is equal to the reference group, the proportion of DIF items is 0 or 0.1, and the sample size is 8400, the Type I error rate appears to be higher when there are 20 items in a block than when there are 10 items in a block.

Focal group theta. In all conditions in which the total sample size is 1200 the IRT-Lord's method Type I error rate is equal across the two levels of focal group theta. Referring to the bottom row of graphs in Figure 9 in which the group sample sizes are equal, differences arise in Type I error rates across levels of focal group ability particularly when the proportion of DIF items is 0.3 and the sample sizes are 4800 or 8400 (the third and sixth graphs of the bottom row). In these cases the Type I error rate is higher when the reference and focal group ability levels are equal than when the focal group ability level is lower than the reference group. Referring to the top row of Figure 9 in which the focal group sample size is smaller than the reference group, differences arise across levels of focal group theta at sample sizes of 4800 and 8400 and in these cases the Type I error rate is higher when the focal and reference group ability levels are different than when they are equal.

Ratio of group sample sizes. Large differences across the Lord's IRT method Type I error rates are noted between conditions in which the group sample sizes are equal (the bottom row of Figure 9) and those in which the focal group is smaller than the reference group (the top row of Figure 9). When the proportion of DIF items is either 0 or 0.1, the Type I error rate is lower when the two groups have equal sample sizes. When the proportion of DIF items is 0.3 and the focal and reference groups have equal ability, the Type I error rate is higher when the groups have equal sample sizes. When the proportion of DIF items is 0.3 and the focal group has lower ability than the reference group the Type I error rate is lower when the groups have equal sample sizes.

Inflation in Type I error rate. Figure 9 reveals that the IRT-Lord's Type I error is inflated in 37 of the 72 study conditions. Most of these occur when the focal group sample is smaller than the reference group (the top row of graphs in Figure 9), the total sample size is

4800 or 8400 regardless of whether there are differences in group ability or not, and across all proportions of DIF items. When the group sample sizes are equal (the bottom row of graphs in Figure 9) and the groups have equal ability levels, the IRT-Lord's method Type I error is inflated when the proportion of DIF items is 0.3 and the sample size is either 4800 or 8400 regardless of the number of items in a block. When the group sample sizes are equal and the groups have different ability levels, the Type I error is inflated when there is no DIF and the sample size is either 4800 or 8400.

In summary, the IRT-Lord's method tends to experience severely inflated Type I error rates when the focal group is smaller than the reference group and the total sample size is 4800 or 8400. When the focal and reference groups are equal in size and the groups have equal ability levels the IRT-Lord's method has inflated Type I error rates when the proportion of DIF items is high and the sample size is 4800 or 8400. When the focal and reference groups are equal in size but the groups have different ability levels the IRT-Lord's method has inflated Type I error rates when there is no DIF present and the sample size is 4800 or 8400.

4.4.2.7 Summary of Type I Error Results

Overall, the three MH DIF approaches maintain better control of Type I error rates than the IRT-Lord's method across all of the study conditions. When the proportion of DIF items is high the MH approaches may demonstrate inflated Type I error rate. However, the sample size conditions in which this occurs varies by MH approach with MH-Block and MH-Modified tending to be inflated at the lower sample sizes while the MH-Booklet method is inflated in the larger sample sizes. The IRT-Lord's method demonstrates severe inflation in the larger total sample sizes when the group sample sizes are unbalanced.

One of the goals of this dissertation was to gain an understanding of the impact of certain factors previously shown to affect DIF detection in complete data on DIF analyses in structurally missing data to provide guidance for test developers and administrators as well as DIF analysts. Of the five factors investigated in this study three are potentially under the control of test developers/administrators or known to DIF analysts prior to conducting DIF analyses (the total sample size, block size, and ratio of group sample sizes) while two would be unlikely to be controlled or known in advance (the proportion of DIF items and the ability levels of the focal and reference groups). It is therefore instructive to summarize how the Type I error rates of the four methods vary by the factors that may be known, averaged over the factors that are unlikely to be known. Figure 10 presents the Type I error rates (y-axis) of the four DIF methods by total sample size (x-axis), block size (columns) and ratio of focal group sample size to reference group sample size (rows), averaged across all levels of the proportion of DIF items and group abilities.

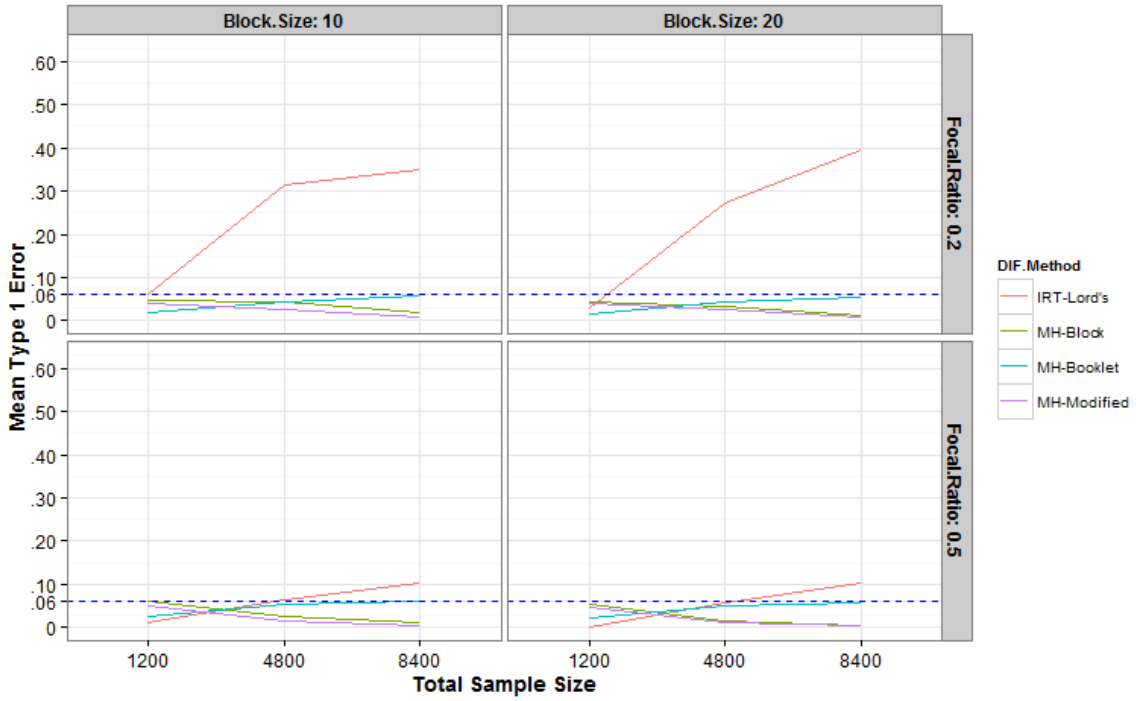

Figure 10: Type I Error Rates across Conditions that are Likely to be Known.

(NOTES – 1. the dotted line represents the upper limit of Bradley's moderate criterion above which Type I error rate is inflated. 2. Focal.Ratio refers to the focal group sample size ratio. 3. Block.Size refers to the number of items in each block.)

The effects of total sample size and focal group sample size ratio on the Type I error rates vary by method. An interaction between sample size and DIF method is evident and is more pronounced when the focal and reference groups are equal in size. As the total sample size increases the Type I error rates of the MH-Booklet and IRT-Lord's methods increase whereas the Type I error rates of the MH-Block and MH-Modified methods decrease. When the total sample size is 1200, the MH-Booklet and IRT-Lord's methods have lower Type I error rates than the other two methods while at total sample sizes of 4800 and 8400 the MH-Block and MH-Modified methods have lower Type I error rates. The IRT-Lord's method Type 1 error rate is greater when the focal group sample size is smaller than the reference

group, particularly when the total sample size is 4800 or 8400. The three MH approaches maintain Type I error across all levels of total sample size, block size and group ratios when averaged across all levels of DIF and ability. When the group sample sizes are equal the IRT-Lord's method has inflated Type I error rate when the total sample size is 8400, and when the group sample sizes are unequal the IRT-Lord's method has inflated Type I error rates when the total sample size is either 4800 or 8400.

4.4.3 Power Results

Power results are presented only for those study conditions in which the Type I error rate met Bradley's (1978) moderate criterion because inflated Type I error rates render associated power non-interpretable.

Power values of 0.80 and higher are typically considered sufficient to detect real effects at the .05 significance level (Cohen, 1988). However, in DIF studies lower power values are often found and are reported (as, for example, Hidalgo & Lopez-Pina, 2004; Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006; and others). It appears that in DIF studies the meaning attached to DIF detection power values is somewhat arbitrary. In this study a DIF detection rate ≥ 0.50 was used as a reference point to signal power equal to or better than a chance DIF detection rate. DIF detection rates ≥ 0.80 were considered to represent high power, and DIF detection rates < 0.50 were considered to be low.

4.4.3.1 Overall Power Results

Table 12 presents the overall power rates by study factor for each of the four DIF analysis methods to provide information about whether or how power varies across each study factor holding the other study factors constant. Note that the number of conditions that contribute to the mean power rate for each study factor (Valid N) varies because conditions

which did not satisfy Bradley's (1978) moderate criterion were not included in the calculation of the mean power rate. There were 48 study conditions in which DIF was simulated; these are the conditions for which power is relevant and was calculated.

Table 12 shows that, over most study factors, in general the MH-Booklet, MH-Block and MH-Modified methods had low power. The IRT-Lord's method varied from low to high power, although its power was calculated on fewer study conditions due to its inflated Type I error in many conditions. The variability in the power of the IRT-Lord's method was greater than the three MH methods, evidenced by higher standard deviations.

A brief description of the general trends in power at the overall study level across the study factors is provided next, followed by more detailed descriptions of power results for each of the four DIF methods across all of the study factors. This section then ends with a summary of power results.

Table 12: ***Overall Power Rates by Study Factor***

	DIF Method											
	MH-Block			MH-Booklet			MH-Modified			IRT-Lord's		
Study Factor	Valid N*	Mean	Standard Deviation	Valid N*	Mean	Standard Deviation	Valid N*	Mean	Standard Deviation	Valid N*	Mean	Standard Deviation
Total Sample Size												
1200	10	.333	.147	16	.049	.020	13	.304	.118	14	.206	.101
4800	12	.540	.180	10	.301	.137	14	.530	.142	6	.889	.045
8400	16	.527	.188	8	.519	.140	16	.517	.155	5	.963	.018
Focal Group Theta												
-1	21	.353	.095	18	.180	.158	24	.371	.115	15	.564	.394
0	17	.636	.166	16	.294	.261	19	.565	.170	10	.458	.347
Focal Group Sample Size Ratio												
0.2	19	.440	.204	18	.192	.185	21	.419	.181	6	.171	.099
0.5	19	.519	.177	16	.280	.246	22	.493	.156	19	.632	.358
Proportion of DIF items												
0	-	-	-	-	-	-	-	-	-	-	-	-
0.1	24	.538	.204	24	.305	.221	24	.533	.176	13	.605	.384
0.3	14	.380	.123	10	.062	.045	19	.360	.105	12	.430	.353
Number of Items per Block												
10	18	.494	.213	17	.239	.219	21	.470	.175	11	.576	.328
20	20	.467	.177	17	.228	.222	22	.444	.170	14	.479	.411

*NOTE: Valid N refers to number of study conditions. Valid N varies because power was not interpreted for study conditions in which the Type I error rate was inflated.

Total sample size. Power increased as the sample size increased for the MH-Booklet and IRT-Lord's methods. Power increased as the sample size increased from 1200 to 4800 for the MH-Block and MH-Modified methods then decreased slightly as the total sample size increased to 8400. Holding other study factors steady, MH-Booklet had very low power to detect the presence of DIF in the smallest sample size (1200) while at the 4800 and 8400 sample sizes IRT-Lord's method had high power, although this power was based on fewer study conditions due to the presence of inflated Type I error in many conditions at these sample sizes.

Focal group theta. The IRT-Lord's method power was slightly lower (by .10) when the focal and reference group mean ability was equal than when there was a difference in mean ability whereas power levels for MH-Block, MH-Booklet and MH-Modified were higher when the focal and reference group mean ability was equal.

Ratio of group sample sizes. Power was higher when the focal and reference group sample sizes were equal than when they were unequal for all four methods although this difference was much larger for IRT-Lord's method than for the three MH methods.

Proportion of DIF items. For all four DIF methods power decreased as the proportion of DIF items increased.

Number of items per block. For the MH-Block, MH-Booklet and MH-Modified methods there was no appreciable effect of block length on power. Power for the IRT-Lord's method was slightly higher (~.10) when there were 10 items per block than when there were 20 items per block.

The power results for all four DIF methods in all 48 study conditions in which power is relevant (i.e. those study condition in which DIF was simulated) are presented in

Table 13. Power is indicated as “NA” in the conditions in which it was not interpretable due to inflated Type I error rates.

Table 13: ***Power Rates for all Study Conditions***

				Equal Group Sizes				Unequal Group Sizes			
Focal Group Theta	Proportion of DIF Items	Block Size	Sample Size	MH-Block	MH-Booklet	MH-Modified	IRT-Lord's	MH-Block	MH-Booklet	MH-Modified	IRT-Lord's
0	0.1	10	1200	0.59	0.10	0.53	0.40	0.38	0.06	0.36	NA
0	0.1	10	4800	0.84	0.53	0.78	0.90	0.76	0.34	0.71	NA
0	0.1	10	8400	0.88	0.70	0.81	NA	0.84	0.57	0.76	NA
0	0.1	20	1200	0.54	0.07	0.50	0.26	0.39	0.04	0.36	0.14
0	0.1	20	4800	0.77	0.50	0.73	0.94	0.72	0.34	0.68	NA
0	0.1	20	8400	0.78	0.68	0.74	0.99	0.76	0.58	0.72	NA
0	0.3	10	1200	NA	0.07	NA	0.30	NA	0.04	NA	0.36
0	0.3	10	4800	NA	NA	0.50	NA	NA	NA	NA	NA
0	0.3	10	8400	0.55	NA	0.47	NA	0.55	NA	0.48	NA
0	0.3	20	1200	NA	0.05	NA	0.15	NA	0.03	0.22	0.12
0	0.3	20	4800	0.50	NA	0.48	NA	NA	NA	NA	NA
0	0.3	20	8400	0.48	NA	0.46	NA	0.48	NA	0.45	NA
-1	0.1	10	1200	0.33	0.06	0.34	0.26	0.22	0.04	0.23	NA
-1	0.1	10	4800	0.45	0.32	0.53	0.89	0.43	0.20	0.46	NA
-1	0.1	10	8400	0.43	0.47	0.51	0.95	0.42	0.32	0.45	NA
-1	0.1	20	1200	0.34	0.05	0.34	0.13	0.21	0.02	0.21	0.10
-1	0.1	20	4800	0.49	0.31	0.55	0.93	0.44	0.19	0.49	NA
-1	0.1	20	8400	0.46	0.49	0.52	0.98	0.43	0.34	0.48	NA
-1	0.3	10	1200	NA	0.05	0.29	0.25	0.17	0.03	0.17	0.20
-1	0.3	10	4800	0.37	NA	0.39	0.86	NA	0.15	0.38	NA
-1	0.3	10	8400	0.36	NA	0.38	0.96	0.33	NA	0.34	NA
-1	0.3	20	1200	NA	0.04	0.24	0.10	0.16	0.02	0.15	0.11
-1	0.3	20	4800	0.35	NA	0.39	0.82	0.34	0.13	0.36	NA
-1	0.3	20	8400	0.34	NA	0.37	0.94	0.32	NA	0.33	NA

NA: Power not interpreted due to inflated Type I error in this study condition

4.4.3.2 MH-Block Method Power

There were 10 conditions in which the power of the MH-Block method was not interpretable due to inflated Type I error. Power tended to increase as total sample size increased with the largest increase occurring between the 1200 and 4800 sample sizes and little or no change occurring between the 4800 and 8400 sample sizes. Power decreased as the proportion of DIF items increased. The power of the MH-Block method did not appear to be influenced by block size with similar performance when there were 10 items as when there were 20 items in a block. Power was slightly higher when group sizes were equal than when they were unequal and higher when group ability levels were equal than when there were true ability differences.

The power of the MH-Block method was low (<.50) or not interpreted due to inflated Type I error in 35 of the 48 study conditions in which DIF was simulated. MH-Block method power met or exceeded the chance DIF detection level (i.e., .50) in 13 of the 48 conditions. The power was high (greater than or equal to 0.80) in 3 of those 13 conditions. All 13conditions occurred when the focal and reference group ability levels were equal. When the focal group ability level was 1 standard deviation below the reference group ability MH-Block power was low or not interpretable.

Within the conditions in which the focal and reference group ability levels were equal, when the reference and focal group sample sizes were also equal the MH-Block power was above the chance detection level or high across all sample sizes and block sizes when 10% of the items in the matching variable were simulated to have DIF. MH-Block power was low or not interpretable in four of the six conditions in which 30% of the items were simulated to have DIF.

When the focal and reference group ability levels were equal and the group sample sizes were unequal, the power of the MH-Block method exceeded the chance DIF detection level in the 4800 and 8400 sample sizes when there were 10% DIF items (regardless of block size). When 30% of the items in the matching variable were DIF there was only one condition that exceeded a chance DIF detection level.

4.4.3.3 MH-Booklet Method Power

There were 14 conditions in which the power of the MH-Booklet method was not interpretable due to Type I error inflation. The overall pattern of the MH-Booklet power results across the study conditions was the same as that of the MH-Block method. Power tended to increase as total sample size increased with the largest increase between the 1200 and 4800 sample sizes, and decrease as the proportion of DIF items increased. The power of the MH-Booklet method did not appear to be influenced by block size with similar performance when there were 10 items as when there were 20. Power was slightly higher when group sizes were equal than when they were unequal and higher when group ability levels were equal than when there were true ability differences.

The power of the MH-Booklet method was low or not interpreted in 42 of the 48 DIF study conditions. The 6 conditions in which the power met or exceeded the chance level occurred when the focal and reference group ability levels were equal and only 10% of the items were simulated to have DIF. There were no conditions in which the MH-Booklet method had high power.

4.4.3.4 MH-Modified Method Power

There were 5 conditions in which the power of the MH-Modified method was not interpretable due to Type I error inflation. The overall pattern of the MH-Modified power

results across the study conditions was very similar to that of the MH-Block and MH-Booklet methods. Power tended to increase as total sample size increased with the largest increase between 1200 and 4800, and decrease as the proportion of DIF items increased. The power of the MH-Modified method also did not appear to be influenced by block size. Power was slightly higher when group sizes were equal than when they were unequal particularly at the smallest sample size and higher when group ability levels were equal than when there were true ability differences.

The power of the MH-Modified method was low or not interpreted in 33 of the 48 study conditions in which DIF was simulated. Similarly to the other two MH methods the MH-Modified method had higher power when the focal and reference group ability levels were equal than when they were unequal. Two of the 15 conditions where the MH-Modified method met or exceeded a chance DIF detection level occurred in the smallest sample size (1200) while 13 occurred when the total sample size was either 4800 or 8400. Fourteen of the 15 conditions occurred when the proportion of DIF items in the matching variable was 10% while the remaining 1 condition occurred when the proportion of DIF items in the matching variable was 30%. The MH-Modified power was high in one study condition.

4.4.3.5 IRT-Lord's Method Power

There were 23 of the 48 study conditions in which the power of the IRT-Lord's method was not interpretable due to Type I error inflation. There were 18 of the 24 unequal group size conditions that were not interpretable and power for the remaining 6 was very low. Therefore there were too few results in the unequal group size conditions to draw meaningful comparisons and the following results pertain only to the equal group

size study conditions. Power tended to increase as total sample size increased with a very large increase occurring between sample sizes of 1200 and 4800. Unlike the MH methods, when the focal and reference group ability levels differed, the mean power of the IRT-Lord's method did not appear to change as the proportion of DIF items increased. When the focal and reference group ability levels were equal the power decreased as the proportion of DIF items increased at the 1200 sample size but it is not possible to draw conclusions about the power rates at the 4800 and 8400 sample sizes because the power was not interpretable at these sample sizes when 30% of the items were DIF. The power of the IRT-Lord's method did not appear to be influenced by block size at the 4800 and 8400 sample sizes, but decreased as block size increased at the 1200 sample size.

The IRT-Lord's method had high power (equal to or greater than 0.80) in 11 of the 24 equal group size study conditions, and low power (less than 0.50) or power that was not interpretable due to inflated Type I error rates in the remaining 13 conditions. In the equal group size and unequal group ability study conditions the IRT-Lord's method had high power to detect DIF in all conditions when the sample size was either 4800 or 8400 but not at the 1200 sample size. In the equal group size and equal group ability study conditions it had high power in the conditions when there were 10% DIF items and the sample size was either 4800 or 8400 (except for one not interpretable condition due to inflated Type I error rate) but not when the sample size was 1200 or when there were 30% DIF items.

4.4.3.6 Summary of Power Results

Figure 11 presents graphs of the power results of the four DIF methods across all study conditions in which DIF was simulated to help elucidate and summarize the general

trends in power results across study factors and methods. The layout of this figure is similar to layout of the Type I error rate graph shown in Figure 5 except there is no column for the conditions in which DIF was not simulated. Please note that the y-axis represents the mean power *given Type I error rate*. That is, power values are not included in this graph in the conditions in which Type I error was inflated (i.e. exceeded Bradley's (1978) moderate criterion) because power was not interpreted in those conditions. The blue dotted line is provided as a point of reference representing the chance DIF detection level of 0.50.

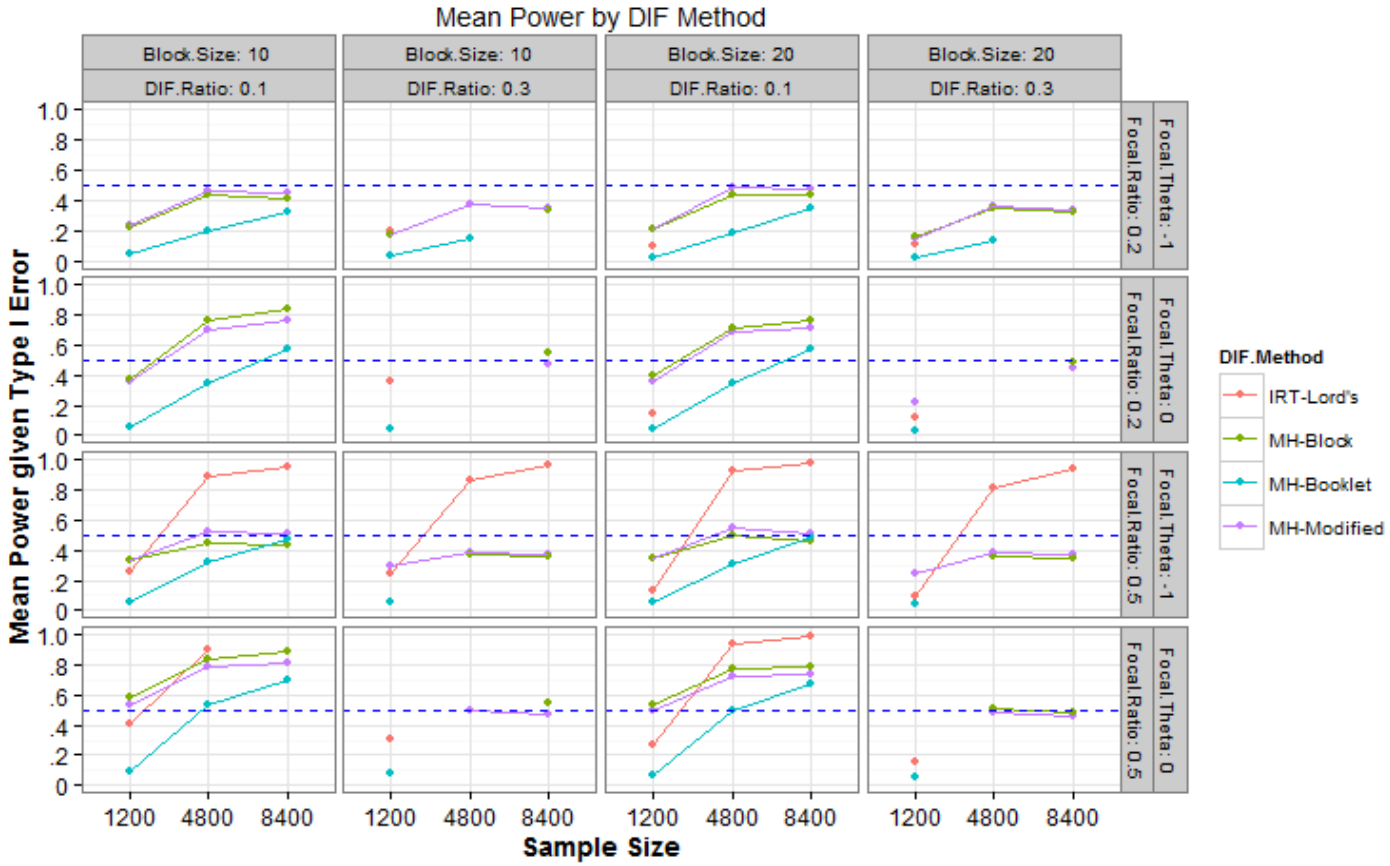

Figure 11: Mean Power Given Type I Error Rates for all Methods across All Study Conditions with DIF.

(NOTES – 1. the dotted line represents a chance level of DIF detection (0.50). 2. DIF.Ratio refers to proportion of DIF items. 3. Focal.Theta refers to focal group mean ability. 4. Focal.Ratio refers to the focal group sample size ratio. 5. Block.Size refers to the number of items in each block.

Referring to Figure 11, five important points can be seen that summarize the power results of this study. First, none of the methods investigated has high power to detect DIF in any conditions when the total sample size is 1200. At this sample size the MH-Block and MH-Modified methods generally have higher power than the remaining two methods, but nonetheless, their power only exceeds a chance DIF detection level when the group sample sizes and mean ability levels are equal and the proportion of DIF items in the matching variable is low (seen in the first and third graphs of the bottom row in Figure 11). Second, the MH-Booklet method has the lowest power of the four methods studied. This is true across all but 2 of the 48 study conditions in which power was analyzed. Third, all three MH methods have low power when the focal group ability level is lower than the reference group ability level (seen in the top row and the third row in Figure 11). Fourth, all three MH methods have low power when the proportion of DIF items in the matching variable is high (seen in the second and fourth columns in Figure 11). Last, the IRT-Lord's method has high power in the larger sample sizes if the ratio of group sizes is equal (seen in the third and fourth rows in Figure 11) except when the proportion of DIF items is high and the group ability levels are equal (the second and fourth graphs of the bottom row). When it has high power, the IRT-Lord's method has the highest power of the 4 methods investigated. It is important to recall that of the four methods investigated the IRT-Lord's method had the greatest variability and its power was calculated on considerably fewer study conditions due to its inflated Type I error.

4.4.4 Modification to the Pooled Booklet Approach

Despite differences in the designs of the Goodman et al. (2011) study and this dissertation (this dissertation investigated more factors in order to be as representative of

real LSA data as possible), it is feasible to average over some of the factors in this dissertation to explore general trends and similarities in Type I error and power results of the pooled booklet method used by Goodman et al. and the MH-Modified method used in this dissertation.

Table 14 shows the conditions that are similar between the Goodman et al. (2011) study and this dissertation, the Type I error rates and power results for the Goodman et al. pooled MH method and the MH-Modified method, and the difference between them. Since Goodman et al. used a focal group mean theta of 0.62 below the reference group mean theta, for the purposes of this comparison I have averaged over the two focal group mean theta values used in this dissertation. In addition, since Goodman et al. investigated Type I error only in the conditions in which there was no DIF simulated, this table presents the Type I error results only for the no DIF conditions in this dissertation. To create their DIF conditions, Goodman et al. simulated moderate DIF in the first item of each block and reported power averaged over all study conditions which averages to approximately 15% DIF. Therefore for comparison purposes I have presented the results of this dissertation for the conditions with DIF in 10% of the items being the closest value to 15%. Goodman et al. sample sizes of 1200, 6000 and 12000 are compared to this study sample sizes of 1200, 4800 and 8400 respectively.

Table 14: ***Type I Error and Power Comparisons of MH-Pooled* and MH-Modified Methods***

	Sample Size		Equal Group Size Conditions			Unequal Group Size Conditions		
Block Size	MH-Pooled*	MH-Modified	MH-Pooled*	MH-Modified	Difference**	MH-Pooled*	MH-Modified	Difference**
				Type I error				
10	1200	1200	0.033	0.040	-0.007	0.034	0.036	-0.002
20	1200	1200	0.033	0.037	-0.004	0.031	0.033	-0.002
10	6000	4800	0.004	0.004	0.001	0.011	0.013	-0.002
20	6000	4800	0.002	0.003	-0.001	0.008	0.010	-0.002
10	12000	8400	0.000	0.001	-0.001	0.002	0.002	0.001
20	12000	8400	0.001	0.000	0.001	0.001	0.002	-0.001
				Power				
10	1200	1200	0.392	0.435	-0.043	0.252	0.295	-0.043
20	1200	1200	0.323	0.420	-0.097	0.178	0.285	-0.107
10	6000	4800	0.668	0.655	0.013	0.568	0.585	-0.017
20	6000	4800	0.712	0.640	0.072	0.612	0.585	0.027
10	12000	8400	0.680	0.660	0.020	0.608	0.605	0.003
20	12000	8400	0.688	0.630	0.058	0.592	0.600	-0.008

*MH-Pooled refers to the pooled MH method used by Goodman et al. (2011)
** Difference = MH-Pooled - MH-Modified

Table 14 reveals very similar Type I error trends and values for the MH-Modified method investigated in this dissertation and the pooled MH method used in the Goodman et al. (2011) study. Similarly to the pooled MH method, the MH-Modified method Type I error decreases most dramatically between the smallest and middle sample sizes; very little change in Type I error occurs between the middle and largest sample sizes. The largest differences between the pooled method and the MH-Modified Type I error are .004 and .007 favouring the pooled method. These occur in the smallest sample size and equal group size conditions. The remaining differences in Type I error are very small (.001 and .002). Overall, the differences in Type I error are small enough to be considered

negligible and to conclude that there is no meaningful difference between the Type I error results of the two methods.

With respect to power, again similar values and trends are seen between the two studies. The largest difference in power is .107 with the MH-Modified method having the higher power. In general, at the smallest sample size the MH-Modified method has slightly higher power. At the middle and largest sample sizes, when the groups are equal in size the pooled method used by Goodman et al. (2011) has slightly higher power and when the group sizes are unequal the results are somewhat variable.

In the next chapter, the overall results of this simulation study are summarized and discussed specifically within the context of the studies of Allen and Donoghue (1996) and Goodman et al. (2011) which this study was intended to follow up and extend. Comparisons are also made with past studies that have investigated the MH and the IRT-Lord's methods in complete data. In addition, the study limitations are discussed and ideas for future research are presented. Finally, the practical implications of the findings on identifying DIF in structurally missing data are discussed.

5 Conclusion

5.1 Review of the Purpose of the Dissertation

DIF analyses are frequently conducted on LSA data even though there is a paucity of research on the accuracy of DIF methods for analyzing DIF in structurally missing data in LSAs. The main purpose of this dissertation was to investigate the accuracy, measured by Type 1 error rates and power, of two DIF methods to analyze DIF in structurally missing data in LSAs.

Previously, only one DIF method, the observed-score MH method, has been investigated in these circumstances. This dissertation expanded the investigation of the MH method by examining three different approaches to forming the MH matching criterion for analyzing structurally missing data. Two of the MH approaches (forming the matching criterion by block and by booklet) had been studied previously. The third approach was a modification introduced in this study with the hypothesis that it may improve the Type 1 error and/or power of the MH method for analyzing BIB data. Additionally, this study is the first to investigate the accuracy of an IRT-based DIF method for identifying DIF in structurally missing data. The DIF methods were selected because they are commonly used in practice, they afford the opportunity to examine one CTT-based method and one IRT-based method, and because past research has indicated that they function optimally under different situations (such as different sample sizes and test lengths) in complete data.

Due to the paucity of research on DIF in structurally missing data little or no empirically-based guidance has been available for practitioners. Therefore one further purpose of this dissertation was to gain an understanding of the impact of certain factors

(previously shown to affect DIF detection in complete data) on DIF analyses in structurally missing data with a goal of informing practice. Understanding how DIF analysis methods perform across these factors can provide valuable information to guide LSA test developers, administrators, and DIF analysts.

In order to address the research purpose and goals, a simulation study was conducted. Five simulation factors (total sample size, percentage of DIF items, ratio of reference and focal group sample sizes, number of items in each block, and true differences in reference and focal group ability) were manipulated to investigate their effect on the Type I error rates and power of the methods investigated. The BIB booklet design, simulation factors and the levels of each factor were chosen to be representative of real LSA data (such as data from PISA, NAEP or TIMSS) and wherever possible to coincide with the two previous studies (Allen & Donoghue, 1996; Goodman et al., 2011) that investigated the MH method performance in structurally missing data.

5.2 Summary of Results

The results of this study present a complex picture of DIF analysis for structurally missing data. In general and averaged over all study conditions, the 3 MH approaches maintained better Type I error rates than the IRT-Lord's method. The IRT-Lord's method Type I error rate was inflated in more conditions than any of the MH approaches, and was particularly inflated when the focal and reference groups were unequal in size.

Once the Type I error rates across the study conditions were known it was possible to determine under which conditions power was relevant and then to consider Type I error and power jointly for each DIF method. These are the results that illuminate the study conditions in which each of the methods may accurately identify DIF items. In this study

accuracy refers to the ability to correctly identify items that are DIF (having high power) while not misidentifying items that are not DIF (having controlled Type I error). There were many study conditions in which none of the methods achieved high power while maintaining low Type I error rate. Importantly, none of the four methods had the combination of high power and low Type I error at the smallest sample size.

None of the MH approaches achieved the combination of high power and low Type I error rates when the reference and focal group mean ability levels differed by one standard deviation. This was a result of low power, not inflated Type I error. Further, none of the MH approaches achieved high power and low Type I error rates when there was a high proportion of DIF items in the matching variable. In the remaining study conditions the MH-Block method had more power than the other two MH approaches. The MH-Booklet approach failed to produce a MH statistic in some replications at the smallest sample size, and had the least power of any of the methods studied.

The IRT-Lord's method did not achieve a combination of high power and low Type 1 error rates in any of the study conditions in which the reference and focal group sample sizes were different. This was a result of low power at the 1200 sample size, whereas at the larger sample sizes it was a result of Type I error inflation.

In the remaining study conditions in which the group sample sizes were equal, the IRT-Lord's method had high power and low Type 1 error rates when the total sample size was either 4800 or 8400 except when there were a large proportion of DIF items and the group ability levels were equal. The exceptions occurred as a result of inflated Type I error rates. When the reference and focal group sample sizes were equal the IRT-Lord's method did not achieve the combination of high power and low Type I error rates in any

of the conditions in which the sample size was 1200. This was a result of low power rather than inflated Type I error rates.

5.3 IRT-Lord's Analyses

This dissertation is the first to investigate the performance of an IRT-based method for identifying DIF in structurally missing data, specifically the IRT-based Lord`s Wald-2 method as executed in the flexMIRT (Cai, 2012) software.

The most important finding related to the IRT-Lord`s method is the extreme Type I error inflation when the focal and reference group sample sizes are unbalanced, particularly when the combined sample size is large. This finding and the finding of low power for the small sample size leads to the conclusion that the IRT-Lord's method did not achieve the combination of high power and low Type I error rates in any of this study's conditions in which the focal and reference group sample sizes were not equal.

IRT-Lord's DIF analysis depends on accurate estimation of item parameters. It was established in the flexMIRT parameter recovery mini-study in this dissertation that flexMIRT was capable of accurately estimating the *b* parameters in complete data in per group sample sizes similar to those investigated in the main study of this dissertation. Therefore it is possible that the inflated Type 1 error rates and low power in the unequal group size conditions are the result of structurally missing data. However, if this were the case it seems likely that there would have been inflated Type I error rates and low power in all study conditions since all conditions involved analyzing structurally missing data. Yet, in many of the equal group size conditions the Type I error rates were not inflated and there was high power for identifying DIF items. The inflated Type I error and low power observed for the IRT-Lord's method in the unequal group size conditions may

therefore be due to disparity in the group sizes rather than to the presence of structurally missing data.

Also in the unequal group size conditions the IRT-Lord's Type I errors decreased as the proportion of DIF increased. This is an unexpected result since DIF studies using IRT-based methods in complete data have shown the Type I error rate increases as the amount of DIF contamination increases (see for example, Wang & Yeh, 2003). Again it appears that the unequal group size conditions, and not the presence of structurally missing data, are resulting in unusual Type I error patterns for the IRT-Lord's method. Alternatively, some interaction between unequal group size condition and the presence of structurally missing data may be causing Type I error inflation.

Last, it is important to note that the Type I errors of the IRT-Lord's method tended to increase as the sample size increased. In hypothesis testing a small effect can still be statistically significant, and as the sample size becomes larger even extremely small effects can be statistically significant. To correct for this problem, it is generally recommended that measures of effect size be used alongside hypothesis tests. As noted earlier, in the case of MH analyses a blended heuristic decision rule (Gómez-Benito et al., 2013; Zumbo, 2008) which combines a statistical significance test and an effect size measure is typically used to identify DIF items using the MH method (Penfield & Camilli, 2007; Zwick et al., 2013). This was done with the MH analyses in this dissertation. However, no empirically validated effect size measure for the IRT-Lord's DIF method currently exists (Kim & Oshima, 2013). Therefore, DIF items were identified by the IRT-Lord's method based on their significance level alone. Consequently, large sample sizes may have resulted in inflated Type 1 error through the identification of small

but non-zero amounts of DIF because an effect size was not used. Similarly, the power of DIF analyses has been shown to be inflated when effect size measures are not used (Jodoin & Gierl, 2001), and the high power observed for the IRT-Lord's method in this study may be partly attributable to the lack of an effect size measure.

On the other hand, the alpha level for a hypothesis test is said to be the probability that the test will lead to a Type I error (Gravetter & Wallnau, 2004). The selection of an alpha level of 0.05 in this study may have been expected to result in a Type I error rate of approximately 0.05. This was true in many of the equal group size conditions (even in the large sample sizes) but was not the case in the large sample size/unequal group size conditions. This finding again tends to support a conclusion that it is the unbalanced sample size or an interaction between unbalanced sample size and structurally missing data that may be leading to the Type I error inflation rather than the lack of an established effect size or the presence of structurally missing data.

5.4 Relationship of MH Results to Past Studies of MH in Structurally Missing Data

This dissertation extended previous research on the performance of the MH method to detect DIF in BIB data. Allen and Donoghue (1996) identified the success of four methods of selecting the MH matching variable (matching by block, booklet, pooled booklet, and extra-information) by comparing the MH transformed log-odds ratio on the delta scale, Δ_{MH} and the standard error of Δ_{MH} of those methods to the results of analysis of the complete data. Based on their comparisons, they recommended that a pooled booklet approach be used for forming the MH matching variable when analyzing DIF in BIB data.

This dissertation evaluated three methods of forming the MH matching variable for BIB data (matching by block, booklet, and modified pooled booklet) by comparing their Type I error rates and power. The results of this study confirm those of Allen and Donoghue (1996) that a pooled booklet approach (in this case the MH-Modified approach) represented an improvement over the MH-Booklet method. The MH-Modified approach resulted in lower Type I error and higher power compared to the MH-Booklet approach. Further, pooling information from all booklets an item was in led to fewer cells with sparse data with the result that there were no conditions in which the MH-Modified method failed to produce a MH statistic as had occurred with the MH-Booklet method.

However, unlike the findings of Allen and Donoghue (1996), this study concludes that pooling the data across booklets to form the matching variable is not superior to the MH-Block method. Although the Type I error values for the MH-Modified approach tend to be slightly lower than the MH-Block approach, both are within a nominal Type I error rate of 0.05 (except in a few conditions with 30% DIF items) and therefore both have acceptable Type I error rates. However, the power observed in this study for the MH-Block method tended to be slightly higher than that observed for the MH-Modified method.

Allen and Donoghue's (1996) analysis revealed low power (~41%) to detect true DIF, even in complete data. They suggested this may have been a result of generating items with a *c* parameter other than zero because the assumptions necessary for the MH statistics are not strictly met. Allen and Donoghue further found that there was relatively little power to detect DIF items when the difficulty parameter was 0, and no power to detect DIF when the DIF item difficulty parameter was 1 or 2. The average values of the

DIF item difficulty parameters in this dissertation varied across study conditions from 0.34 to 0.47 which is in the range at which Allen and Donoghue found low power. It is possible that the low power observed in this dissertation is a result of the non-zero *c* parameters and the values of the *b* parameters used to simulate the data (which overall had a mean of 0). Nonetheless, the item parameters used in this study were real parameters from the TIMSS 2007 mathematics assessment which are typical LSA item parameters. This may provide an indication that the low power observed in this study and in the Allen and Donoghue study may be more related to the nature of the data that typically arises in large scale assessments (that is, the data follows a 3PL model) and the unsuitability of the MH method for analyzing 3PL data rather than a result of structurally missing data.

Goodman et al. (2011) investigated the performance of the pooled MH method recommended by Allen and Donoghue (1996) in structurally-missing data with a BIB design. Similarly to Allen and Donoghue, Goodman et al. found low power for the pooled booklet method and suggested that the low power may have been a result of the large number of matching levels in the pooled MH method. Therefore the modification proposed in this dissertation reduced the number of matching levels to investigate whether power would increase as a result. However, a comparison of results across similar study conditions revealed that there was no improvement in Type I error and very little improvement in power. A very small gain in power occurred in the smallest sample size conditions, while no improvement or a decline in power was seen at the middle and largest sample sizes. Decreasing the number of levels of the matching variable did not appreciably affect the Type I error rate or power when pooling information across booklets in the conditions in this study.

This dissertation also confirms the findings of Goodman et al. (2011) regarding low power of the MH method at the smallest sample size. Goodman et al. found that the pooled booklet approach did not perform well in BIB data when the total sample size was 1200 whereas Allen and Donoghue (1996) found that the pooled booklet approach performed well in all of their study conditions. The difference in results between these studies can be attributed to the different sample sizes investigated in the two studies. The results of Goodman et al.'s pooled booklet approach and of the MH-Modified approach in this dissertation confirm that when the total sample size is 1200 both of the pooled booklet MH approaches have low power, but as the sample size increases the power increases. When the sample size reaches 4800 the MH-Modified method has at least a moderate degree of power, but only in conditions in which the focal and reference group ability levels are equal and there were a small proportion of DIF items.

Allen and Donoghue (1996) found the presence of DIF in the matching variable increased the value of the standard error of Δ_{MH} of the studied item resulting in the appearance of less DIF against the focal group. However, as their results were averaged over all proportions of items with DIF in their study it was not possible to ascertain how their results may have varied as the proportion of DIF items varied. Similarly, Goodman et al. (2011) reported Type I error and power results averaged over all levels of proportion of DIF items so it is also not possible to ascertain the impact of increasing the proportion of DIF items on Type I error and power from their results. This dissertation therefore extended these studies by investigating how the Type I error and power of the MH methods varied as the proportion of items with DIF increased thus providing information about the impact of increasing amounts of DIF in the MH matching variable. Type I error

rates of all three MH methods considered in this study did not tend to vary as the proportion of DIF items rose from 0% to 10% but did increase when 30% of the items were simulated to have DIF. The increase in Type I error rates was accompanied by a substantial decrease in power between the 10% DIF conditions and the 30% DIF conditions. This confirms observations of Allen and Donoghue that there appears to be less DIF against the focal group when there is a larger proportion of DIF items in the matching variable since the appearance of less DIF may be manifested as lower power.

Allen and Donoghue (1996) found that the number of items per block was only marginally important when comparing the mean and standard deviation of the differences between Δ_{MH} values based on complete data and block and pooled booklet analyses. Goodman et al. (2011) found that increasing the number of items per block led to small improvements in Type I error rates and power for the pooled booklet analysis of BIB data in the larger sample size conditions (6000 or 12000). However, in the smaller sample size condition (1200) power decreased as the number of items per block increased for the pooled booklet analysis of BIB data. The results of this dissertation tend to confirm the findings of Allen and Donoghue in that, averaged over all other study conditions, the Type I error rates and power of all three MH methods investigated remain approximately the same regardless of block size. Unlike the findings of Goodman et al., in the smallest sample size conditions (1200), the power of the MH-Modified method did not tend to decrease as the number of items per block increased. Across similar study conditions this was the only notable improvement of the MH-Modified method over the pooled booklet method recommended by Allen and Donoghue.

Neither the Allen and Donoghue (1996) nor the Goodman et al. (2011) studies varied the ability levels of the focal and reference groups. Therefore this dissertation extends the previous research by examining the influence of group ability levels on MH Type I error and power for analyzing BIB data. Results show that the Type I errors of the three MH approaches were fairly stable across both levels of focal group ability when there was little or no DIF, but when 30% of the items had DIF the MH Type I error tended to be higher when the group ability levels were equal than when the focal group had lower ability. An important finding of this study is the inadequate power of all three MH approaches investigated to identify DIF in BIB data across all other study conditions when the focal group ability is one standard deviation below the reference group ability.

Allen and Donoghue`s (1996) study did not vary the reference and focal group sample sizes (which were 5100 and 1050 respectively) therefore it is not possible to discern how their results may have varied if the ratio of group sample sizes had varied. Goodman et al. (2011) investigated two levels of sample size ratios: an equal group size condition and a condition in which the reference group was three times the size of the focal group. The MH-Modified results of this dissertation have confirmed the general findings of Goodman et al. that unequal group sizes tend to result in lower power for a pooled MH method with BIB data, particularly at the smallest sample size.

5.5 Relationship of Results to Past Studies of DIF in Complete Data

The next two sub-sections draw comparisons between this study and studies of DIF in complete data.

5.5.1 MH Method

A key issue with respect to using the MH method to analyze DIF in structurally missing data is that of how to form the matching variable. Once a decision has been made about the composition of the matching variable the MH method proceeds without any missing data (that is, only the items in each block or each booklet are included in the analysis, as described in Sections 3.4.1.1 to 3.4.1.3). Therefore, the Type I error and power results of this study may be expected to be similar to those seen in the research on analyses of complete data in similar study situations.

Given the large number of studies on the Type I error and power of the MH method for analyzing complete data, it is helpful to compare the results of this study to a summary or meta-analysis of past findings. A meta-analytic study of the MH method (Guilera, Gómez-Benito, Hidalgo, & Sánchez-Meca, 2013) analyzed a total of 55 original studies of MH Type I error rate and/or power to identify variables that contribute to the MH Type I error rate and power. Guilera et al. made the following conclusions with respect to the five factors studied in this dissertation: (1) test length has minimal impact on the Type I error and power of the MH method; (2) as the proportion of items with DIF increases, particularly above 20%, the Type I error tends to increase and the power decreases; (3) Type I error rates and power tend to increase with sample size; (4) the effect of unequal group sample sizes tends to increase Type I error rate and decrease power, although the overall effect is not large; and, (5) up to one standard deviation difference in group mean ability levels tends to increase the Type I error rate but has little impact on power.

The general findings on the power of the three MH methods investigated in this study are very similar to the meta-analytic findings of Guilera et al. (2013) with respect to the effects of test length (in this study the number of items in a block), proportion of DIF items, sample size, and the ratio of group sample sizes.

Unlike the findings of Guilera et al. (2013) the findings of this dissertation indicate that a difference in group mean ability levels reduces the power of all three MH methods to correctly identify DIF items. However, results of this study are similar to those of other studies (for example, Clauser, Mazor, & Hambleton , 1993; Güler & Penfield, 2009; Narayanan & Swaminathan, 1994) that used the same ability distributions as were used in this dissertation and found that correct identification rates of DIF items was lower when group ability levels differed by one standard deviation.

The general trends with respect to Type I error rates of the three MH methods investigated in this study are very similar to the meta-analytic findings of Guilera et al. (2013) with respect to the effects of test length and proportion of DIF items. However, this study has different findings with respect to the effects of sample size, the ratio of group sample sizes, and differences in group mean ability levels on the Type I error rates of the MH methods studied.

The Type I error of the MH-Booklet analyses tended to increase as the sample size increased, similar to the meta-analytic findings of Guilera et al. (2013). However, the Type I error rates of the MH-Block and MH-Modified methods tended to decrease as sample size increased. Further research may be needed to investigate why this occurred, although one potential explanation may be that Guilera et al. examined Type I error rate based solely on statistical significance whereas this dissertation employed a blended

decision rule based not only on statistical significance but also on an effect size measure. Using a blended rule may have controlled the Type I error rates better in the larger sample sizes associated with the MH-Block and MH-Modified methods as compared with the smaller sample sizes of the MH-Booklet method (since the number of examinees administered each booklet is only one-quarter the number of examinees administered each block in the BIB design used in this study).

With respect to the effect of unbalanced group sample sizes, Guilera et al. (2013) found that the effect of unequal group sample sizes tended to increase Type I error rate whereas this study found that the effect of unequal group sample sizes tended to vary with the total sample size and the proportion of DIF items in the matching variable. Few studies have investigated the effect of unbalanced group sample sizes on MH Type I error. Paek and Guo (2011) found that increasing the reference group sample size had little impact on Type I error. Miller and Oshima (1992) found that Type I error tended to increase when the group sizes were unequal, whereas Herrera and Gómez (2008) and Monahan and Ankenmann (2005) found that Type I error decreased when group sizes were unequal. These studies also showed that there are complex interactions between group sample size ratios and other factors such as total sample size, item difficulty and discrimination. The results of this study tend to confirm the complex interactions noted in these previous studies of DIF in complete data.

5.5.2 IRT-Lord's Method

This dissertation used the IRT-Lord's Wald-2 method which was amended by Langer (2008) to include the use of MML estimation and a supplemented EM algorithm and to allow for concurrent estimation of the focal group population parameters. This

updated method has been investigated in two previous studies (Langer, 2008; Woods, Cai, & Wang, 2013) that analyzed DIF in complete data.

In analyses of polytomous data with either 25% or 50% of the items simulated to have DIF, Woods et al. (2013) found inflated Type I error rates of the Wald-2 method, which were particularly high when 50% of the items were DIF. Inflation tended to increase with sample size, and both equal and unequal group size conditions were inflated using this method. Similar to the Woods et al. findings in complete data, this dissertation concludes that using the IRT-Lord's Wald-2 method in a DIF sweep (referred to as the "all-other" procedure) results in inflated Type I error rates in structurally missing data. However, there are two differences in the findings of this dissertation. First, in this dissertation the IRT-Lord's Wald-2 method had inflated Type I error rates when no DIF existed whereas Woods et al. did not have a no-DIF condition. Second, in this dissertation the inflation is higher when the group sample sizes are unbalanced than when they are equal.

Langer (2008) analyzed polytomous data (using a graded response model) and dichotomous data (using a 3PL model) and found no inflation in Type I error rates. However Langer's study simulated a smaller percentage of DIF items than Woods et al. (2013) and Langer only investigated equal group size conditions. This dissertation extends the two previous studies by investigating the use of the IRT-Lord's Wald-2 procedure in structurally missing dichotomous data and finding inflated Type I error rates, particularly in unequal group size conditions and low power at the smallest sample size.

More generally, looking beyond only those studies that specifically investigated the IRT-Lord's Wald-2 method amended by Langer (2008), research has shown that other

unidimensional IRT-based DIF methods that can be used to analyze dichotomous data (such as the original Lord's chi-square, Raju's area measures, and likelihood ratio tests) tend to perform similarly to each other and produce congruent DIF results (Kim & Cohen, 1995) when using an "all other" method similar to the Wald-2 method. Simulation studies show that when there is no DIF or when a test contains only a single DIF item the Type I error rate of those methods tends to be well controlled. However, as the proportion of DIF items increases the Type I error rate increases (see, for example: Ankenmann, Witt, & Dunbar, 1999; Kopf, Zeileis, & Strobl, 2013; Stark, Chernyshenko, & Drasgow, 2006; and Woods, 2009). This is particularly true in the case of uniform DIF in which all of the DIF items are simulated to favour one group (W.-C. Wang, Shih, & Sun, 2012), as is the case in this study. While this pattern was observed in this dissertation in some study conditions when the reference and focal groups were equal in size, the opposite pattern was observed when the focal and reference group sample sizes were not equal.

Using a method in which all items were initially constrained to be equal across groups and then freeing items one at a time (similarly to the IRT-Lord's Wald-2 method used in this study), Stark et al. (2006) found that the IRT-based likelihood ratio method was robust to the presence of true differences in ability between the groups. They also found that power increased as sample size increased but that Type I error rates also increased as sample size increased. The IRT-Lord's results of this study tend to confirm these results.

5.6 Limitations and Future Research Directions

5.6.1 Generalizability of Results

This study represents a preliminary yet extensive investigation into the accuracy of DIF detection methods in structurally missing data due to booklet design. The booklet design, the study factors and their levels were intended to be as representative as possible of real LSA data while still trying to stay within reasonable boundaries. As noted in the literature review however, there are many other factors that can have an impact on DIF analyses which were not investigated here, as well as other types of data designs, types of data, and other uses of structurally missing data that may be considered. It is important to consider the extent to which the results of this study may generalize to these other situations.

In this study I elected to replicate the BIB design used by Goodman et al. (2011) in order to extend the Goodman et al. study and be as similar as possible for the purpose of drawing comparisons across the studies. However, there are many sorts of large scale educational assessments that use different data designs than the one used in this study. It is not known to what extent the results of this study would generalize to other BIB designs. Future simulation studies may wish to vary aspects of the BIB design to investigate the impact of different BIB designs on identification of DIF items in large scale educational assessments.

This study was a preliminary study of an IRT DIF method and an extension of previous studies that had used MH to analyze uniform DIF in dichotomously-scored unidimensional data containing a moderate degree of DIF, for two groups only. Therefore these are the conditions under which this study was carried out. Of course, real LSA data

may also contain polytomously-scored data, may be multidimensional, and/or may exhibit uniform or non-uniform DIF of varying degrees. Further, it is sometimes desirable to analyze DIF for more than two policy-relevant groups simultaneously. Future studies may be designed to explore DIF analyses in structurally missing data with a broader range of data types, and for multiple groups with a view to establishing the degree of generalization of the results obtained here.

Other factors have been shown to have an impact on DIF detection, such as the values of the DIF item parameters and the degree of DIF. This dissertation used item parameters from the TIMSS 2007 mathematics assessment. The extent to which the results of this study would generalize to items from other LSAs or other content areas than mathematics is not known. Further, this dissertation simulated a moderate amount of DIF in the b parameter only. Studies of the MH method indicate that it is more powerful when there is a greater amount of DIF (i.e., when the value of the b parameter is shifted further upward for the focal group) therefore it may be of interest to replicate this study with varying degrees of DIF. Finally, related to the item parameters, this study simulated DIF similarly to Goodman et al. (2011) in that DIF was created in the first item(s) in each block rather than randomly across all items. Therefore the characteristics of the specific items at the beginning of each block may have had an impact on the study results. Future studies may wish to randomly select which items get DIF to remove any systematic impact of specific DIF items.

Although this research focused on issues related to large scale educational assessments that use BIB designs and produce structurally missing data there are many other uses and applications of structurally missing data for which DIF analyses may be

required. For example, in the field of health research structurally missing data can be found when data is combined across multiple health regions or across multi-site health studies that collect some similar and some dissimilar information on health questionnaires or in longitudinal studies in which not all questions are asked of all participants at each stage of data collection. In health and social science research the use of overlapping survey or questionnaire designs has been recommended to reduce costs and ease stress and burden on respondents (see, for example, Aasland, Olff, Falkum, Schweder, & Ursin, 1997). Recently it has been suggested that the number of items on contextual background questionnaires and surveys that accompany large scale educational assessments could be increased by creating different "questionnaire booklets" that would be rotated randomly throughout the respondents similarly to the current system of assigning and rotating the assessment booklets (Adams, Lietz, & Berezner, 2013). Another application of structurally missing data occurs in computerized adaptive testing and multistage computerized testing in which examinees are routed to different individual questions or blocks of questions depending on their response to earlier questions. It is not known how the results of this study may generalize to these other types of structurally missing data therefore future research could investigate DIF analyses in these types of applications of structurally missing data.

5.6.2 Limitations and Future Research Specific to IRT-Lord's Method

The results of the IRT-Lord's analyses in this study are deserving of further study. In particular, three areas for future research are suggested by this study's results. The first is related to the lack of an effect size measure for the IRT-Lord's method. It is possible that in the conditions in which power was not calculated for the IRT-Lord's method as a

result of inflated Type 1 error rates, if an effect size measure existed and could be used the Type 1 error rates may have been reduced to be within Bradley's (1978) moderate limit. In this case, power would have been interpreted and different conclusions about the accuracy of the IRT-Lord's method may have been reached. Unfortunately, until an effect size measure for the IRT-Lord's method has been developed DIF analysis results based on this method should be interpreted with caution as both Type 1 error and power may be overstated, especially when the group samples sizes are unequal.

One area of future research arising from the results of the IRT-Lord's analyses is related to the observed Type I error inflation and power in the unequal sample size conditions. This study investigated a very unbalanced sample size ratio which represents the case of an extremely small focal group relative to a much larger reference group. Future studies may investigate a variety of group size ratios to further examine the impact of unbalanced sample sizes on Type I error and power of the IRT-Lord's method.

Another area of future research arising from the results of the IRT-Lord's analyses is the unexpected finding that the Type I error decreased as the proportion of DIF items increased in the unequal group size conditions. It would be beneficial to conduct studies aimed at determining the mechanism of the apparent suppression that occurred in this study. In that regard, a comparison between a complete data analysis and analysis of structurally missing data across similar study conditions may help to determine whether the decrease in inflation was due to the unequal group sizes or due to structurally missing data.

One further consideration and potential limitation of this study related to the IRT-Lord's analyses is the impact of placing constraints and prior distributions on item

parameters to avoid problems with estimations or failure to converge. As noted in the methods section, specific prior distributions were selected for this study so as not to be overly restrictive yet still guide the parameter estimation process, and they were based on realistic values for multiple choice item parameters found in large scale assessments. Nonetheless, it is important to consider the possibility that different results might have been obtained for the IRT-Lord's method if different constraints had been placed. For example, Langer (2008) found that placing a prior on the lower asymptote resulted in conservative Type I error values. Therefore it would be beneficial to replicate this study using different prior values for the parameters. In addition, it may be possible to modify the IRT-Lord's analysis syntax to account for the unbalanced sample sizes, and this should also be explored in future research.

5.6.3 Identifying DIF in Small Samples

The sample sizes investigated in this study were carefully selected to represent commonly encountered per jurisdiction, per item, and per booklet sample sizes of LSAs such as PISA and NAEP. Even when an overall LSA sample size is large the *per item*, *per block* and *per booklet* sample sizes can be quite small and may fail to meet current recommendations for sample size for DIF analyses. Small sample sizes represent an ongoing challenge for DIF analysis and this dissertation has shown that at the smallest sample size none of the DIF methods investigated had sufficient power to detect DIF. In addition, at the smallest sample size the MH-Booklet method failed to provide a DIF statistic in many replications. One recommendation therefore, is that future studies investigate the use of DIF analysis methods appropriate for small sample sizes, with structurally missing data. Examples of DIF methods that may be promising for small

samples are: empirical Bayes approach for MH (Zwick, Thayer, & Lewis, 1999); Bayesian updating methods for MH (Zwick, Ye, & Isham, 2012); a modified Angoff Delta plot method (Magis & Facon, 2012); and the Cochran-Mantel-Haenszel method (Meyer, Huynh, & Seaman, 2004) which can be used with both dichotomous and polytomous items.

5.6.4 Relative Contributions of Factors and Potential Interactions

This dissertation provides an exploration of how the power and Type I error rates of the four methods are affected by the study factors investigated. As such, this dissertation was not designed to provide statistical information about the extent to which each of the studied factors contributes to variation in power and Type I error for each of the DIF methods. Nor was it designed to provide information about how the factors may interact to affect the power and Type I error rates. The results of this study point toward complex main effects and interactions amongst the factors investigated. Future studies may be designed specifically to further probe these complex issues, for example, through the use of factorial ANOVAs.

5.6.5 Purification

It is also important to note that purification methods are often recommended to be used in DIF analyses. Purification refers to a two-step analysis process whereby items that are identified as DIF items in the first step of the analysis are omitted from the matching variable and a second DIF analysis is conducted with the "purified" matching variable. In other words, the first step is intended to assist in the selection of items that are invariant across groups which can then be used to form a more valid matching variable for the second analysis. The use of purification procedures has been found to lead to less Type I

error and greater power with the MH method (Guilera et al., 2013). Purification methods have also been shown to be effective with IRT-based DIF methods although their effectiveness decreases with increasing proportions of DIF in the test (e.g. W.-C. Wang, Shih, & Sun, 2012; Woods, 2009). This dissertation did not use a purification method, and therefore represents a study of how the methods would perform in a "worst case scenario", when nothing is known about the DIF status of the items. This is similar to the first step in a purification process. While it is important to understand how each of these methods perform as a first step in purification, future studies may be directed at understanding how these methods perform in a two-step purification process for analyzing structurally missing data.

5.6.6 Future Research on Other DIF Methods

The results of this study reinforce the recommendations made by both Allen and Donoghue (1996) and Goodman et al. (2011) to continue researching other methods of analyzing DIF in structurally missing data. In addition to the low power at the smallest sample size, the MH method power suffered when there were true ability differences between the groups. And, as noted above, the low power of the MH method may be related to the nature of the data – that is, real LSA data can reasonably be assumed to follow a 3PL model and in this case the MH method may not be appropriate. Although the IRT-Lord's method performed better than the MH methods when the group ability levels differed, it did not perform well in any condition when the group sample sizes were different. Other DIF methods such as SIBTEST (Shealy & Stout, 1993) which may perform better in conditions in which the groups have different ability levels due to its regression coefficient method, logistic regression (Swaminathan & Rogers, 1990) or other

IRT DIF methods or IRT DIF software programs should be explored with structurally missing data. These other methods also have the advantage of being able to detect non-uniform DIF which may occur in large scale assessments that produce structurally missing data whereas the MH method is best suited to detecting uniform DIF only.

5.7 Recommendations for DIF Analysts

To be accurate DIF methods should maintain Type 1 error and exhibit good power: they should not mistakenly identify non-DIF items as DIF items and they should accurately identify true DIF items as DIF items. DIF analysts want to be assured that the methods they are using minimize Type I errors and maximize power to avoid serious implications of inaccurate DIF detection. The results of this study provide useful information and three specific suggestions for DIF analysts to help them choose the most accurate DIF analysis method given the particular data they have.

5.7.1 Select a Block-wise MH Matching Variable

One important question addressed in this dissertation was how to select a matching variable for MH analyses of structurally missing data. This is a critical measurement issue because the selection of the matching variable contributes to measurement accuracy in terms of achieving the best possible Type I error rates and power. The results of this dissertation indicate that forming the matching variable by block (the MH-Block method) leads to better Type I error rates and power than forming the matching variable by booklet or through a modified pooled booklet approach.

In the conditions studied in this dissertation the MH-Booklet method had the lowest power of the three MH methods for analyzing structurally missing data. Recall that the MH-Booklet analyses have the advantage of being based on a larger number of items

than the MH-Block method, but the disadvantage of being based on a smaller sample size. In addition to having the lowest power, the MH-Booklet method failed to produce a MH statistic in some of the small sample study conditions. Low power and failure to produce a MH statistic are problems that are a direct result of low sample size. It appears that any benefit of the additional reliability of the matching variable gained by including more items in the analysis as compared to the MH-Block method is overcome by the reduction in sample size. The MH-Booklet method is therefore not recommended for use in situations similar to the conditions in this study.

The MH-Block and MH-Modified methods performed very similarly in the conditions in this study although the MH-Modified method had modestly better Type I error and the MH-Block method had modestly better power. The MH-Modified approach cannot be recommended as a means of improving the low power previously observed for the pooled booklet approach. In addition, from a practical perspective, it may be computationally simpler to prepare data for analysis by MH-Block than by pooled booklet approaches such as the MH-Modified method. Based on joint Type I error and power analysis results the computationally simpler MH-Block method is preferable to the more complex MH-Modified method.

5.7.2 Use Two Methods to Identify DIF in Structurally Missing Data

Neither the MH-Block method (which is the preferred MH method based on the results of this dissertation) nor the IRT-Lord's method will provide analysts with the required assurance of a combination of good Type 1 error and power in all of the conditions studied in this dissertation. As noted earlier the power of the MH method suffered when there were true differences in group ability and the Type I error of the IRT-

Lord's method suffered and its power may have been inflated when the group sample sizes were unequal. As a result, the MH method cannot be recommended when true ability differences exist between the groups and the IRT-Lord's method cannot be recommended when the group sizes are unequal. In the absence of as yet unknown better methods of detecting DIF in structurally missing data it may be recommended that the MH-Block method be used in conditions in which the group ability levels might reasonably be expected to be similar and the IRT-Lord's method be used when the group sizes are equal and a difference in group abilities might be expected. Neither method can be recommended when the group sample sizes are unequal and the groups can be anticipated to have different ability levels.

While group sizes are known in advance of DIF analysis, group ability levels may not be. As a result it will be difficult to know in advance which method will be more accurate. It is therefore recommended that both procedures should be used to confirm the presence or absence of DIF. Items may be considered to be DIF and recommended for follow-up expert review if they are identified as DIF by both methods. Similarly, items may be considered to be DIF free if they are not flagged as DIF by both methods. This recommendation has been made previously in the context of analyzing DIF in complete data (Ercikan et al., 2004; Hambleton, 2006) and it is reiterated and reinforced here in the context of structurally missing data.

5.7.3 Analyzing DIF in Small Samples

An important finding of this dissertation was that none of the methods investigated had high power to detect DIF when the combined groups sample size was as small as 1200. Please recall from Chapter 3 that there is a difference between the total sample size

and the sample size included in each analysis for the four analysis methods as a result of the BIB design. Total sample sizes of 1200, 4800 and 8400 refer to sample sizes of 100, 400 and 700 in each MH-Booklet analysis. Those same total sample sizes refer to sample sizes of 400, 1600 and 2800 for the MH-Block and MH-Modified analysis whereas all 1200, 4800 or 8400 examinees are included in the IRT-Lord's analyses. Despite the differences in the number of examinees included in the analysis for each particular analysis method, when the total sample size was 1200 none of the methods had high power. Therefore it is recommended that total sample sizes of greater than 1200 be used whenever DIF analyses are required to be carried out for structurally missing data. In cases where total sample sizes of more than 1200 are not available, DIF analysis results should be interpreted with caution as there may be insufficient power to identify DIF when it does exist.

Another recommendation may be to use DIF methods that are more powerful in small sample sizes in addition to the MH-Block and IRT-Lord's methods (such as the methods suggested in Section 5.6.3). It should be noted however, that these methods have not yet been investigated for use with structurally missing data. Nonetheless they may be able to provide additional information to complement the MH and IRT-Lord's methods that have now been investigated with structurally missing data.

5.8 Recommendations for Test Developers and Administrators

None of the methods investigated in this dissertation maintained Type I error rates while simultaneously exhibiting high power to identify DIF across the variety of realistic situations simulated in this study. Put another way, some of the methods could detect DIF in some of the conditions but no one method investigated could detect DIF in all of the

study conditions. Even when combining two DIF methods there may conditions in which DIF in structurally missing data goes undetected, for example, when the groups' sample sizes and ability levels are not equal. Therefore, a key finding for test developers is that post-hoc or secondary DIF analyses of structurally missing data cannot be relied on to reduce bias or provide evidence of fairness and validity for group comparisons. Rather, extra care and attention should be paid at the item development and test construction stages to minimize the potential for bias.

Block size did not have a strong influence on Type I error and power for the BIB design investigated in this dissertation for either the MH or the IRT-Lord's methods. These results suggest that whether there are 10 or 20 items in a block should not be of primary concern (for the purposes of future DIF analyses) when developing tests that result in structurally missing data if the test design is similar to the one used in this study.

As noted earlier, sample sizes represent an ongoing challenge for DIF analysis, and test developers and administrators should bear this challenge in mind when determining sample sizes for large scale assessments that will result in structurally missing data. Even though the overall sample size (such as the country sample sizes in international assessments, the state sample size in NAEP or the per province sample size in the Canadian provinces that take part in PISA) may be very large it may nonetheless be too small to enable accurate DIF analyses of policy-relevant groups.

Two examples from the DIF literature help to elucidate the problems related to small sample sizes that may be encountered in DIF analyses of structurally missing data. Hauger and Sireci (2008) analyzed TIMSS 1999 science items for DIF between students tested in the dominant language and those tested in a second language in the United States,

Singapore and Iran. The total per country sample sizes ranged from approximately 5000 to 8700. Nonetheless, some of the science items had been administered to fewer than 200 of Hauger and Sireci's target examinees. Those items were not included in the analysis due to insufficient sample size. In a second example, Le (2009) investigated gender DIF in 60 test language groups from 50 countries that participated in PISA. Despite an overall sample size of about 83,000 examinees, one of the language groups that was analyzed for gender DIF had a sample size of approximately 450 males and females combined. In circumstances such as these, DIF analyses between policy-relevant groups may be compromised due to insufficient sample sizes. Three potential situations may result: the DIF analysis may not be conducted due to small sample sizes, the analysis may proceed with a sample size that may be too small to enable accurate detection of DIF items, or items may be dropped from the analysis resulting in loss of information and potential impact on the analysis of the remaining items.

The results of this dissertation raise some cause for concern about achieving adequate total sample sizes that will result in DIF analyses with enough power to detect DIF if it exists in structurally missing data. The total sample size may have to be considerably larger than current practice (which was discussed in Section 3.2 and exemplified in the preceding paragraph) to achieve the sub-sample sizes (that is, the sizes of policy-relevant groups within the total sample size) needed for accurate DIF analyses. And, as the conditions become more taxing for DIF analyses, for example when the group ability levels differ or when the group sample sizes are unbalanced or both of these conditions occur simultaneously, greater still sample sizes will be required to reach a reasonable power to detect DIF. However, achieving even larger sample sizes than

currently used will create additional burden on test developers and administrators. In the case of large-scale assessments that use structurally missing data it may not be feasible or reasonable to expect this would be possible to achieve. Nonetheless, the sample size findings in this dissertation are very important and deserving of further consideration.

Last, it should be noted that the sample size of policy-relevant groups may be quite small in comparison to a reference population. The presence of unbalanced focal and reference group sample sizes in this study led to inflated Type I error and low power of DIF analyses. Thus it may be recommended to oversample small policy-relevant groups to ensure more balanced group sizes for the purposes of DIF analyses.

5.9 Implications of Inaccurate DIF Detection

It is important to consider the implications of inaccurate DIF detection; that is, failing to identify real DIF items or misidentifying non-DIF items. Three perspectives are worth mentioning: a practical test development perspective, a research perspective, and a policy perspective.

There are practical implications and costs associated with falsely identifying DIF items and failing to identify true DIF items. The statistical finding of DIF is followed up and decisions are made about whether to keep, modify, or discard DIF items. As part of the decision-making process items that are identified as DIF may be extensively reviewed for content, curricular, cultural or linguistic differences by expert panels seeking to verify the potential for DIF and determine its source. These reviews can be costly and time-consuming. The misidentification of non-DIF items as DIF items (Type I errors) can have two important repercussions. First, item reviews are triggered unnecessarily resulting in potentially wasted expense and experts' time. Second, psychometrically-sound items may

be modified or discarded needlessly. This can upset the content balance and the construct equivalence of the test thereby affecting test validity. Failing to identify true DIF items through a lack of power also has important repercussions. Since DIF signals the potential for bias, undetected DIF represents undetected potential for bias. Items may be falsely determined to be functioning equivalently across groups, will not be subjected to expert review, and may remain on a test to the detriment of one group or another. These consequences will have an impact on test fairness and validity.

Large-scale assessments of achievement that result in structurally missing data (such as PISA, TIMSS, PIRLS and NAEP) are used, among other things, for educational research purposes. They are accompanied by contextual background questionnaires which capture data about factors related to educational systems, school settings, teaching practices, and student home background. Researchers seek to establish relationships between these factors and scores on the assessment as a means of determining which factors are associated with high or low achievement. If the achievement data contain undetected DIF the relationships may not hold for sub-groups within the data since the meaning attached to their achievement scores may be different. Other research is also carried out specifically to explore and/or compare the relationships between contextual factors and achievement for different sub-groups (such as students tested in their first language compared to second language learners, or those who are immigrant students compared to those who are not). Again, the results of undetected DIF may confound and undermine the findings of these research activities.

Other educational research activities use DIF methods to seek evidence of the extent to which results of assessments that use structurally missing data (such as PISA,

TIMSS, PIRLS and NAEP) are comparable across groups, for example cross-cultural or cross-linguistic groups. If the DIF method used in this research has a high Type I error rate items will be misidentified as DIF and conclusions may be drawn that the assessment is potentially biased when it may not be. Conversely, if the DIF method fails to detect DIF that exists in the data conclusions may be drawn that the assessment is valid for drawing comparisons across groups when in fact it may not be.

Finally, it is important to consider the accuracy of DIF detection from a policy perspective. The uses of large-scale assessments that have structurally missing data and the inferences drawn from them can have far-reaching and tremendously important implications for participating countries, regions or jurisdictions. They are used to examine whether educational systems are meeting society's needs, and whether the needs of policy-relevant groups of examinees (such as gender, race, language or cultural groups) are being met equitably. The results of these comparisons are used to derive educational policy and to evaluate and inform educational programming. International assessments such as PISA, TIMSS, and others are used to compare and rank order countries according to their scores on the assessment, sometimes with drastic impact on the educational policy of lower-achieving countries, such as "borrowing" educational policy from higher-achieving countries (Chung, 2010) and spurring massive educational reform and new educational standards (Ertl, 2006).

Clearly, cross-jurisdiction or cross-group comparisons arising from assessments that use structurally missing data can have significant implications and should be supported by a high degree of validity evidence. The validity evidence about score comparability gathered through DIF analyses needs to be accurate to support the types of

comparisons that lead to important policy decisions. It is important to note that the presence of DIF does not necessarily indicate overall bias in an assessment or that the absence of DIF assures an absence of bias. Nonetheless, it must be recognized that the failure to identify DIF if it exists, or the misidentification of items as DIF when they are not, may result in misinformed policy and misdirected educational programming.

5.10 Contributions of the Study

The results of this study provide significant contributions to the body of knowledge about analyzing DIF in structurally missing data. In many of this study's conditions (which had been selected to be as representative as possible of real situations that may arise when analyzing DIF in structurally missing data in the context of large scale assessments of achievement) the DIF methods investigated did not have adequate power to identify DIF that existed in the data. Further, when the total sample size was as small as 1200 none of the methods investigated had adequate power to identify DIF.

This study extended the only two previous studies of DIF in structurally missing data by investigating three ways of forming the MH matching variable for structurally missing data. Findings indicate that forming the MH matching variable by block provided the best overall accuracy in terms of Type I error and power under the conditions of this study. This dissertation is also the first study to investigate a latent variable approach, the IRT-Lord's method for detecting DIF in structurally missing data. It has contributed by illuminating study conditions in which the IRT-Lord's method has a good Type I error rate and power, and also by finding severely inflated Type I error and low power in unequal group size conditions. This study also contributes by suggesting particular areas

of future research for investigating the IRT-Lord's method for identifying DIF in structurally missing data.

This study has also contributed to the field by making recommendations for DIF analysts, test developers and test administrators. Three specific recommendations for DIF analysts have been made to help improve the accuracy of DIF analyses in structurally missing data. This dissertation has shown that some sample sizes typically employed in large scale assessments that make use of structurally missing data may be too small to allow accurate DIF detection. Therefore a specific recommendation has been made that test developers and administrators give serious consideration to minimum sample size requirements that will enable the most accurate DIF detection possible. Last, this dissertation has highlighted the need to use extra care and attention during item development to reduce the risk of bias because post hoc or secondary DIF analyses cannot be relied on to accurately detect DIF in a variety of realistic structurally missing data conditions.

5.11 Concluding Remarks

The results of large-scale educational assessments that use booklet designs that result in structurally missing data are used to guide educational policy. DIF analyses provide important validity evidence to support fairness and equity in testing for policy-relevant groups and to support comparisons of results across jurisdictions in large-scale assessments. The accuracy of DIF analyses is central to providing this important form of validity evidence. Nonetheless, there has been surprisingly little research attention paid to the accuracy of DIF analyses in structurally missing data.

This study has provided valuable evidence about the accuracy of DIF analysis methods in scenarios that were simulated to be as close as possible to real structurally missing data scenarios. Some of the methods could detect DIF in some of the conditions but no one method investigated could detect DIF in all of the study conditions. The most important findings of this study are that no one method examined in this study will accurately find DIF across a wide variety of realistic situations, and that the sample sizes of some existing large scale assessments are not large enough to allow accurate DIF detection using the methods that have been studied to date. Equally important, the results of this study point to an urgent need to continue investigating the accuracy of these and other DIF methods to support the current uses of large-scale assessments that produce structurally missing data.

References

Aasland, O. G., Olff, M., Falkum, E., Schweder, T., & Ursin, H. (1997). Health complaints and job stress in Norwegian physicians: The use of an overlapping questionnaire design. *Social Science & Medicine*, *45*(11), 1615–1629.

Adams, R., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. Retrieved from http://works.bepress.com.ezproxy.library.ubc.ca/ray_adams/38/

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement*, *33*(2), 231–251.

Angoff, W. H. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An Investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, *36*(4), 277–300.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*(1), 5–37. doi:10.1016/j.jsp.2009.10.001

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.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152.

Bratley, P., Fox, B. L., & Schrage, L. E. (1987). *Guide to simulation* (2nd ed.). New York: Springer.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement*, *19*(4), 309–321.

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

Chung, J. H. (2010). Finland, PISA, and the implications of international achievement studies on education policy. *International Perspectives on Education and Society*, *13*, 267–294.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *6*(4), 269–279.

Clauser, B., Mazor, K. M., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement*, *31*(1), 67–78.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J: L. Erlbaum Associates.

DeBoer, J. (2010). Why the fireworks?: Theoretical perspectives on the explosion in international assessments. In A. W. Wiseman (Ed.), *The impact of international achievement studies on national education policymaking: Vol. 13. International perspectives on education and society* (pp. 297-330). doi: 10.1108/S1479-3679(2010)0000013014

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education*, *15*(1), 15–31.

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics*, *34*(2), 149–170.

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 Associates.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*(2), 189–199.

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicológica*, (1), 107–132.

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

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301–321.

Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education, 32*(5), 619–634.

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement, 64*(6), 925 –936. doi:10.1177/0013164404267288

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5*(3), 43–53.

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*(4), 278–295.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*(3), 225–245.

Finch, H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: Power and type I error rates. *Applied Measurement in Education, 24*(4), 281–301. doi:10.1080/08957347.2011.607054

Finch, W. H. (2011). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement*, *71*(4), 663 –683. doi:10.1177/0013164410385226

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues & Practice*, *28*(3), 39–53.

Glas, C. A. W., & Geerlings, H. (2009). Psychometric aspects of pupil monitoring systems. *Studies in Educational Evaluation*, *35*(2-3), 83–88.

Gómez-Benito, J., Hidalgo, M. D., & Zumbo, B. D. (2013). Effectiveness of combining statistical tests and effect sizes when using logistic discriminant function regression to detect differential item functioning for polytomous items. *Educational and Psychological Measurement*, *73*, 875-897. Retrieved from http://epm.sagepub.com.ezproxy.library.ubc.ca/content/early/2013/06/20/0013164413492419

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

Goodman, J. T., Willse, J. T., Allen, N. L., & Klaric, J. S. (2011). Identification of differential item functioning in assessment booklet designs with structurally missing data. *Educational and Psychological Measurement, 71*(1), 80 –94. doi:10.1177/0013164410387341

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*(2), 197.

Gravetter, F. J., & Wallnau, L. B. (2004). *Statistics for the Behavioral Sciences* (6th Edition.). Belmont, CA: Thomson/Wadsworth.

Guilera, G., Gómez-Benito, J., Hidalgo, M. D., & Sánchez-Meca, J. (2013). Type I error and statistical power of the Mantel-Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods*, *18*(4), 553.

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement*, *46*(3), 314–329.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, *44*(11), S182–S188.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications, Inc.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125. doi:10.1177/014662169602000201

Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. *International Journal of Testing*, *8*(3), 237–250.

Herrera, A. N., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. *Quality & Quantity*, *42*(6), 739–755. doi:10.1007/s11135-006-9065-z

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903–915.

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: Lawrence Erlbaum Associates.

Hopstock, P. J., Pelczar, M. P., & Xie, H. (2011). *Technical report and user's guide for the Program for International Student Assessment (PISA): 2009 data files and database with U.S. specific variables* (No. NCES 2011-025). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from nces.ed.gov/surveys/pisa/pdf/2011025.pdf

Houts, C., & Cai, L. (2012). *flexMIRT TM users manual version 1.0: Flexible multilevel item factor analysis and test scoring*. Seattle, WA: Vector Psychometric Group.

Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 221–244). New York, NY: Springer-Verlag New York, Inc.

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*(4), 329–349.

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software*, *20*(10), 1–14.

Kaplan, D. (1995). The impact of BIB spiraling. Induced missing data patterns on goodness-of-fit tests in factor analysis. *Journal of Educational and Behavioral Statistics*, *20*(1), 69–82.

Kim, J., & Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement*, *73*(3), 458–470.

Kim, S.-H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, *8*(4), 291–312.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131–143.

Klinger, D. A., DeLuca, C., & Miller, T. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy, 76*, 1 – 34.

Kopf, J., Zeileis, A., & Strobl, C. (2013). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. Retrieved from http://epub.ub.uni-muenchen.de/17481/

Kubinger, K. D., Hohensinn, C., Hofer, S., Khorramdel, L., Frebort, M., Holocher-Ertl, S., … Sonnleitner, P. (2011). Designing the test booklets for Rasch model calibration in a large-scale assessment with reference to numerous moderator variables and several ability dimensions. *Educational Research and Evaluation*, *17*(6), 483–495. doi:10.1080/13803611.2011.632666

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Doctoral Dissertation). Retrieved from ProQuest Dissertations & Theses. (Publication number 3331000)

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing*, *9*(2), 122–133.

Leite, W., & Beretvas, S. N. (2010). The performance of multiple imputation for Likert-type items with missing data. *Journal of Modern Applied Statistical Methods*, *9*(1), 64–74.

Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and type I error in the detection of differential item functioning. *Educational and Psychological Measurement*, *72*(5), 847–861. doi:10.1177/0013164411432333

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions. *Psicológica: Revista de Metodología Y Psicología Experimental*, *30*(2), 343–370.

Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, *18*(2-3), 292 –326. doi:10.1177/0049124189018002004

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: Lawrence Erlbaum Associates.

Magis, D., Beland, S., & Raiche, G. (2012). difR: Collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics. R package version 4.2.

Magis, D., & Facon, B. (2012). Angoff's delta method revisited: Improving DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 302–321. doi:10.1111/j.2044-8317.2011.02025.x

Mazzeo, J., & von Davier, M. (n.d.). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results.* Retrieved from https://edsurveys.rti.org/PISA/documents/MazzeoPISA_Test_DesignReview_6_1_09.pdf

Mellenbergh, G. J. (2005). Item bias detection: Classical approaches. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioural science* (Vol. 2, pp. 967–970). New York, NY: John Wiley & Sons Ltd.

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

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement*, (16), 381–388. doi:10.1177/014662169201600410

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297–334.

Mislevy, R. J., & Wu, P. K. (1988). *Inferring examinee ability when some item responses are missing* (Research Report No. RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Wu, P. K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing.* (Research Report No. RR-96-30-ONR). Princeton, NJ: Educational Testing Service.

Monahan, P. O., & Ankenmann, R. D. (2005). Effect of unequal variances in proficiency distributions on type-I error of the Mantel-Haenszel chi-square test for differential item functioning. *Journal of Educational Measurement, 42*(2), 101–131.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). TIMSS 2011 Assessment Frameworks. *International Association for the*

Evaluation of Educational Achievement. Retrieved from http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED512411

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315–328.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257 –274. doi:10.1177/014662169602000306

OECD (2009), *PISA 2006 Technical Report*, PISA, OECD Publishing. doi: 10.1787/9789264048096-en

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from http://timss.bc.edu/timss2007/techreport.html

Paek, I., & Guo, H. (2011). Accuracy of DIF estimates and power in unbalanced designs using the Mantel–Haenszel DIF detection procedure. *Applied Psychological Measurement*, *35*(7), 518–535.

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement*, *29*(2), 150 –151. doi:10.1177/0146621603260686

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, Psychometrics, pp. 125–167). Amsterdam, The Netherlands: Elsevier.

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502.

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

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning. *Educational and Psychological Measurement*, *69*(1), 18 –34. doi:10.1177/0013164408318756

Rogers, H. J. (2005). Differential item functioning. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 1, pp. 485–490). New York, NY: John Wiley & Sons Ltd.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105–116.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142 –151. doi:10.3102/0013189X10363170

Sargent, R. G. (2005). Verification and validation of simulation models. In *Proceedings of the 2005 Winter Simulation Conference* (pp. 130–143). Retrieved from http://dl.acm.org.ezproxy.library.ubc.ca/citation.cfm?id=1162736

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

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

Sijtsma, K., & Junker, B. W. (2006). Item response theory: past performance, present developments, and future expectations. *Behaviormetrika*, *33*(1), 75–102.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, *20*(2), 148–166.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. F. Merenda, & Spielberger, C. D. (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–115). Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. doi:10.1111/j.1745-3984.1990.tb00754.x

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.

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

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. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement*, *28*(5), 317 –331. doi:10.1177/0146621604264870

Vanneman, A., Hamilton, L., Baldwin Anderson, J., & Rahman, T. (2009). *Achievement gaps: How Black and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress* (No. NCES 2009-455). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/studies/2009455.pdf

Wang, W. C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*(3), 221–261.

Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *72*(4), 687–708. doi:10.1177/0013164411426157

Wang, W. C., & Su, Y. H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*(2), 113–144.

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479–498. doi:10.1177/0146621603259902

Wiberg, M. (2009). Differential item functioning in mastery tests: A comparison of three methods using real data. *International Journal of Testing*, *9*(1), 41–59.

Willse, J. T., Goodman, J. T., Allen, N., & Klaric, J. (2008). Using structural equation modeling to examine group differences in assessment booklet designs with sparse data. *Applied Measurement in Education*, *21*(3), 253–272.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. doi:10.1177/0146621607314044

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*(3), 532–547.

Yoes, M. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model*. Saint Paul, MN: Assessment Systems Corporation. Retrieved from http://www.assess.com/docs/Yoes%201995%20report.pdf

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223–233.

Zumbo, B. D. (2008). *Statistical methods for investigating item bias in self-report measures.* Florence: Universita degli Studi di Firenze E-prints Archive. http://eprints.unifi.it/archive/00001639/.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement*, *24*(4), 293–308.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational and Behavioral Statistics, 15*(3), 185–197.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*(3), 233–251.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*(1), 55–66.

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

Zwick, R., Ye, L., & Isham, S. (2012). Improving Mantel–Haenszel DIF estimation through Bayesian updating. *Journal of Educational and Behavioral Statistics, 37*(5), 601–629.

Zwick, R., Ye, L., & Isham, S. (2013). *An investigation of the efficacy of criterion refinement procedures in Mantel-Haenszel DIF analysis* (No. ETS RR-13-16).

Princeton, NJ: Educational Testing Service. Retrieved from http://origin-www.ets.org/Media/Research/pdf/RR-13-16.pdf

Appendix 1: Item Parameters Used to Generate the Data

Item	Parameter			Item	Parameter			Item	Parameter		
	a	b	c		a	b	c		a	b	c
1*	0.66	-1.36	0.15	51	0.91	-0.48	0.21	101*	1.36	1.00	0.09
2*	0.80	0.29	0.17	52	1.19	0.15	0.21	102*	1.59	0.30	0.15
3*	0.68	-0.41	0.11	53	0.44	0.39	0.17	103*	1.11	-0.06	0.17
4*	0.86	-0.91	0.14	54	1.00	0.14	0.20	104*	1.21	0.61	0.17
5*	0.51	-1.74	0.15	55	1.30	1.05	0.11	105*	1.79	0.21	0.08
6*	0.97	-1.65	0.15	56	1.65	0.69	0.25	106*	0.90	0.64	0.17
7*	1.05	-0.44	0.17	57	1.16	0.04	0.20	107*	1.31	-0.32	0.10
8*	1.20	-1.12	0.15	58	1.06	-0.46	0.20	108*	1.52	0.43	0.20
9*	1.26	-0.20	0.14	59	1.25	-0.72	0.16	109*	1.48	0.67	0.16
10*	1.16	-0.77	0.17	60	0.91	0.07	0.19	110*	1.47	0.35	0.14
11	0.89	-1.05	0.11	61*	1.15	0.51	0.16	111	0.85	-0.04	0.16
12	0.68	0.14	0.21	62*	0.62	-0.20	0.13	112	1.50	-0.32	0.20
13	0.82	-1.65	0.14	63*	0.84	-0.82	0.13	113	1.44	-0.37	0.14
14	0.71	-0.40	0.11	64*	0.69	0.11	0.13	114	1.07	0.43	0.11
15	0.76	-1.07	0.12	65*	1.33	0.52	0.11	115	1.22	0.24	0.15
16	0.95	0.90	0.19	66*	1.05	0.12	0.04	116	1.54	0.91	0.12
17	0.75	-0.63	0.14	67*	1.51	-0.14	0.06	117	1.18	0.67	0.18
18	0.59	-1.00	0.17	68*	0.98	-0.42	0.06	118	1.40	0.04	0.15
19	0.82	-1.81	0.12	69*	2.37	-0.34	0.21	119	1.07	0.56	0.11
20	0.67	0.27	0.15	70*	1.85	0.10	0.09	120	0.99	1.09	0.12
21*	0.90	-0.60	0.13	71	1.47	-1.04	0.13	121*	0.78	-0.77	0.16
22*	0.77	-0.54	0.14	72	1.39	-0.79	0.09	122*	0.85	-0.02	0.16
23*	0.80	-0.48	0.21	73	1.38	-0.73	0.06	123*	1.13	-0.62	0.12
24*	0.89	-0.06	0.14	74	0.87	0.08	0.09	124*	1.17	0.61	0.24
25*	0.95	0.42	0.15	75	0.80	-0.02	0.04	125*	1.00	0.60	0.16
26*	0.45	0.95	0.17	76	1.04	-0.86	0.07	126*	0.91	0.64	0.14
27*	0.92	-0.89	0.12	77	1.08	-0.25	0.03	127*	0.90	-0.42	0.20
28*	1.16	1.07	0.15	78	1.24	0.52	0.09	128*	0.93	0.69	0.23
29*	0.69	-0.89	0.15	79	1.46	0.22	0.12	129*	1.34	0.30	0.17
30*	0.66	0.07	0.21	80	1.82	0.13	0.10	130*	0.78	-0.23	0.19
31	0.67	-0.75	0.16	81*	1.62	0.14	0.09	131	1.13	-0.59	0.13
32	1.26	0.36	0.20	82*	1.20	-0.19	0.05	132	1.23	0.21	0.24
33	1.25	-0.46	0.12	83*	1.02	-0.20	0.05	133	1.31	0.17	0.19
34	0.94	-0.50	0.15	84*	0.90	0.87	0.10	134	0.73	-0.70	0.15
35	0.91	0.58	0.14	85*	1.32	-0.34	0.05	135	1.39	-0.02	0.18
36	1.08	-0.36	0.15	86*	1.19	0.56	0.10	136	1.03	0.16	0.24
37	0.57	0.64	0.19	87*	0.90	0.16	0.04	137	1.76	0.95	0.11
38	0.44	0.22	0.14	88*	1.04	0.91	0.11	138	1.45	0.44	0.19
39	1.14	0.31	0.18	89*	1.15	-0.03	0.18	139	0.94	-0.33	0.14
40	1.39	0.21	0.15	90*	1.31	-0.09	0.12	140	1.21	0.19	0.18
41*	1.31	0.36	0.15	91	1.44	0.40	0.15	141*	0.95	0.11	0.23
42*	0.65	-0.20	0.20	92	1.28	-0.11	0.12	142*	1.24	-0.15	0.11
43*	1.21	0.13	0.21	93	0.99	0.85	0.10	143*	1.28	-0.31	0.22
44*	1.07	-0.24	0.18	94	1.16	0.03	0.12	144*	1.18	0.55	0.14
45*	1.23	-0.22	0.13	95	0.97	0.83	0.16	145*	1.26	0.78	0.24
46*	1.29	-0.27	0.13	96	0.97	-0.07	0.16	146*	1.21	-0.04	0.21
47*	0.91	0.63	0.11	97	0.62	0.66	0.15	147*	0.96	-1.30	0.15
48*	0.72	-0.51	0.22	98	1.31	0.09	0.10	148*	1.20	-0.24	0.18
49*	1.16	0.29	0.16	99	1.31	0.22	0.15	149*	0.80	-0.94	0.12
50*	0.79	0.69	0.22	100	1.77	0.13	0.11	150*	1.45	-0.89	0.18

*denotes items used in the 10-item per block condition

Appendix 1: Item Parameters (Continued)

Item	Parameter			Item	Parameter			Item	Parameter		
	a	b	c		a	b	c		a	b	c
151	0.79	-0.37	0.11	161*	1.09	0.51	0.18	171	0.87	0.67	0.12
152	0.96	0.13	0.26	162*	0.93	0.21	0.15	172	1.73	0.92	0.18
153	1.23	0.47	0.22	163*	0.80	-0.37	0.18	173	0.77	0.52	0.16
154	1.25	0.62	0.09	164*	0.84	1.04	0.14	174	0.95	0.21	0.19
155	1.23	0.97	0.10	165*	1.02	-0.62	0.14	175	1.19	-0.63	0.14
156	1.07	0.36	0.15	166*	1.00	-0.53	0.14	176	0.69	-1.17	0.15
157	1.50	0.52	0.22	167*	0.97	-0.03	0.09	177	1.61	0.34	0.17
158	0.95	-1.17	0.14	168*	0.88	-1.38	0.15	178	0.71	-0.25	0.21
159	0.63	1.55	0.17	169*	1.34	-0.58	0.13	179	1.17	0.69	0.15
160	0.61	-0.55	0.19	170*	1.03	0.68	0.20	180	1.23	0.94	0.18

*denotes items used in the 10-item per block condition

Appendix 2: Verification of Simulation Procedures

R Code to check data file dimensions:			**Purpose of check is to ensure:**
Number of files created	=	Total number of files expected to be created	No missing files; all conditions as expected
Number of unique study seeds	=	Total number of study seeds	Each replication of each study condition has a unique seed
Length of group index vector	=	Length of sample size for simulation condition	Every data file grouping vector length matches sample size for the simulation condition
Length of a parameter vector	=	Block size x block count (i.e. number of items in this simulation condition)	Every data file number of a parameters matches number of items in the simulation condition
Length of b parameter vector (reference group b pars)	=	Block size x block count (i.e. number of items in this simulation condition)	Every data file number of reference group b parameters matches number of items in the simulation condition
Length of b_dif parameter vector (focal group b pars)	=	Block size x block count (i.e. number of items in this simulation condition)	Every data file number of focal group b parameters matches number of items in the simulation condition
Length of c parameter vector	=	Block size x block count (i.e. number of items in this simulation condition)	Every data file number of c parameters matches number of items in the simulation condition
Number of rows in BIB data	=	Sample size in complete data for simulation condition	Every BIB data file has correct sample size for the simulation condition (used in all MH DIF analyses)
Number of rows for focal group in BIB data	=	Total sample size x focal ratio for simulation condition	Every BIB focal group data file for flexMIRT analyses has the correct focal group sample size for the simulation condition
Number of rows for reference group in BIB data	=	Total sample size x (1 – focal group sample size) for simulation condition	Every BIB reference group data file for flexMIRT analyses has the correct reference group sample size for the simulation condition

R Code to check data file dimensions:			**Purpose of check is to ensure:**
Number of columns in BIB data	=	Block size x block count(i.e. number of items in this simulation condition)	Every BIB data file has correct number of items for the simulation condition (used in all MH DIF analyses)
Number of columns in BIB focal group data	=	Block size x block count(i.e. number of items in this simulation condition)	Every BIB focal group data file for flexMIRT analyses has the correct number of items for the simulation condition
Number of columns in BIB reference group data	=	Block size x block count(i.e. number of items in this simulation condition)	Every BIB reference group data file for flexMIRT analyses has the correct number of items for the simulation condition
Number of rows in full data	=	Total sample size for simulation condition	Every full data file (prior to removing structurally missing data) contains correct sample size for the simulation condition
Number of rows in full focal group data	=	Total sample size x focal ratio for simulation condition	Every full focal group data file (prior to removing structurally missing data) has the correct focal group sample size for the simulation condition
Number of rows in full reference group data	=	Total sample size x (1 – focal group sample size) for simulation condition	Every full reference group data file (prior to removing structurally missing data) has the correct reference group sample size for the simulation condition
Number of columns in full data	=	Block size x block count(i.e. number of items in this simulation condition)	Every full data file (prior to removing structurally missing data) contains correct number of items for the simulation condition
Number of columns in full focal group data	=	Block size x block count(i.e. number of items in this simulation condition)	Every full focal group data file (prior to removing structurally missing data) has the correct number of items for the simulation condition
Number of columns in full reference group data	=	Block size x block count(i.e. number of items in this simulation condition)	Every full reference group data file (prior to removing structurally missing data) has the correct number of items for the simulation condition

R Code to check data value consistency:			**Purpose of check is to ensure:**
b parameters for items without DIF	=	Original b parameters	Every data file non-DIF items should have original value of b parameter (no DIF added)
b parameters for items with DIF	=	b parameter value +0.48 IF AND ONLY IF the item is supposed to be a DIF item	Every data file DIF items should have original value of b parameter plus 0.48
Number of items without DIF	=	Block size x block count x (1-DIF ratio)	Every data file has the correct number of non-DIF items for the simulation conditions
Number of items with DIF	=	Block size x block count x DIF ratio	Every data file has the correct number of DIF items for the simulation conditions
BIB focal group data	=	Data that corresponds to the group index for the focal group	BIB Focal group data is correctly identified
BIB reference group data	=	Data that corresponds to the group index for the reference group	BIB Reference group data is correctly identified
Full focal group data	=	Data that corresponds to the group index for the focal group	Full Focal group data is correctly identified
Full reference group data	=	Data that corresponds to the group index for the reference group	Full Reference group data is correctly identified
Calculations and visual checks to verify contents of data files:			**Purpose of check is to ensure:**
Calculated matching score in Excel for modified pooled booklet MH analysis in randomly selected data files			R code was correctly calculating the modified matching score (x_j)
Analyzed MH DIF in randomly selected data files using DIFAS software (Penfield, 2005)			R code MH analysis produces same results as other software; checks accuracy of DIF results
Visually inspected randomly selected BIB data files			BIB data files contain dichotomous data and NA values that follow the simulation booklet design shown in Figure 2 of dissertation
Visually inspected randomly selected data files used in the 3 MH analysis methods			Data used in the MH analyses has the expected structure (no missing values, correct number of items according to method and simulation condition, correct number of focal and reference group members)

Calculations and visual checks to verify contents of data files:	**Purpose of check is to ensure:**
Visually inspected randomly selected data files used in the flexMIRT analyses	Data used in the flexMIRT analyses has the expected structure (contains missing values and dichotomous data that follows the simulation booklet design shown in Figure 2 of dissertation; correct number of focal and reference group members; correct number of items)
Visually inspected randomly selected flexMIRT output files	Correct identification of focal and reference group thetas; No error messages or warnings; Reasonableness of group theta estimates; Reasonableness of item parameter estimates; Number of DIF items reasonable according to Type I error and power calculations.
Calculated power and Type I error rates on randomly selected output files in Excel	Correct calculations of power and Type I error rates obtained through the simulation process.