


4-5-2017

Extending the Model with Internal Restrictions on Item Difficulty (MIRID) to Study Differential Item Functioning

Yong "Isaac" Li

University of South Florida, liy1@mail.usf.edu

Follow this and additional works at: <http://scholarcommons.usf.edu/etd>

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), and the [Educational Psychology Commons](#)

Scholar Commons Citation

Li, Yong "Isaac", "Extending the Model with Internal Restrictions on Item Difficulty (MIRID) to Study Differential Item Functioning" (2017). *Graduate Theses and Dissertations*.
<http://scholarcommons.usf.edu/etd/6724>

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.

Extending the Model with Internal Restrictions on Item Difficulty (MIRID) to Study Differential
Item Functioning

by

Yong “Isaac” Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in Curriculum and Instruction with an emphasis in
Educational Measurement and Research
Department of Educational and Psychological Studies
College of Education
University of South Florida

Co-Major Professor: Yi-Hsin Chen, Ph.D.
Co-Major Professor: Jeffrey D. Kromrey, Ph.D.
John Ferron, Ph.D.
Stephen Stark, Ph.D.

Date of Approval:
March 9, 2017

Keywords: differential item functioning, validity, item response modeling, Rasch models, the
MIRID

Copyright © 2017, Yong “Isaac” Li

DEDICATION

To my late mother and sister. I miss you.

ACKNOWLEDGEMENTS

I wish to express my most profound gratitude to my co-major professor, Dr. Yi-Hsin Chen, who brought me into the beautiful world of cognitive diagnostic models and componential IRT modeling. Through his limitless patience and generosity, he guided me along the way of academic pursuit. Whenever I was lost or felt hopeless, he was there and ready to lend me a hand. Without his wisdom, insightful teaching, encouragement, and conviction, I would not have been able to come this far. It has been a pleasant and inspiring experience for me to learn from him and collaborate with him.

I reserve my special thanks and appreciation for my co-major professor, Dr. Jeffrey Kromrey. The epitome of erudition and calmness for people around, I have massive regret for not having been able to learn more from him. But I wish to thank him for steering my research in the right course and granting me important opportunities to gain experience in research. In particular, I am grateful for him to spend valuable hours reviewing and advising my manuscripts and computer code time and again.

My committee members, Dr. John Ferron and Dr. Stephen Stark, have given this study their time and scholarly contributions. I am extremely thankful for their thoughtful suggestions and careful review that have improved the quality of my research. At key moments of this journey, they both provided important, brilliant directions which are characteristic of their academic excellence.

Also, I am indebted to the Research Computing staff at USF, Dr. John Desantis, Dr. Anthony Green, and Joseph Botto, for their tireless assistance and trouble-shooting with regards to my usage of the computing cluster.

Last but not least, I thank my wife, Tina, and daughter, Anwyn, for their patience, understanding, and suffering in these years.

TABLE OF CONTENTS

LIST OF TABLES	iii
LIST OF FIGURES	v
ABSTRACT	viii
CHAPTER ONE INTRODUCTION	1
Differential Item Functioning in the Context of the MIRID	1
DIF as the Consequence of Construct Multidimensionality	4
Purpose of the Study	6
Significance of the Study	9
Definitions	10
CHAPTER TWO LITERATURE REVIEW	13
The Generalized Linear Mixed Models	14
The Linear Regression Model	14
Linear Mixed Models	15
Generalized Linear Models	16
Generalized Linear Mixed Models (GLMMs)	17
Nonlinear Mixed Models (NLMMs)	20
Item Response Modeling in the GLMMs Framework	20
The Model with Internal Restrictions on Item Difficulty (MIRID)	25
MIRID in the Generalized Statistical Models Framework	25
Estimation Methods and Computer Programs for the MIRID	35
Differential Item Functioning	40
The Mantel-Haenszel Procedure	42
Simultaneous Item Bias Test (SIBTEST)	42
The Model-based DIF Approach for the MIRID	43
Individual-item Level Model Specification (DIF)	43
Item-group Level Models (DFFm and DFFc)	45
Component Weight Model (DWF)	46
CHAPTER THREE METHOD	48
Design of the Research	48
The Scope	48
The Simulation Study	49

Implementation	55
Data Generation	55
Estimation	58
Analysis Procedures.....	60
Evaluation Procedures	61
 CHAPTER FOUR RESULTS	 65
Parameter Recovery of the Standard MIRID.....	65
Results for the Proposed Differential Functioning Models	66
Recovery of the True DIF (Delta) Parameters	67
Recovery of Zero-value DIF Parameters	75
Type I Error Control and Power of the MIRID Differential Functioning Models	80
Results from Fitting the Mismatched Differential Functioning Models.....	96
False Detection Rates of the Mismatched MIRID Differential Functioning Models	97
Non-DIF Parameter Recovery of the Mismatched Models	113
 CHAPTER FIVE DISCUSSION.....	 137
Summary.....	137
Findings.....	138
Item-Level DIF	139
Group-Level DIF	141
Mismatching DIF Model and DIF Source	142
Implications.....	143
Implications for Content Researchers.....	144
Implications for Methodology Researchers	146
Limitations and Future Studies	148
Conclusions.....	150
 REFERENCES	 151
 APPENDIX A: Examples of Analysis Code in SAS.....	 161
 APPENDIX B: Estimation Bias and RMSEs of the Zero-value DIF Parameters of the MIRID DIF, DFFc, DFFm, and DWF Models.....	 163
 APPENDIX C: Estimation Bias and rmseS of the Model Parameters of the MIRID DFFc, DFFm, and DWF Models.....	 165
 APPENDIX D: Type I Error Rates and Power Obtained from Estimating Item DIF Parameters by Component and by Item.....	 169

LIST OF TABLES

Table 1. Item Families and Component Items	29
Table 2. Item Predictor Matrix.....	29
Table 3. Component weight Matrix for One Item Family	30
Table 4. Parameter Estimates from the Example Study	34
Table 5. Simulation Conditions	56
Table 6. Recovery of Item Location Parameters of the Standard MIRID	68
Table 7 Bias of the Non-zero DIF Parameter Estimates under the MIRID DIF Model	69
Table 8. RMSEs of the Non-zero Delta Parameter Estimates under the MIRID DIF Model.....	70
Table 9. Bias of the DIF-related Item Location Parameters under the MIRID DIF	72
Table 10 RMSE of the Delta Parameters under the MIRID DFFc, DFFm, and DWF Models...	74
Table 11 Bias of the Delta Parameters under the MIRID DFFc, DFFm, and DWF Models.....	75
Table 12 Type I Error Rates for MIRID DIF over 500 Replications.....	83
Table 13 Power of the MIRID DIF over 500 Replications.....	85
Table 14 Type I Error Rates for the MIRID DFFc over 500 Replications	86
Table 15 Power of the MIRID DFFc over 500 Replications	88
Table 16 Type I Error Rates for the MIRID DFFm Model over 500 Replications	89
Table 17 Power of the MIRID DFFm over 500 Replications.....	91
Table 18 Type I Error Rates for the MIRID DWF over 500 Replications	92
Table 19 Power of the MIRID DWF over 500 Replications	94

Table 20 Hypothesiswise Type I Error Rates for the Four Proposed MIRID Models	95
Table 21 Experimentwise Type I Error Rates after Hochberg Adjustment for the Four Proposed MIRID Models.....	95
Table 22 False Detection Rates when the MIRID DIF Model Was Applied to the DFFc Data.	101
Table 23 False Detection Rates when the MIRID DIF Model Was Applied to the DFFm Data	102
Table 24 False Detection Rates when the MIRID DIF Model Was Applied to the DWF Data.	103
Table 25 False Detection Rates when the MIRID DFFc Model Was Applied to the DIF Data.	106
Table 26 False Detection Rates when the MIRID DFFc Model Was Applied to the DFFm Data	107
Table 27 False Detection Rates when the MIRID DFFc Model Was Applied to the DWF Data	108
Table 28 False Detection Rates when the MIRID DFFm Model Was Applied to the DIF Data	111
Table 29 False Detection Rates when the MIRID DFFm Model Was Applied to the DFFc Data	112
Table 30 False Detection Rates when the MIRID DFFm Model Was Applied to the DWF Data	113
Table 31 False Detection Rates when the MIRID DWF Model Was Applied to the DIF Data	116
Table 32 False Detection Rates when the MIRID DWF Model Was Applied to the DFFc Data	117
Table 33 False Detection Rates when the MIRID DWF Model Was Applied to the DFFm Data	118

LIST OF FIGURES

Figure 1. Average bias of the non-zero DIF parameter estimates under the MIRID DIF model by sample size	69
Figure 2. Average RMSEs of the non-zero DIF parameter estimates under the MIRID DIF model by sample size	70
Figure 3. Recovery of the location parameter of component items under the MIRID DIF model by sample and impact.....	73
Figure 4. RMSEs of the non-zero DIF parameter estimates under the MIRID DFFc by sample size	76
Figure 5. Average RMSEs of the non-zero DIF parameter estimates under the MIRID DFFm by sample size	76
Figure 6. RMSEs of the non-zero DIF parameter estimates under the MIRID DWF by sample size	76
Figure 7. Average bias in estimation of zero-value DIF parameters in the MIRID DIF model...	77
Figure 8. Average RMSEs in estimation of zero-value DIF parameters in the MIRID DIF model	77
Figure 9. Average bias in estimation of zero-value DIF parameters in the MIRID DFFc, DFFm, and DWF models (smaller delta conditions)	78
Figure 10. Average bias in estimation of zero-value DIF parameters in the MIRID DFFc, DFFm, and DWF models (larger delta conditions).....	78
Figure 11. Average RMSEs in estimation of zero-value DIF parameters in the MIRID DFFc, DFFm, and DWF models (smaller delta conditions)	79
Figure 12. Average RMSEs in estimation of zero-value DIF parameters in the MIRID DFFc, DFFm, and DWF models (larger delta conditions)	79
Figure 13. The MIRID DIF model experiment-wise Type I error rates after Hochberg adjustment by sample size	84

Figure 14. The MIRID DFFc model experimentwise Type I error rates after Hochberg adjustment by sample size	87
Figure 15. The MIRID DFFm model experimentwise Type I error rates after Hochberg adjustment by sample size	90
Figure 16. The MIRID DWF model experimentwise Type I error rates after Hochberg adjustment by sample size	93
Figure 17. Bias of the 30 estimated DIF parameters of the MIRID DIF model when fitted to data with different sources of differential functioning	99
Figure 18. RMSE of the 30 estimated DIF parameters of the MIRID DIF model when fitted to data with different sources of differential functioning.....	100
Figure 19. Bias of the three estimated DIF parameters of the MIRID DFFc model when fitted to data with different sources of differential functioning.....	104
Figure 20. RMSE of the three estimated DIF parameters of the MIRID DFFc model when fitted to data with different sources of differential functioning	105
Figure 21. Bias of the ten estimated DIF parameters of the MIRID DFFm model when fitted to data with different sources of differential functioning.....	109
Figure 22. RMSE of the ten estimated DIF parameters of the MIRID DFFm model when fitted to data with different sources of differential functioning.....	110
Figure 23. Bias of the three estimated DIF parameters of the MIRID DWF model when fitted to data with different sources of differential functioning.....	114
Figure 24. RMSE of the three estimated DIF parameters of the MIRID DWF model when fitted to data with different sources of differential functioning	115
Figure 25. Bias of the Estimated Item Locations when the DIF Model Was Fitted to Different Models.....	119
Figure 26. RMSE of the Estimated Item Locations when the DIF Model Was Fitted to Different Models.....	120
Figure 27. Bias of the Estimated Item Locations when the DFFc Model Was Fitted to Different Models.....	121
Figure 28. RMSE of the Estimated Item Locations when the DFFc Model Was Fitted to Different Models.....	122

Figure 29. Bias of the Estimated Item Locations when the DFFm Model Was Fitted to Different Models.....	123
Figure 30. RMSE of the Estimated Item Locations when the DFFm Model Was Fitted to Different Models.....	124
Figure 31. Bias of the Estimated Item Locations when the DWF Model Was Fitted to Different Models.....	125
Figure 32. RMSE of the Estimated Item Locations when the DWF Model Was Fitted to Different Models.....	126
Figure 33. Bias of the Estimated Component Weights when the DIF Model Was Fitted to Different Models.....	128
Figure 34. RMSE of the Estimated Component Weights when the DIF Model Was Fitted to Different Models.....	129
Figure 35. Bias of the Estimated Component Weights when the DFFc Model Was Fitted to Different Models.....	130
Figure 36. RMSE of the Estimated Component Weights when the DFFc Model Was Fitted to Different Models.....	131
Figure 37. Bias of the Estimated Component Weights when the DFFm Model Was Fitted to Different Models.....	132
Figure 38. RMSE of the Estimated Component Weights when the DFFm Model Was Fitted to Different Models.....	133
Figure 39. Bias of the Estimated Component Weights when the DWF Model Was Fitted to Different Models.....	134
Figure 40. RMSE of the Estimated Component Weights when the DWF Model Was Fitted to Different Models.....	135

ABSTRACT

Differential item functioning (DIF) is a psychometric issue routinely considered in educational and psychological assessment. However, it has not been studied in the context of a recently developed componential statistical model, the model with internal restrictions on item difficulty (MIRID; Butter, De Boeck, & Verhelst, 1998). Because the MIRID requires test questions measuring either single or multiple cognitive processes, it creates a complex environment for which traditional DIF methods may be inappropriate. This dissertation sought to extend the MIRID framework to detect DIF at the item-group level and the individual-item level. Such a model-based approach can increase the interpretability of DIF statistics by focusing on item characteristics as potential sources of DIF. In particular, group-level DIF may reveal comparative group strengths in certain secondary constructs. A simulation study was conducted to examine under different conditions parameter recovery, Type I error rates, and power of the proposed approach. Factors manipulated included sample size, magnitude of DIF, distributional characteristics of the groups, and the MIRID DIF models corresponding to discrete sources of differential functioning. The impact of studying DIF using wrong models was investigated.

The results from the recovery study of the MIRID DIF model indicate that the four delta (i.e., non-zero value DIF) parameters were underestimated whereas item locations of the four associated items were overestimated. Bias and RMSE were significantly greater when delta was larger; larger sample size reduced RMSE substantially while the effects from the impact factor were neither strong nor consistent. Hypothesiswise and adjusted experimentwise Type I error

rates were controlled in smaller delta conditions but not in larger delta conditions as estimates of zero-value DIF parameters were significantly different from zero. Detection power of the DIF model was weak. Estimates of the delta parameters of the three group-level DIF models, the MIRID differential functioning in components (DFFc), the MIRID differential functioning in item families (DFFm), and the MIRID differential functioning in component weights (DFW), were acceptable in general. They had good hypothesiswise and adjusted experimentwise Type I error control across all conditions and overall achieved excellent detection power.

When fitting the proposed models to mismatched data, the false detection rates were mostly beyond the Bradley criterion because the zero-value DIF parameters in the mismatched model were not estimated adequately, especially in larger delta conditions. Recovery of item locations and component weights was also not adequate in larger delta conditions. Estimation of these parameters was more or less affected adversely by the DIF effect simulated in the mismatched data. To study DIF in MIRID data using the model-based approach, therefore, more research is necessary to determine the appropriate procedure or model to implement, especially for item-level differential functioning.

CHAPTER ONE

INTRODUCTION

Modeling cognitive or behavioral constructs underlying item responses with decomposed processes has become an actively researched area in educational and psychological measurement. Different from the traditional practice of trait organization, such componential approaches recognize intermediate item responses that represent processes as well as the final responses and aim to explain final responses with properties of the intermediate responses. Observations on the “components” supply additional information on more dimensions than can be obtained by focusing on the trait alone. A prominent componential approach, the linear logistic test model (LLTM; Fischer, 1973, 1977), has been adopted by practitioners from many disciplines and served as the platform for development of newer psychometric models, such as the model with internal restrictions on item difficulty (MIRID; Butter, De Boeck, & Verhelst, 1998). Originally a member of the Rasch family of models, the MIRID has had many extensions which provide a new context for investigating measurement issues common to education and psychology. This dissertation concentrated on one of them, differential item functioning (DIF).

Differential Item Functioning in the Context of the MIRID

At the heart of test fairness and construct validity is the issue of differential item functioning, which has received extensive research in the past decades. A general definition provided by Chang, Mazzeo, & Roussos (1996) considers an item as having DIF when conditional on the latent trait being measured, one group of respondents having on average

higher probability than the other group to give a particular response to the item. Commonly seen in the literature are DIF analyses to answer the question whether particular items became unfairly easier for members of the focal group than for the reference group.

Numerous procedures have been devised and implemented for DIF detection. Among the most often used are the Mantel-Haenszel (MH) (Holland & Thayer, 1988) method, which is non-parametric, and several model-based procedures, such as Lord's chi-square method (Lord, 1980), Raju's (1990) area measures, and the likelihood ratio test (Thissen, Steinberg, & Wainer, 1993). These procedures have been proven successful in discovering DIF but not as much in helping to understand its possible causes. Moreover, it is unclear the extent to which these traditional approaches are effective when faced with the unique characteristics of the MIRID.

A confirmative approach to examine how underlying processes affect a complex behavioral outcome, the MIRID assumes that the construct of interest can be decomposed into mental processes represented by different items and that there is a definitive between-item relationship similar to that of the LLTM with disparate groups of items retaining one or more properties. For example, performance on questions of addition, subtraction, and multiplication are expected to influence response to items subsuming all these operations. Tests designed in the framework of the MIRID are made of a number of item families, each of which consists of one or more component items measuring individual processes (subtasks) as well as a composite item requiring all these subtasks to answer. Every item family corresponds to a "situation" describing the construct and shares the same number of component items. The difficulty parameter of the composite item is defined as weighted summation of the parameters of all component items in this family plus an intercept. In other words, the MIRID assumes that the difficulty of a

composite item is explained perfectly by the difficulty parameters of all the component items in its family and there is no room for error.

The unique data structure resulting from this linear relationship gives rise to a complex DIF environment where different types of DIF may exist. A basic form occurs when multiple items from different item families and different components exhibit DIF. The sporadicity and lack of pattern therein would make the cause of this kind of individual item DIF difficult to explain. However, we are faced with another kind of DIF when classes of items sharing the same properties presumably contribute to the differential effect in a substantive way. Modeling this form of DIF (“differential facet functioning (DFF)”); Englehard, 1972) summarizes individual-item DIF in a parsimonious fashion on the basis of commonality amongst these items. In the MIRID, there are two facets of item groups (“domains”), components and situations (item families), and one or more categories of each or both can potentially cause DIF.

The DFF exhibited by item families (“DFFm”) can be labeled “situational” since each family of items describes a common setting. In a hypothetical case of measuring altruistic abstinence, the questions could inquire about sacrificing for children (the common setting) where women would be expected to outscore men. Consequently, items of the same family would have their location parameters differing between males and females and violate the null hypothesis of equal component item parameters across groups. The other type of group-level DIF is found with items within the same component (“DFFc”) and can be labeled “componential”, which comes into being when *ALL* or *MOST* of the items under one component or multiple components carry parameters that favor certain manifest groups over the others. Again, with the same example, suppose the construct of altruistic abstinence can be broken down into such factors as willpower,

faith, and life satisfaction, one would expect that respondents of certain cultural background tend to answer more strongly questions measuring a particular factor than the others.

The fourth potential source of DIF in the MIRID is the weight parameter accompanying each component (component weight), including the intercept. This form of componential DIF (differential weight functioning or “DWF”) occurs when component items contribute to the difficulty of composite item varyingly from group to group; that is, a component (or its items on average) may be more important for the focal group than the reference group. Greater complexity ensues when more than one type of DIF happens. For example, when there is differential effect with one item family and a component at once, unequal location parameters coincide with unequal component weights across groups to create DIF parameters on two dimensions that will be challenging to detect.

Any type of DIF in component items will lead to DIF in associated composite items whose parameters must be estimated through the linear relationship between component and composite items. When DIF occurs in component weights alone, only composite items will exhibit item-level DIF.

In summary, the MIRID presents different types of possible DIF scenarios for manifest groups, including at both individual-item level and at item-group level, which further breaks down into componential DIF, situational DIF, and component weight DIF, as well as concurrence of any of these DIF types. Such complexity must be heeded during investigation.

DIF as the Consequence of Construct Multidimensionality

In measurement practice, the construct of interest can be viewed as comprising more than one dimension. This does not imply necessarily applying multidimensional psychometric models; rather, it provides a framework for study of differential item functioning. In the context of the

MIRID, the properties shared by component items or situational items can be viewed as secondary dimensions to the primary or the target trait being tested. Therefore, DFF and DWF may be thought of as the consequence of secondary dimensions not accounted for in normal assessments.

Numerous studies adopted the DIF framework of secondary dimensions (Ackerman, 1992; Bolt & Stout, 1996; Douglas, Roussos, & Stout, 1996; Finch, 2005; Roussos & Stout, 1996; Shealy & Stout, 1993a, 1993b; Xie & Wilson, 2008). According to Shealy and Stout (1993a & 1993b), a secondary dimension is considered auxiliary and to cause benign DIF if it complements the primary dimension intended to be measured; on the contrary, if the item property is irrelevant to the construct, it is a nuisance dimension that leads to adverse DIF. Substantive analysis may be called upon to determine whether the DIF is benign or adverse. By retaining the auxiliary dimension of items and eliminating items with adverse DIF, construct validity and fairness of the test will be improved at once. Although the MIRID was conceived as a unidimensional model, it can be considered to some extent multidimensional if each component is treated as a dimension of the trait of interest. Thus the multidimensional DIF framework proposed by Shealy and Stout (1993a) can be applied in this research to untangle the complexity.

By applying the paradigm of multidimensionality as the potential cause of differential functioning, differential functioning of items in the MIRID can be studied in the statistical framework of generalized linear and nonlinear mixed models (GLMMs and NLMs) by adding grouping or interaction covariates. The nonlinearity in the difficulty of composite items results from the product of two parameters to be estimated: the latent item predictor and component weight, which makes up the fixed effects part of the MIRID. Such a model-based approach is

primarily based on Meulders and Xie (2004), who modeled DIF by including person-by-item interactions as predictors in the NLMM. Their work extended from a general DIF approach, differential facet functioning (Englehard, 1992), which allows various procedures to explore DIF at the level of item groups. The person property means group membership and the item property the subtask it measures so that their interaction reflects the difference in ability between the focal and the reference group.

Purpose of the Study

DIF studies are an important means to preserve test fairness and construct validity and have produced a voluminous literature and numerous detection methods. The MIRID and its extensions can become powerful tools to study cognitive and affective attributes underlying latent traits. However, due to its unique componential structure, applying conventional detection methods may lead to incorrect conclusions failing to account for the relationship between component and composite items. Like with other less applied psychometric models, in-depth knowledge on their statistical properties and appropriate and effective implementation procedures, such as ways of parameterization, methods to study differential functioning must be developed before they are ready for applied data application.

No research on differential functioning in the context of the MIRID has been published so far. Wang and Jin (2010) postulated an approach of a likelihood ratio test based on nested models to study DIF in component items and composite items. Their method of DIF detection would need to be repeated for every studied item and will neither point out the potential source of nor explain the differential functioning. They did not carry out the study and no other research on this topic has been found.

In the past decades, research in differential functioning has gone through three phases of evolution in focus and efforts (Zumbo, 2007). In the first phase (“conceptual”), the emphasis was to distinguish between item bias and item impact by identifying item characteristics that were either intended to be assessed and thus causes of group differences in performance as a result of impact or unintended to be assessed so as to making the item unfairly easier and biased for one group over another. The focus of the next phase (“statistical”) was on establishing procedures to detect DIF with sufficient power and acceptable Type I error rates. Nevertheless, many standard DIF procedures do not lend themselves to identification of potential causes behind DIF after statistical analysis has flagged certain items and created the disjoint between techniques and meaning. In the current third phase (“substantive”), however, the efforts in DIF studies are poured into discovering reasons behind identified DIF for distinct groups of equal trait levels by ways of purposeful modeling and content analysis.

The substantive approach to studying DIF is suitable for the complexity and various types of differential functioning with the MIRID. It avoids the often adopted practice of removing from the test any items flagged by statistical detection for the DIF exhibited may be benign instead of adverse, which are often confounded in reality. Removing items with adverse DIF improves test fairness. Conversely, keeping DIF items on auxiliary secondary dimensions improves construct validity of the test as it indicates that these items are capable of differentiating groups on valid grounds that are part of the construct being measured. If these dimensions as possible explanation of the benign DIF expectedly confirm the design theory behind the MIRID instrument and increase construct knowledge, such items or their improved version need to be kept in the test. On the other hand, keeping these “good” items saves the unnecessary cost that may be incurred from modifying or replacing them.

The objective of this research was two-fold. One was to propose and examine a model-based approach to detecting and potentially explaining in the context of the MIRID differential functioning by taking into account its possible discrete sources, including individual items (DIF), item facets formed by components (DFFc) and item families (DFFm), and component weight (DWF). The proposed approach is formulated by extending the standard MIRID, a member of the Rasch family of models, to include differential effects in the nonlinear mixed models and was fitted to data structure of the MIRID. Since the extended MIRID does not include an item discrimination parameter, only uniform DIF was studied.

The other objective was to investigate the effect from applying a DIF model to study differential functioning caused by a different source. For example, applying a MIRID DIF model to a data set where there is differential functioning present with one component weight. Would the DWF be conducive to statistically significant parameter estimates of individual item DIF and thus mislead the researchers? Similarly, would considerable DIF on one or two items lead to significant nonzero estimate of differential effects with an item family when the DFFm model is applied? Addressing these questions would provide insight into potential impact from fitting the wrong DIF model in conducting DIF investigation and alert researchers about the importance of following the correct procedures in DIF study with the MIRID as well as about the importance of substantive analysis.

In empirical settings, more than one type of differential functioning can occur as a result of the unique data structure of the MIRID. For example, one component may be more important for the focal group than the reference group (DWF) when several individual items exhibit DIF favoring either group (DIF). However, it was decided that as the initial MIRID DIF exploration

this research would lay the foundation by tackling each source separately; the investigation of their concurrence is left for future research.

Given the aforementioned research purposes, this study sought to answer the following questions:

- 1.) Does the proposed MIRID differential functioning models maintain Type I error control? When it is under control, what is the power of the MIRID DIF, DFFc, DFFm, and DFW models in detecting differential functioning of different sources?
- 2.) How accurate are the parameter estimates of these models, including the DIF parameters, item locations, component weights, and impact?
- 3.) How do the following factors affect the performance of the proposed differential functioning approach, including sample size, DIF magnitude, and group differences in trait level?

To investigate the effect of applying the incorrect model to study differential functioning in the context of the MIRID, the following questions were addressed based on the analysis results:

- 4.) How well are the model parameters estimated if the wrong models are fitted to the data? Are they more adversely impacted under some conditions than others?
- 5.) Are any of the estimates of the incorrectly specified DIF parameters statistically significant? Which differential effects in the data produce the most misleading findings when the unmatched model is fitted?

Significance of the Study

This model-based DIF approach in the context of the MIRID may be able to identify DIF in individual items as well as item groups simultaneously in keeping with the model's structure

made up of component and composite items. It may differentiate between DIF that exists in item families and differential functioning exhibited by one or more components while taking into account of the group difference in the latent trait. This approach may be capable of identifying group weakness and strength on a part of the measured construct as a consequence of the presence of benign DIF. This utility, aided with substantive analysis, may enable interpretation of certain types of DIF by locating possible causes. Hypothetically, for instance, while a traditional DIF detection procedure locates a number of individual items with significant DIF, the proposed approach would be able to identify significant group-level DIF in one component even if only some of the associated individual items display small amounts of differential functioning. By means of this, differential functioning in separate items is summarized and explained by using item properties shared by the item group.

The MIRID is a promising model to uncover the operational mechanism behind cognitive and psychological responses. Developing a compatible and pragmatic DIF investigation approach will increase the understanding and use of this componential modeling tool. From the perspective of applied research, the contribution from successfully developing the DIF approach will be the improvement of psychometric qualities of the MIRID through enhancing the fairness and construct validity at once and thus make it more accessible to researchers.

Definitions

Linear Logistic Test Model (LLTM): A statistical model which was first introduced by Fischer (1973) as a member of Rasch family of models. It re-expresses item difficulty as a weighted summative composite of the cognitive attributes identified a priori as underlying item responses. Parameters to estimate include coefficients of the every attribute.

The MIRID: The model with internal restrictions on item difficulty was developed on the basis of the LLTM (Butter et al., 1998). Instead of every item embodying one or more properties (attributes) to some extent like with the LLTM, the MIRID supposes one or more groups of items each of which reflects an attribute. The other items not representing the supposed item properties have their location parameters defined as weighted sums of difficulty of the former type of items.

Components: Item properties (a.k.a. attributes, strategies, mental processes, etc.) in the MIRID are called components.

Component Items: Since a component is embodied by a group of items in the MIRID, these items are labeled as component items, each of which belongs with only one component.

Composite items: The other type of items in the MIRID that requires all component processes to answer and whose parameter is linearly related to those of its associated component items.

Differential facets functioning (DFF): DIF shown by groups (facets) of items. In this study, it refers to DIF from either components or item families or both.

Differential item functioning (DIF): With statistical evidence, the presence of differential performance on an item by two or more groups of examinees conditioning on their trait levels. In this study, it also refers to the sporadic DIF exhibited by individual items.

Differential weight functioning (DWF): DIF shown by component weights. This definition is limited to the MIRID only.

Item families: A group of items led by a composite item and its associated component items, each of which reflects only one component. A family of items may describe a situation (or scenario) of the measured construct.

Component weights: The importance of each component item in the linear relationship that determines the location of the composite item. Items within a component share the same component weight.

CHAPTER TWO

LITERATURE REVIEW

This chapter is divided into four sections. Firstly, the statistical framework of the generalized linear/nonlinear mixed models (GLMMs/NLMMS) and its relationship with item response models are introduced to provide the backdrop for the MIRID, which is presented in the second section along with its estimation methods. Next, the issues around differential item functioning are discussed in the third section. On the basis of these sections, the definition and specification of the MIRID DIF approach are given in the final part.

The purpose of traditional item response theory (IRT) models is to estimate from response data parameters of individual persons and items located on the same latent scale. A modern perspective conceptualizes item response models in a broader, generalized statistical framework, namely, the generalized linear mixed models (GLMMs) and nonlinear mixed models (NLMMS). Such a framework allows item and person parameters to be estimated in either fixed or random terms, introduces into the model effects from item and person properties, and is capable of incorporating a range of existing measurement models. The power of this modeling framework lies in the fact that in addition to location of persons and items on the scale of the latent trait, item characteristics (e.g., cognitive processes, format) and person attributes (e.g. demographics, psychological differences) can be integrated in the statistical model as either fixed or random effects. Under the traditional paradigm, however, this explanatory stage of analysis is not conducted until IRT calibration has been completed and is often performed separately in the

form of regression. In the context of item response models, the NLMMs are equivalent to the GLMMs plus the item discrimination parameter and are essentially the same family of models.

The Generalized Linear Mixed Models

Four types of statistical models are reviewed in this section, including the simplest linear regression model, the more complex but more general linear mixed models and generalized linear models, and finally the generalized linear mixed models (GLMMs), which are extensions of the other three models. After showing the connections between these models, the formulation of the GLMMs for dichotomous data will be presented. Since GLMMs are closely related to NLMMs as a special case with a slope parameter of one (Kackman, 2000), this discussion will concentrate on the GLMMs.

The Linear Regression Model

One of the elementary statistical techniques, linear regression is often used to model the relationship between a single variable y , the dependent or outcome variable, and one or more independent variables, also called regressors or covariates, x_1, \dots, x_k , with K as the number of independent variables. When $K = 1$, it is simple regression but when $K > 1$ it becomes multiple regression. By assuming a linear relationship between the dependent and independent variables, regression analysis describes the structure of the data, makes predictions over future observations, and explains the effect on the outcome variable from the covariates included in the model.

The linear regression model can be represented in matrix terms as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (1)$$

where with n observations $\mathbf{y} = (y_1, \dots, y_n)^T$, the unknown regression parameters $\boldsymbol{\beta} = (\beta_0, \dots, \beta_k)^T$, the error term $\boldsymbol{\epsilon} = (\epsilon_0, \dots, \epsilon_n)^T$, and the design matrix is

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ 1 & x_{31} & x_{32} & \dots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}.$$

The estimation of β can be carried out using the least square approach, which defines its best estimate as one that minimizes the sum of the squared errors. The error term ϵ is typically assumed to be independent and identically normally distributed with mean of zero and variance of σ^2 , that is to say, $\epsilon \sim N(0, \sigma^2 I)$. However, this is not always a reasonable assumption.

Linear Mixed Models

In linear regression models, effects from the predictor variables are considered unchanging (fixed), such as treatment and control in a biological experiment, and all observations are assumed independent of each other. However, for analysis of data in a nested structure, particularly, clustered (a.k.a. hierarchical) data or longitudinal (or repeated measures) data, this assumption is inappropriate. In such data, level-one observations (individuals or repeated observations) are nested within level-two observations (clusters or subjects), which may be nested within even-higher clusters. To account for the correlation within data, randomness needs to be included in modeling of cluster effects. Statistical models containing both fixed effects and random effects are mixed models. In matrix notation, linear mixed models can be represented as

$$y = X\beta + Z\gamma + \epsilon, \quad (2)$$

where y is a vector of n observations, β is a vector of fixed effects, and γ is a vector of random effects. The random effects represent the influence of subjects/persons on their repeated observations that is not captured by the observed covariates. These are treated as random effects because the sampled subjects are thought to represent a population of subjects. X is the design

matrix for the fixed effects relating observations y to β , and Z is the design matrix for the random effects relating y to γ . γ and ϵ are assumed to be unrelated with mean of zero and covariance matrices \mathbf{G} and \mathbf{R} , respectively, both of which are sources of random variation within the model.

The expectation and variance of y are presented as

$$E[y] = X\beta \quad (3)$$

$$Var[y] = \mathbf{ZGZ}^T + \mathbf{R} . \quad (4)$$

When both random sources are assumed to be normally distributed $\gamma \sim N(0, \mathbf{G})$ and $\epsilon \sim N(0, \mathbf{R})$, the observed dependent variable is also normally distributed as $y \sim N[X\beta, Var(y)]$.

Generalized Linear Models

The linear regression model describes the relationship between the dependent variable and the fixed effect through a linear function (linearity), which assumes constant variance (homoscedasticity) and normal distribution of error terms (normality). Relaxing these assumptions but including in the model only fixed effects extends the linear regression model into generalized linear models (GLMs) (cf. Nelder. & Wedderburn, 1972; McCullagh & Nelder, 1989).

The class of GLMs allows for several types of dependent variables such as continuous, dichotomous, counts, etc., which are assumed to be generated from a particular member of the exponential distribution family, such as binomial, normal, and Poisson, and incorporate disparate statistical methods like linear regression, logistic regression, and Poisson regression. Three key components of a generalized linear model are identified as the linear predictor, a link function, and a form of the measurement variance as a function of the predicted value. The linear predictor

is denoted as $\eta = X\beta$, where X is the design matrix and β the fixed effects. The link function $f_{link}(\cdot)$ converts the expected value of the outcome variable to the linear predictor, that is,

$$f_{link}[E(Y)] = f_{link}[\mu] = \eta \quad (5)$$

This transformed expected value is predicted by a linear combination of observed variables.

Finally, the last key component specifies the variance of the dependent variable as a function of the mean:

$$Var(Y) = Var(\mu) = Var[f_{link}^{-1}(\eta)] \quad (6)$$

When the distribution of the outcome variable is assumed normal, the inverse of the identity link function is η ; when the distribution is binomial, the inverse link becomes

$$\mu = \frac{e^\eta}{1 + e^\eta}. \quad (7)$$

To capture non-systematic variability, a variance function is defined for the GLMs. For normal data it is one; but for binomial data, assuming dispersion parameter is one,

$$Var(Y) = \mu(1 - \mu) \quad (8)$$

Generalized Linear Mixed Models (GLMMs)

A GLMM is a particular type of the linear mixed models which extends the generalized linear models by incorporating both fixed and random effects in the linear predictor (Breslow & Clayton, 1993; McCulloch & Searle, 2001; Stroup, 2012).

As in the mixed models, the fixed and random effects are combined to form a linear predictor,

$$\eta = X\beta + Z\gamma \quad (9)$$

where X is the design matrix for the fixed effects β and Z the design matrix for the random effects γ . With a vector of residuals ϵ added, the observed outcome data can be modeled as

$$y = X\beta + Z\gamma + \epsilon = \eta + \epsilon \quad (10)$$

The random effects γ are assumed to be normally distributed with a mean of zero and variance matrix \mathbf{G} (so called G-side variance), which are denoted as $\gamma \sim N(0, \mathbf{G})$.

As with the linear mixed model, common link functions available for GLMMs f_{link} , depending on distributions, include identity (for normal distribution), logit, and probit (for binomial distribution).

Unlike GLMs, which specify for y a probability distribution from the exponential family, GLMMs assume a conditional response distribution that depicts the relationship between linear predictor and observations,

$$y|\gamma \sim [f_{link}^{-1}(\eta), \mathbf{R}], \quad (11)$$

that is, the conditional distribution of y given random effects γ , often called the error distribution, has a mean of $f_{link}^{-1}(\eta)$ and variance \mathbf{R} (referred to as R-side variance). Related, the expected values of the dependent variables of a GLMM are

$$E[y|\gamma] = \mu = f_{link}^{-1}(X\beta + Z\gamma) = f_{link}^{-1}(\eta) \quad (12)$$

That is, the conditional mean of the outcome variable depends on the linear predictor through the inverse link function. In addition, the covariance matrix \mathbf{R} depends on the conditional mean μ through a variance function $\mu(1 - \mu)/n$.

Mixed models for continuous normal dependent variables have been well researched (e.g., Laird & Ware, 1982). The power of the GLMMs lies with its ability to handle non-normal categorical data. In the special case of binary outcomes (dichotomous data), the GLMM logit link function is formulated as

$$f_{link}[\mu] = \text{logit}(\mu) = \log \left[\frac{\mu}{1 - \mu} \right] = \eta. \quad (13)$$

The conditional expectation equals the conditional probability of receiving a positive score given the random effects:

$$E[y|\gamma] = \mu = P(y = 1|\gamma). \quad (14)$$

The model can be formulated as

$$P(Y_{ji} = 1|\gamma_j, x_{ji}, z_{ji}) = f_{link}^{-1}(\eta_{ji}) = \Psi(\eta_{ji}) \quad (15)$$

where j represents the higher-level unit (cluster, subjects) and i as the level-one unit (repeated observations, items) nested within j . The inverse link

$$\Psi(\eta_{ji}) = [1 + \exp(-\eta_{ji})]^{-1} \quad (16)$$

happens to be the logistic cumulative distribution, which simplifies parameter estimation by relating to the probability density function in a simple way:

$$\psi(\eta_{ji}) = \Psi(\eta_{ji})[1 - \Psi(\eta_{ji})] \quad (17)$$

The alternative to this logistic model is the probit model, which is based on standard normal distribution and uses the normal cumulative distribution and probability density function. In conclusion, the differences between the four closely connected classes of models can be summarized in the following way. The ordinary linear regression model contains no random effects and assumes normal distribution of the error terms. The generalized linear models utilize a link function to relate the linear model to the outcome variable, which allows the error distribution to be other than normal. The homoscedasticity assumption of the linear regression extends into specifying that the variance of the dependent variable is a function of its predicted value (the mean). Furthermore, the linear mixed models assume that the function relating μ to the fixed and random effects can be linear, that the variance is not a function of the mean, and that the random effects follow a normal distribution. All these assumptions become untenable with

non-normal dependent variables (e.g., binary outcomes) so that linear models cannot be directly applied.

Nonlinear Mixed Models (NLMMs)

Some IRT models are nonlinear because of their multiplicative functions in their specification (e.g., a product of a slope parameter and a threshold). Although some authors consider that GLMMs include NLMMs (Lindstrom & Bates, 1990), the class of generalized linear mixed models is said to be a subset of nonlinear mixed models (McCulloch & Searle, 2001; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). Often the two terms combine to refer to a broad family of models that incorporate such characteristics as fixed and random effects on the outcome, independent observations from exponential distributions, and linear predictors through a link function. Nonlinearity occurs when the fixed or the random effects or both are modeled in a nonlinear fashion; or in the case of the proposed DIF modeling approach, the nonlinearity in the difficulty of composite items resides in the product of two parameters in estimation: the latent item predictor and component weight, which makes up the fixed effects part of the MIRID.

Item Response Modeling in the GLMMs Framework

Regular item response theory (IRT) models can be conceptualized within the GLMMs framework, including binary data models such as the Rasch model and componential models like the Logistic Linear Test Model (LLTM). Since the MIRID was developed on the basis of the Rasch model and the LLTM, the section below will describe formulation of the Rasch in the GLMM framework after the rationale for doing so is given. Because the standard MIRID does not involve the item discrimination parameter, there is no nonlinear term in the specification.

The Rationale of a Generalized Statistical Approach to Item Response Modeling

The purpose of conventional item response models is to measure certain affective or cognitive outcomes in relation to individuals in order to evaluate, compare, or predict their “performance” on the measured variable. This modeling approach gives individual estimates of a person parameter, which, dependent on the outcome variable, can be person’s ability, proficiency, psychological traits, attitudes, etc. At the same time, each of the items on the assessment instrument (e.g., a test, a survey, etc.) administered also receives estimate of its parameter, which is often labeled as “location” or “difficulty”. Estimated person parameters and item parameters imply that the persons and the items have been placed on the same scale of the construct being measured. This modeling approach “describes” the locations occupied by individual items and persons alike.

In other academic disciplines, conventional statistical methods are often used to test hypotheses in connection with design effects, for example, in sciences and medical research, and attempt to answer the question of “why.” Such studies are explanatory, whose principal mission is to explain the outcome variable in association with the design factors under investigation. The broad framework of GLMMs are of explanatory nature and item response models defined within this framework are ‘explanatory,’ too (De Boeck & Wilson, 2004). Since there are multiple items on an instrument, item responses are inherently repeated observations and conform to a structure where items are nested within persons. This new angle of looking at item response models forms the basis for the explanatory approach, which relates IRT to the broad statistical literature on mixed models.

This approach brings into the model item and person characteristics to complement the location parameters. That is, characteristics such as the cognitive operations an item taps into,

item content, students' SES, anxiety level, etc., can be added to the model as regression predictors (covariates). The GLMMs framework satisfies the measurement goal by providing an estimate of the location parameter on the measurement scale for each person and each item based on the probability of a correct response. In addition, the estimates of the regression coefficients give us the understanding of the correlation between item responses and the predictors. In other words, the regression function explains the extent to which item and person properties affected item responses. Depending on research interests and questions, different covariates can be incorporated to adapt or extend standard item response models to serve a specific scientific query or special data set. Therefore, this generalized approach achieves both the descriptive and explanatory purposes of modeling.

The estimated regression weights in the generalized IRT models are in fact the effects of explanatory variables on how individuals responded to items. The item and person location parameter estimates in this model are obtained in a different way from a descriptive model, although both sets are fixed point estimates on the measurement scale. Conventional models treat items and persons as unchanging entities with only one location parameter each. The GLMMs approach combines the effects from all included predictors, which vary across items and persons, to estimate the location parameters, often resulting in greater accuracy and better model fit. Conceptualized within this statistical framework, traditional and newly created item response models can be fitted with computer programs designed for GLMMs and NLMMs. Details of such estimation and software can be found in later sections.

Recasting the Rasch Model within the GLMMs

Item response theory models as types of latent trait models were developed outside the GLMMs in the fields of educational and psychological measurement. Statisticians have sought to

merge the two classes of models. For example, Mellenbergh (1994) developed generalized linear item response theory (GLIRT) that is analogous to the generalized linear models. Moustaki and Knott (2000) proposed generalized latent trait models to analyze manifest variables with different distributions. Rijmen et al. (2003) introduced a nonlinear IRT framework based on the mixed logistic model. The explanatory item response theory models by De Boeck and Wilson (2004) clarified the differences between various item response models and statistical models and placed them in a broad statistical framework that enables a generalized statistical approach to data analysis which takes advantage of the flexibility of available statistical computing packages.

In binary data analysis with link function being either logistic or probit, and the random effects assumed to be normal, the close relationship between the basic Rasch model and the GLMMs is the most evident. Under the Rasch model, the responses to items ($i = 1, 2, \dots, I$) by subjects (persons) ($j = 1, 2, \dots, J$) are assumed to be conditionally independent Bernoulli observations, where the conditional probabilities of getting a score of 1 are modeled as follows:

$$p(Y_{ji} = 1 | \theta_j, \beta_i) = \pi_{ji} = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)}. \quad (18)$$

where π_{ji} is the probability of success on item i by person j ; β_i is the item parameter of item i ; θ_j is the person parameter (ability) of person j . The person parameter is a latent variable that is treated as fixed in the Rasch conception. To enter this model into the realm of the GLMMs, we need to 1) consider the θ_j values as randomly sampled from a normally distributed population and 2) regard item responses as nested within persons.

Equation 9 gave the GLMMs linear equation in matrix terms. In summation format, this equation can be re-written for subject j and item i as follows:

$$\eta_{ji} = \sum_{k=0}^K \beta_k X_{ik} + \sum_{p=0}^P \gamma_{jp} Z_{ip}, \quad (19)$$

where k represents the fixed-effect predictors (items) and p the random-effect predictors (persons). To comply with the tradition of psychometrics, γ as the personal parameter is replaced with θ and the item parameter β takes on the negative form. Since no person predictor is included in the Rasch model, the random part of the equation reduces to θ_j as the intercept. For the fixed-effect part, $X_{ik} = 1$ only when $i = k$, so only one term of this sum is kept. After these changes, the linear predictor for the recast Rasch model is

$$\eta_{ji} = \theta_j - \beta_i = \text{Ln} \frac{\pi_{ji}}{1 - \pi_{ji}}, \quad (20)$$

which is the expected value on the logit scale with $\theta_j \sim N(0, \sigma_\theta^2)$. This Rasch model can also be considered a regression model as follows:

$$\eta_{ji} = \theta_j - \beta_1 X_{i1} - \dots - \beta_k X_{ik} - \dots - \beta_l X_{il} \quad (21)$$

Since the mean of θ_j is specified as zero, the random effects are defined as the deviations from the mean effect. The mean of β is also constrained to be zero to ensure that the model is identifiable (otherwise, X would not be of full column rank).

The GLMMs can be extended to handle response data with more than two categories (1/0). However, since polytomous data are out of the scope of this study, extensions in this regard are not reviewed here but their details can be found in such studies as Tuerlinckx and Wang (2004), Fox (2007), and Natesan, Limbers, and Varni (2010). Like the standard Rasch model, the polytomous models introduced by these authors can be seen as members of the multivariate generalized mixed models. Because GLMMs are by nature hierarchical models

suitable to analyze data of nested structure when items are considered nested within persons, they are also labeled hierarchical models or multilevel models, which are a class of the GLMMs.

Within the framework of the GLMMs, explanatory item response modeling provides additional utility to the data description brought by conventional IRT modeling. Not only does it serve the measurement purpose, it also provides insight as to why the level of measurement is achieved in terms of the item or person properties being investigated. The added benefits of the GLMMs framework call for more attention to this modeling approach.

The Model with Internal Restrictions on Item Difficulty (MIRID)

This section reviews the conception and formulation of the standard binary MIRID (a Rasch model) from the perspective of generalized linear mixed models.

MIRID in the Generalized Statistical Models Framework

The GLMM framework created space for development of nonstandard, “specialized” item response models, one of which is the model with internal restrictions on item difficulty (MIRID). By incorporating latent item characteristics, the MIRID can be applied to instruments consisting of item families created with component and composite items. In essence, the MIRID is designed to explain item responses by modeling the assumed latent linear relationship between different types of component items and composite items within each situation. Since its official publication in 1998 as a Rasch type of item response model, various extensions have been proposed that have turned the MIRID far more generalizable although these extensions are not considered here. For example, Wang and Jin (2010a) formulated two types of polytomous MIRID for ordinal response data, one for the cumulative logits and the other for adjacent-category logits. In addition, the authors proposed (2010b) a multilevel, two-parameter MIRID

with random weights. Also recently, Lee (2010) suggested that two generalizations be added to the model: the random item effects and the multidimensionality.

The primary utility of the MIRID is investigate affective and cognitive outcomes using item specific componential difficulties and component weights that are more realistic to model than the components themselves being latent. By designing component items to represent “subtasks” as predictors of the corresponding composite item, the MIRID can be used to test theories on how complex psychological constructs can be broken down and influenced by their parts.

The power of item response modeling in the GLMMs/NLMMs framework lies in its ability to allow covariates to enter the model at either subject or item level as independent variables to explain their effects on item responses. Outside this framework, such an analysis is typically conducted in two phases: first, item and person parameter estimation under the regular item response theory structure and second, a regression analysis to bring the research variables into the model to explain and predict their effects on the latent outcome variable.

The generalized linear mixed models framework for item response data reviewed here is mainly based on Rijmen, Tuerlinckx, De Boeck, and Kuppens, (2003) and De Boeck and Wilson (2004). In this framework, the basic Rasch model is regarded as a regression model where the logit of a correct response (η_{ji}) functions as the expected value, the person parameter (θ_j) as the intercept in the regression, and item parameter ($-\beta_j$) as the regression weight of X_{ik} (see Equation 20). The typical predictors in the Rasch model are person parameter and item predictors, one for each item. When $k = i$, $X_{ik} = 1$; otherwise $X_{ik} = 0$. The full Rasch model in regression format taking into account all items is spelled out in Equation 21. The values of the item parameters ($-\beta_j$) do not vary across persons.

Recast in this mode, item and person predictors are used to explain the effects of items and persons and therefore the basic Rasch model becomes a case of an explanatory item response model (De Boeck & Wilson, 2004). In addition to item and person predictors, item and person properties can be incorporated in the regression model. For person properties, the predictors can be both manifest variables (e.g. gender, SES, etc.) and latent variables that are regressed on external personal variables (Adams, Wilson, & Wu, 1997) such as motivation, attitude towards school, etc. Item properties can be the cognitive processes an item is written to tap into. When covariates reflecting both item and person properties are introduced into the model, it becomes “doubly explanatory” (De Boeck & Wilson, 2004).

The MIRID belongs with the category containing only item property predictors, along with the linear logistic test model (LLTM; Fischer, 1973). The relationships between the two models will be described later following the introduction of MIRID.

The Rasch MIRID

The MIRID model was proposed originally by De Boeck (1991) to explore the componential structure of an affective or cognitive construct measured using a test or questionnaire. Later, Butter (1994) and Butter, De Boeck and Verhelst (1998) developed it into full formulation. As their version of MIRID was devised to fit binary response data based on the basic Rasch model, it is labeled as the dichotomous Rasch MIRID. By design, the MIRID models are not suitable for regular assessments but only for a particular type of data which consist of component items and composite items. The multiple mental processes in a cognitive or affective construct can each be considered a subtask or a single operation when measured. At the lowest cognitive level, hypothetically, one can imagine such a construct as a hand calculation problem involving three subtasks, addition, subtraction and multiplication. The item

encompassing all three subtasks is a composite item, whereas the other three items each measuring one subtasks are component items. Together the four items form an item family.

Table 1 below illustrates this structure. Hypothetically, each item family could represent a hand calculation problem fully spelled out in composite items four, eight, and twelve. Component one to three correspond to the three subtasks, addition, subtraction, and multiplication, represented by the three component items from each family (items one to three, five to seven, and nine to eleven). On the affective side, a hypothetical example could be evaluating a complex latent trait, such as “grit”, which comprises components like perseverance, concentration, and motivation. Item families could be designed to measure this trait from disparate real-life contexts such as work, study, exercise, etc., often labeled as “situations” as the items can be written for specific environments.

The MIRID assumes that the difficulty of the composite item can be decomposed as a weighted sum of the difficulties of the component items. This linear relationship creates internal restrictions on the difficulty of the composite item, hence the name MIRID. The purpose of the MIRID is to investigate the underlying relationship between the processes behind a complex psychological construct and examine the internal validity of the component and composite items appearing on the same assessment.

Formulation of the Rasch MIRID

Within the generalized mixed models framework, the MIRID formulation contains both fixed and random effects. One piece of the fixed effects is reflected by an item predictor matrix A as shown in Table 2-2, where k represents one of a total of K components, m as one of a total of M item families, and β_{mk} is the difficulty for component k in item family m . This matrix

summarizes the K component item parameters across M item families as well as a vector of constant.

The other piece of fixed effects is shown in Table 3 as the component weight matrix Ψ reflecting component item weights for every item family. In this table, the identity matrix reflects the component items under every component; ω_k is the weight of component k ; ω_0 as the intercept is a normalization constant.

Table 1.

Item Families and Component Items

		Component 1			Component 2			Component 3		
Family1	Item1	1	0	0	0	0	0	0	0	0
	Item2	0	0	0	1	0	0	0	0	0
	Item3	0	0	0	0	0	0	1	0	0
	Item4	1	0	0	1	0	0	1	0	0
Family2	Item5	0	1	0	0	0	0	0	0	0
	Item6	0	0	0	0	1	0	0	0	0
	Item7	0	0	0	0	0	0	0	1	0
	Item8	0	1	0	0	1	0	0	1	0
Family3	Item9	0	0	1	0	0	0	0	0	0
	Item10	0	0	0	0	0	1	0	0	0
	Item11	0	0	0	0	0	0	0	0	1
	Item12	0	0	1	0	0	1	0	0	1

Table 2.

Item Predictor Matrix

	Predictor 1	Predictor 2	...	Predictor $K-1$	Predictor K	Constant
Family 1	β_{11}	β_{12}	...	$\beta_{1(K-1)}$	β_{1K}	1
Family 2	β_{21}	β_{22}	...	$\beta_{2(K-1)}$	β_{2K}	1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
Family $M-1$	$\beta_{(M-1)1}$	$\beta_{(M-1)2}$...	$\beta_{(M-1)(K-1)}$	$\beta_{(M-1)K}$	1
Family M	β_{M1}	β_{M2}	...	$\beta_{M(K-1)1}$	β_{MK}	1

Table 3.

Component weight Matrix for One Item Family

	Component 1	Component 2	...	Component K	Composite
Item 1	1		...		ω_1
Item 2	0	1	...		ω_2
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
Item R	0		...	1	ω_K
Intercept	0	0	...	0	ω_0

The product of the two pieces, component weight matrix and item predictor matrix, becomes the fixed effects of the model, as shown in Equation 22, which is exemplified in Equation 23 with a two-family three-component structure. The right-hand side of this equation shows the item parameter matrix for the six component items and two composite items.

$$(Fixed\ Effect)_{ji} = A_f \Psi_r = \beta'_i, \quad (22)$$

where $\beta'_i = \beta_{mk}$ for component items and $\beta'_i = \sum_{k=1}^K \omega_k \beta_{mk} + \omega_0$ for composite items with $i = 1, 2, \dots, K + 1, \dots, M(K + 1)$ as defined in 2-23 with three components.

$$\begin{aligned} & \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} & 1 \\ \beta_{21} & \beta_{22} & \beta_{23} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & \omega_1 \\ 0 & 1 & 0 & \omega_2 \\ 0 & 0 & 1 & \omega_3 \\ 0 & 0 & 0 & \omega_0 \end{pmatrix} \\ & = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \sum_{k=1}^3 \omega_k \beta_{1k} + \omega_0 \\ \beta_{21} & \beta_{22} & \beta_{23} & \sum_{k=1}^3 \omega_k \beta_{2k} + \omega_0 \end{pmatrix}. \end{aligned} \quad (23)$$

Definition of the fixed effects imply that the values of the latent item predictors are also the item difficulties of the component items. For composite items, their fixed effects are explained in terms of latent item predictors and their weights (Smits & Moore, 2004). Note that in generalized terms, the difficulty of the composite item is assumed to be

$$\beta_{m0} = \sum_{k=1}^K \omega_k \beta_{mk} + \omega_0, \quad (24)$$

where m is one of the M item families.

The random effects in the MIRID mirror those in the Rasch model in that the person parameter is allowed to vary randomly as expressed by $\theta_j \sim N(0, \sigma_\theta^2)$. In regression format, the dichotomous MIRID model is defined as

$$\log \left[\frac{P(x_{jmk} = 1 | \theta_j)}{P(x_{jmk} = 0 | \theta_j)} \right] = \eta_{ji} = \theta_j - \beta_i'. \quad (25)$$

An Example

The MIRID model has been applied in both the cognitive and affective domains. To illustrate the circumstances where it can be applied and how it can be applied, the methods and empirical data from a previous study (Smits & De Boeck, 2003) are described briefly in the following paragraphs. The standard MIRID, as well as its extensions like polytomous MIRID, has been applied to this data set.

In this study about measuring a construct, guilt, the cited theory suggests that the feeling of guilt in a given situation can be decomposed mainly into three components: 1) feeling of a norm being violated; 2) a tendency to worry about what one has and has not done; 3) a desire to make restitution for one's misdeeds (Barrett, 1995; Gilbert, Pehl, & Allan, 1994; Tangney, 1995). On the basis of this theory, the researchers interviewed a group of teenagers and asked them to describe a situation where they felt guilty in one of the three contexts: work or study situation, personal relationships, and leisure time.

From the interviews ten scenarios were collected and summarized as environments conducive to the feeling of guilt. The first two of the ten have to do with breakup and trumpet and are given below as examples:

1. During some time you were having a love affair, but you're not really in love with him or her. Making an end at this relation, you discover he or she supposed the relation was serious. He or she is very grieved.
2. A few years ago, you started playing trumpet in a brass band. The schooling you needed is completely paid by the brass band. At the moment, you're a good musician, you stop playing in the brass band, because you find yourself not fitting in the group of musicians.

Survey questions written for these scenarios were administered to 270 high school students, who were asked to answer the following four questions on a four-point rating scale (0 = "No"; 1 = "Not Likely"; 2 = "Likely"; and 3 = "Yes"):

1. Do you feel like having violated a moral, an ethic, a religious and/or a personal code? (the norm-violation component)
2. Do you worry about what you did or failed to do? (the worrying component item)
3. Do you want to do something to rectify what you did or failed to do? (the restitution component item)
4. Do you feel guilty about what you did or failed to do? (the guilt composite item)

These three component items (1st to 3rd) and the corresponding composite item (4th) form an item family for every one of the ten scenarios/situations. The table below is reproduced from the study to illustrate the output of parameter estimates. Judging by the estimated values, for

example, it is clear that the componential contributions from situation 10 were lower with all three components than the contributions from situation 5. That is, the teenagers in the sample were less likely to feel guilty in situation 10 than 5. Using hand calculation, the reconstructed composite item parameter in situation 5 amounts to $(.245 * 563) + (.591 * .775) + (.300 * 1.094) - .082 = .842$. The authors noted that since the sum of the component weights is only slightly larger than 1.00 (1.136) and the intercept is nearly zero (-.082), the linear function got nearer to a weighted average.

Table 4.

Parameter Estimates from the Example Study

Situation	Component 1		Component 2		Component 3		Intercept	Composite
	Parameter	Weight	Parameter	Weight	Parameter	Weight		
1	-.245	.245	.507	.591	.089	.300	-.082	.184
2	-1.536		-1.971		-2.062			-2.242
3	-.745		-.060		-.321			-.396
4	-.122		.009		.053			-.091
5	.563		.775		1.094			.842
6	1.272		1.006		.830			1.073
7	.352		.604		.170			.412
8	-.193		-.279		-.108			-.327
9	-.077		.814		1.461			.819
10	-.623		-.718		-.541			-.821

Note: Standard errors of estimation are omitted.

MIRID and LLTM

In the GLMMs, both the LLTM and the MIRID belong with the same class that employs only item properties as predictors. In statistical terms, either model was proved to be generalization of the other; under certain conditions, the two models can be equivalent (Butter et al., 1998; Bechger et al., 2002; Maris & Bechger, 2004). It is therefore important to describe briefly the close relationships between the two models.

Researchers were interested to investigate the factors behind the level of test performance and item difficulty as well as the relationships between these factors. A psychometric model that serves this purpose was the linear logistic test model (LLTM; Fischer, 1973), which was built up to explain item difficulty parameters with respect to the underlying cognitive processes the items were posited to measure. Items' association with componential processes is considered an item

property which can be incorporated as a predictor in the generalized statistical modeling framework.

The standard LLTM is also a dichotomous Rasch model whose item parameters are modeled as linear contributions in a way similar to how the difficulty of composite item in the MIRID is derived. However, the so-called complexity factor values (q_{ik}) replaces the component item parameters in Equation 24.

$$\beta_i = \sum_{k=1}^K \omega_k q_{ik} + \omega_0 . \quad (26)$$

The values of the complexity factor are given a priori as specified by the Q-matrix. For item i , if component (or process) $k = i$, $q_{ik} \neq 0$, otherwise $q_{ik} = 0$.

The starkest distinction between the two models is the fact that there is no composite item in LLTM as every item parameter reflects certain item properties, the extent of which is assumed known and given by the entries in the Q-matrix. On the opposite, the MIRID does not require any prior knowledge about the size of the componential contributions q_{ik} but estimate the parameters of the component items (β_{mk}). That is, the MIRID estimates the entries in the Q-matrix which are provided in the LLTM. The weights of the contribution (ω_k) are parameters to be estimated in both models.

Estimation Methods and Computer Programs for the MIRID

Likelihood-based methods are commonly used in statistical estimation. In this category there are three popular procedures, joint maximum likelihood estimation (JMLE), conditional maximum likelihood estimation (CMLE), and marginal maximum likelihood estimation (MMLE). Their popularity is in part due to the well-understood properties of maximum likelihood estimators: asymptotic consistency and normality, as well as estimation efficiency.

The basic MIRID was devised with CMLE for parameter estimation but the authors of its various extensions proposed and applied their modeling approaches using MMLE procedure. With CMLE, the person parameters are eliminated by being conditioned on a sufficient statistic for the latent trait θ . For MMLE, the marginal likelihood is maximized by first integrating out the person parameters and using the first- and second-derivatives to derive item parameter estimates. The estimates of person parameters can then be obtained using the estimated item parameters. Other authors proposed estimation approaches with resampling-based Bayesian methods, particularly the Markov Chain Monte Carlo (MCMC) procedure, for more complex MIRID models (such as multilevel ones). This section will focus on MMLE only as it is the approach used in this research and discuss its basic concepts and main usage. After, the software to implement this method will be introduced.

Marginal Maximum Likelihood Estimation

In the standard Rasch model, item difficulties (β s) are treated as fixed effects and person parameters θ s are regarded as random effects, as defined below:

$$\eta_{ji} = \text{logit}(\pi_{ji}) = \theta_j - \beta_i, \quad (27)$$

which portrays the relationship between J persons and I items. In the Rasch model, the concept of sufficient statistics (Andersen, 1980) is defined by the total score of an individual, that is,

$$s_j = \sum_{i=1}^I y_{ji}.$$

In addition to item effects, person effects can be brought into the likelihood as independent random draws from a density defined over a population denoted by $g(\theta_j|\psi)$. In psychometrics, this population density is typically assumed to be normally distributed with a mean of zero and unknown variance with the population parameters ψ estimated along with item parameters. The marginal maximum likelihood (MML) can be defined as

$$L(\beta, \psi) = \prod_{j=1}^J \int \prod_{i=1}^I \Pr(Y_{ji} = y_{ji} | \theta_j) g(\theta_j | \psi) d\theta_j, \quad (28)$$

which is to be integrated with respect to the random effects. Parameter estimates are derived through maximizing this likelihood. As the normal distribution of the random effects can be denoted as $\phi(\theta_j | \mu_\theta, \sigma_\theta^2)$, where μ_θ is often set as zero and σ_θ^2 represents the variance to be estimated, Equation 4 can be redefined as follows

$$L(\beta, \sigma_\theta^2) = \prod_{j=1}^J L_j(\beta, \sigma_\theta^2) = \prod_{j=1}^J \int \Pr(y_j | \beta, \theta_j) \phi(\theta_j | 0, \sigma_\theta^2) d\theta_j \quad (29)$$

where $L_j(\beta, \sigma_\theta^2)$ represents individual contribution to the marginal likelihood. Since the integral in this formulation does not have a closed-form solution (Tuerlinckx et al, 2004), numerical integration techniques are employed to approximate the integral to get at maximizing the likelihood.

Different maximization methods are available for MMLE, such as the Gauss-Hermite quadrature method (Abramowitz & Stegun, 1974), which is popular in item response modeling research. In a nutshell, the Gauss-Hermite quadrature method approximates the integral by replacing it with a single finite number of rectangles as close in total size as the area under the integrand. The Gaussian quadrature approximation is defined as follows (Naylor & Smith, 1982):

$$L_j(\beta, \sigma_\theta^2) = \int \Pr(y_j | \beta, \theta_j) \phi(\theta_j | 0, \sigma_\theta^2) d\theta_j \approx \sum_{m=1}^M \Pr(y_j | \beta, \sqrt{2}\sigma_\theta q_m) \frac{w_m}{\sqrt{\pi}}, \quad (30)$$

where q_m and w_m are the m th quadrature node and weight, respectively. In Gaussian quadrature approximation, the quadrature points are centered at zero for each random effect so that the current random effects variance matrix is used as the scale matrix. That is, every person has the same rescaled nodes, which may be unrealistic for individuals located at either end of the normal

distribution. The adaptive Gaussian method (Pinheiro & Bates, 1995) corrects this shortcoming by using the empirical Bayesian estimate of θ_j calculated along with its asymptotic variance ($\hat{\tau}_j^2$) for each person. This Bayesian estimate $\hat{\theta}_j$ needs to be added to the node q_m , which must be multiplied by $\sqrt{2}\hat{\tau}_j$. The marginal likelihood for person j can be defined with adaptive Gaussian quadrature as follows:

$$\begin{aligned}
 L_j(\beta, \sigma_\theta^2) &= \int \Pr(y_j|\beta, \theta_j)\phi(\theta_j|0, \sigma_\theta^2)d\theta_j & (31) \\
 &= \int \frac{\Pr(y_j|\beta, \theta_j)\phi(\theta_j|0, \sigma_\theta^2)}{\phi(\theta_j|\hat{\theta}_j, \hat{\tau}_j^2)}\phi(\theta_j|\hat{\theta}_j, \hat{\tau}_j^2)d\theta_j,
 \end{aligned}$$

Although adaptive Gaussian method requires fewer quadrature nodes than the non-adaptive Gaussian method, it takes more computing resource since empirical Bayesian estimates must be calculated at each step of the optimization process. However, it was shown that the two methods yielded similar results (De Boeck & Wilson, 2004).

The approximated likelihood function then goes through optimization through some iterative numerical methods, such as the Newton-type algorithms, of which the Newton-Raphson technique and Fisher scoring algorithms are the most widely used. Both of these techniques are direct approaches based on the first and second derivatives. Because they are expensive to compute and unreliable in convergence, a few quasi-Newton algorithms were proposed that require only the gradient (first derivative) and thus are more efficient. A representative indirect maximization approach is the Expectation-Maximization (EM) algorithm.

The more complex MIRID models such as those of multilevel and of crossed random effects adopted the Bayesian Markov Chain Monte Carlo (MCMC) estimation as a valid

alternative (Lee, 2010; Hung, 2011). However, this method has its own points of complexity, like selection of priors and possibly lengthy burn-in time, convergence evaluation, etc.

Overall, the differences between maximum likelihood approaches and Bayesian MCMC are obvious (Tuerlinckx et al., 2004). First, the distinctions between fixed and random effects are clear with CML and MML; with MCMC, all effects are in essence random. In addition, maximum likelihood estimation gives standard errors based on an asymptotic normal approximation to the likelihood but only parameter intervals of posterior distributions are yielded by MCMC. Thirdly, convergence evaluation is less straightforward with MCMC than with CML and MML approaches.

Software

The software that implements the conditional maximum likelihood (CML) approach was the MIRID CML program (Smits et al., 2001), which was created for the original Rasch-MIRID and dichotomous data only. The SAS NLMIXED procedure can also apply the CML approach and was found to produce different person parameter estimates than the MIRID CML program and essentially the same item parameter estimates (Smits & De Boeck, 2003). Since then, MMLE as implemented by SAS NLMIXED has been the engine behind much of the psychometric modeling research, including many studies to extend and generalize the MIRID. This is logical since the MIRID and its extensions can be viewed as special cases of generalized nonlinear mixed models. Also, the well-understood properties of MMLE and the ease in implementation and output interpretation with SAS NLMIXED remain attractive.

For the extended models, such as the random-weights MIRID, polytomous MIRID, OPLM MIRID, random-weights polytomous MIRID, the two-parameter MIRID, the NLMIXED procedure was the tool for parameter estimation built upon marginal maximum likelihood (MML)

approach. In terms of integration methods, both Gauss-Hermite quadrature and adaptive Gaussian approaches were chosen in different studies. The number of quadrature points in combination with the Gaussian approximation equals the number of points used in each dimension of the random effects, one of which is the intercept. The number used during the evaluation of the integral were specified to be 15 or 20 in different studies to obtain a reasonable precision in describing the distribution of the random effects without increasing the estimation time significantly.

As for the optimization techniques to carry out the maximization, NLMIXED provides a number of options. The default is a dual quasi Newton-Raphson algorithm. In contrast to the original Newton-Raphson technique that calculates standard errors of the parameter estimates from the second derivative matrix of the likelihood function, the quasi-Newton approach computes only the first derivatives and thus takes much less time to run.

It has been shown that complex, multilevel, and nonlinear models can be difficult or impossible to estimate using existing MLE-based software (Congdon, 2003; Fox, 2010). However, because the DIF models proposed in this dissertation are not overly complex, MMLE using NLMIXED is chosen for parameter estimation in this research and the MCMC approach is not necessary.

Differential Item Functioning

When examinees at the same ability level from the reference and focal group have different probabilities of answering it correctly, the item is considered to have DIF (Pine, 1977). In measurement practice, the construct of interest can be viewed as comprising more than one dimension. This does not imply necessarily applying multidimensional psychometric models; rather, it provides a framework for study of differential item functioning. DIF at item-group level

can be viewed as a consequence of secondary dimensions to the primary or the target trait being tested but unaccounted for in normal assessments. In keeping with the multidimensionality paradigm, Shealy and Stout (1993) suggested that DIF can be studied for a group of items with the method of “differential bundle functioning”. When groups of items sharing the same properties presumably contribute to the differential effect in a substantive way, this form of DIF is labeled as “differential facet functioning” (Englehard, 1972), which summarizes individual-item DIF in a parsimonious fashion on the basis of commonality amongst these items.

DIF studies at item-group level have been conducted, such as Kim and Huynh (2010) who discovered that some items administered in paper-based mode favored students without disabilities. Similarly, Wainer, Sireci, and Thissen (1991) demonstrated how to model DIF at the testlet score level (“differential testlet functioning”). Other studies adopted the DIF framework of secondary dimensions (Ackerman, 1992; Bolt & Stout, 1996; Douglas, Roussos, & Stout, 1996; Finch, 2005; Roussos & Stout, 1996; Shealy & Stout, 1993a, 1993b; Xie & Wilson, 2008). According to Shealy and Stout (1993a & 1993b), a secondary dimension is considered auxiliary and to cause benign DIF if it complements the primary dimension intended to be measured; on the contrary, if the item property is irrelevant to the construct, it is a nuisance dimension that leads to adverse DIF. Substantive analysis may be called upon to determine whether the DIF is benign or adverse. By retaining the auxiliary dimension of items and eliminating items with adverse DIF, construct validity and fairness of the test will be improved at once. Although the MIRID was conceived as a unidimensional model, it can be considered to some extent multidimensional if each component or item family (“situation”) is treated as a dimension of the trait of interest. Thus the multidimensional DIF framework proposed by Shealy and Stout (1993) can be applied in this research to untangle the complexity.

Numerous DIF detection methods have been developed and implemented, among which the most representative and widely used are the Mantel-Haenszel method and SIBTEST, whose brief descriptions are given below.

The Mantel-Haenszel Procedure

Holland (1985) and Holland and Thayer (1988) first described and applied the MH procedure in DIF studies. With MH, detection is conducted via comparing the odds ratios of item endorsement frequencies across reference and focal groups after examinees have been matched on a measure of the latent trait, which is typically the total score. Next, both groups are divided into a number of strata on the basis of the test score levels. Within each stratum, a 2 (groups) x 2 (item score) contingency table is constructed for the studied item with group membership as a function of item response frequency. The odds of endorsing an item in each stratum are obtained and aggregated across the strata to compute the MH statistic, which is distributed as a chi-square with 1 *df*. If the observed MH exceeds the critical value of 3.84, the item is flagged as exhibiting DIF, and the process is repeated for all the remaining items.

Simultaneous Item Bias Test (SIBTEST)

SIBTEST was developed on the basis of the multidimensionality DIF framework (Shealy & Stout, 1993a, 1993b). This procedure is able to detect both uniform and nonuniform DIF in multiple items at once. In practice, the test is divided into two, one "suspect" subtest containing the item(s) suspected of DIF and the other assumed to be DIF-free, the score on the DIF-free subtest serving as the matching variable. A weighted mean difference in subtest score between the two groups as well as its standard error are calculated and their ratio becomes the DIF statistic to be tested against the null hypothesis of it being zero.

The Model-based DIF Approach for the MIRID

By applying the paradigm of multidimensionality to potentially explain differential functioning, the MIRID DIF can be studied in the statistical framework of generalized linear and nonlinear mixed models (GLMMs and NLMMs) with the inclusion of grouping and interaction covariates. The nonlinearity in the difficulty of composite items results from the product of two parameters to be estimated: the latent item predictor and component weight, which makes up the fixed effects part of the MIRID. Such a model-based approach is primarily based on Meulders and Xie (2004), who modeled DIF by including person-by-item interactions as predictors in the NLMM. Their work extended from a general DIF approach, differential facet functioning (Englehard, 1992), which allows various procedures to explore DIF at the level of item groups.

The proposed DIF approach also falls within the framework of explanatory measurement (De Boeck & Wilson, 2004), which involves person properties and/or item properties to explain the effects of persons and/or items and is grounded in the GLMMs and NLMMs. In this framework, traditional item response models like the Rasch model are viewed as “descriptive” due to the lack of covariates representing item or person properties; however, a model like the LLTM is considered “item explanatory” because of the inclusion of item attributes but no person properties. In case of DIF modeling, the person property means group membership and the item property is embodied in the cognitive operation(s) it measures so that their interaction reflects the difference in ability between the focal and the reference group.

Individual-item Level Model Specification (DIF)

The proposed DIF models extend the standard MIRID by incorporating covariates to explain various potential sources of DIF. Since the standard MIRID is of the Rasch family, the proposed DIF models here contain no discrimination parameters and focus on only uniform DIF.

The most fundamental type is sporadic differential functioning exhibited by individual items which can be modeled by extending the specification of a component item in Equation 25. Let j denote one of J persons (students, examinees) in the MIRID with M item families and K components. There are $(K + 1)M$ items in total, with mk denoting an item in family m under component k and $m0$ representing a composite item in family m . Let x_{jmk} be the binary response to the component item which takes value 1 for success and 0 for failure by person j . Assuming x_{jmk} as an independent random variable with Bernoulli distribution and the probability of correct response as $p_{jmk} = P(x_{jmk} = 1)$, the item-level MIRID DIF model can be expressed as follows:

$$\text{logit}(p_{jmk}) = \eta_{jmk} = \theta_j + G_j\gamma_g - (\beta_{mk} + G_j\delta_{mk}), \quad (32)$$

where

θ_j is the trait level (ability) of person j ($\theta_j \sim N(0, \tau)$), a random effect.

γ_g accounts for the average group difference in the latent trait to solve the problem of ability matching in traditional DIF studies; that is, there is no longer need for separate estimation of trait level of group members. In this model, γ_g is a fixed effect from the focal group.

G_j denotes group membership of person j , with a value 0 indicating reference group and 1 focal group.

β_{mk} represents the difficulty parameter for component item in family m under component k , a fixed effect.

δ_{mk} is the fixed effect reflecting the magnitude of DIF for component item in family m under component k . If negative, it implies DIF effect favoring the reference group.

$G_j\delta_{mk}$ can be viewed as the interaction term between an item and a person's group membership, which is doubly explanatory.

With this individual-item DIF model, when a constant is added to all J ability parameters and all $K \times M$ item location parameters, η_{jmk} remains unchanged. On the other hand, a constant could be added to γ_g and all of the $K \times M$ DIF effects to arrive at the same η_{jmk} . Such a model is not identified. According to Paek (2002), identifiability in DIF modeling can be obtained by constraining the mean of the distribution of θ_j to be zero or setting one or more of the δ_{mk} as 0. Such a constraint assumes a priori knowledge of at least one item being group invariant, which is not uncommon in empirical situations.

The logit of success for a composite item with DIF can be modeled as

$$\eta_{jmo} = \theta_j + G_j\gamma_g - \left(\sum_{k=1}^K \omega_k(\beta_{mk} + G_j\delta_{mk}) + \omega_0 \right), \quad (33)$$

where DIF effects from all component items within family m are summed up in the parentheses at the right side of the equation.

Item-group Level Models (DFFm and DFFc)

The two domains (facets) of the MIRID, components and situations, can have their group-level DIF modeled simultaneously for a component item. The component differential model (DFFc) and the item family differential model (DFFm) can be specified as such:

$$\eta_{jmk} = \theta_j + G_j\gamma_g - (\beta_{mk} + G_j\delta_k), \quad (34)$$

$$\eta_{jmk} = \theta_j + G_j\gamma_g - (\beta_{mk} + G_j\delta_m), \quad (35)$$

with δ_k and δ_m representing item facet DIF effect for a component and a situation, respectively. Analogous to the item-level DIF model, identification can be achieved here by constraining the mean of the distribution of θ_j to be zero.

Accordingly, the DFFm for a composite item is expressed as:

$$\eta_{jm0} = \theta_j + G_j\gamma_g - \left(\sum_{k=1}^K \omega_k \beta_{mk} + G_j K \delta_m + \omega_0 \right). \quad (36)$$

Similarly, the DFFc for a composite item is defined as

$$\eta_{jm0} = \theta_j + G_j\gamma_g - \left(\sum_{k=1}^K (\omega_k \beta_{mk} + G_j \delta_k) + \omega_0 \right). \quad (37)$$

Note that item facet DIF does not interact with component weights in above specifications.

Component Weight Model (DWF)

It is a likely scenario that the importance of one component weighs more for one group than the other. As a result, this component contributes more to the parameters of the composite items. This differential effect is captured with δ_w and can be seen only in the formulation of the composite DIF model:

$$\eta_{jm0} = \theta_j + G_j\gamma_g - \left(\sum_{k=1}^K (\omega_k + G_j \delta_w) \beta_{mk} + \omega_0 \right). \quad (38)$$

Given the presence of individual item DIF, the component weight model can be applied after the item-group model (component DIF only) to compare which one fits the data better.

It is quite likely more than one type of differential functioning exists within a MIRID data set. For example, one component may be more important to the focal group while a few items in different components contain adverse DIF. However, such a scenario will be left for future research.

In conclusion, the model-based MIRID DIF approach includes multiple models targeted at different potential DIF sources, which can be extended to accommodate the presence of more than one type of DIF. This approach conforms to the structure of the standard MIRID and thus detects only uniform DIF. These models, particularly the DFF models, facilitate explanation of differential functioning effects from a substantive perspective. Modeling DIF effect through interaction terms of group and facets or group and items is an approach under the framework of generalized linear mixed models that can be implemented with general-purpose statistical software. Because a large number of parameters affect the quality of marginal maximum likelihood estimation, it is suggested to start the DIF study by exploring DIF in individual items and conducting a substantive analysis so that parameters related to non-DIF items are not included in the model.

CHAPTER THREE

METHOD

Design of the Research

The Scope

The objective of this research was to propose and examine a model-based approach to detecting and potentially explaining in the context of the MIRID differential functioning by taking into account its possible discrete sources, including individual items (DIF), item facets formed by components (DFFc) and item families (DFFm), and component weight (DWF). The proposed approach is formulated by extending the standard MIRID, a member of the Rasch family of models, to include differential effects in the nonlinear mixed models and was fitted to the data structure of the MIRID. Since the extended MIRID does not include an item discrimination parameter, only uniform DIF was studied.

In the context of the MIRID, the most common type of differential functioning would be the one scattered in a number of individual items that does not exhibit any marked pattern, which can naturally be detected by the DIF model. In addition, the proposed approach incorporates the “item bundling” technique to model differential effect at the item bundle level. Items sharing the same properties form item bundles, also called facets, which, in the case of the MIRID, are item families (“situations”) and components, and the differential functioning by these item facets are labeled “DFFm” and “DFFc”, respectively, in the dissertation. The fourth kind of differential functioning occurs with component weights, meaning the importance of certain component(s)

differs between the focal and reference group, and is called “DWF” henceforth. There are as well other types of possible differential effects within the MIRID, for example, DIF with composite items alone, but they are beyond the research scope here.

In empirical settings, more than one type of differential functioning can occur as a result of the unique data structure of the MIRID. For example, one component may be more important for the focal group than the reference group (DWF) when several individual items exhibit DIF favoring either group (DIF). However, it was decided that as the initial MIRID DIF exploration this research would lay the foundation by tackling each source separately; the investigation of their concurrence is left for future research.

The Simulation Study

A Monte Carlo simulation study was carried out to achieve the objective of the research that was two-fold: assess the efficacy of the proposed models in detecting the four types of differential functioning; and investigate the impact from fitting the “wrong” model to the data generated from a different source of differential functioning. First, simulation conditions were constructed by keeping some variables constant, allowing others to vary, and manipulating a few in order to generate item parameters and response data. The four DIF models were then fitted to all simulated data sets. Based on derived estimates, such measures as bias, RMSE (to check parameter recovery), power, and Type I error rates (for efficacy in DIF detection) were computed and evaluated in order to answer the research questions.

With no prior DIF studies available in the context of the MIRID, direct guidance for research design was not available from the literature. However, the widely acknowledged factors that potentially affect DIF detection in traditional DIF studies were considered in designing the simulation such as sample size, test length, proportion of items on the test containing DIF, the

magnitude of DIF, and the difference in ability distribution between the reference and focal group (Mazor, Clauser, & Hambleton, 1992; Rogers & Swaminathan, 1993). In addition, other factors that must be looked at due to the unique framework of the MIRID were also considered, including number of components, number of item families (situations), and correlation between components. These factors were involved in other MIRID-related research (e.g., Butter, De Boeck, & Verhelst, 1998; Wang & Jin, 2010a, 2010b). Since it would not be feasible to examine the impact of all of the aforementioned factors in one DIF study, in the simulation only a few of them were altered or allowed to vary while others were fixed in all conditions.

Manipulated variables. Four factors were altered to construct simulation conditions: sample size, magnitude of differential functioning, distributional properties of the reference and focal groups, and sources of differential functioning (individual items, components, item families, and component weights). Note that for the three sources at item-group level (DFFc, DFFm, and DWF), the differential magnitude was reflected in data generation through manipulating parameters of individual items as explicated later.

DIF studies with item response models typically involve more parameters to be estimated, and the larger the sample size the better the detection results. For example, in their research on LLTM, Green and Smith (1987) simulated data with sample sizes of 30, 200, and 1000 and concluded that a sample size of at least 1000 would lead to optimal estimation and sample sizes smaller than 200 would result in poor accuracy. In this research, only two levels of sample size were set, 250 or 1,500 for either manifest group, providing a total sample of 500 or 3,000, in order to contrast the effects of small and large sample sizes. Such a choice was inspired by the original MIRID research (Butter et al., 1998) where sample sizes of 300 and 3,000 were part of the simulation design. The issue of unequal groups was not addressed.

Magnitude of DIF is a crucial factor. A common practice in DIF research is to set the magnitude at different levels so that detection power can be tested thoroughly. Accordingly, the amount of individual-item DIF was altered as either 0.2 or 0.7 to represent low and high level of DIF magnitude with the high level expected significant enough to be detected. With no previous research describing the amount of differential functioning in item groups or component weight which can be reasonably anticipated in a practical MIRID setting, extra consideration was given to determining suitable magnitude of DFF and DWF. Through simulation, Nandakumar (1993) showed that items tested for DIF as a group produced more effective detection results with moderate to large differential functioning because of the amplifying effect. This research provided a wonderful opportunity to confirm this finding in a different, unconventional data structure. As there is not a location parameter for an item group in the MIRID, differential functioning supposedly manifested at the item group level had to be represented by modeling it at individual item level, which in the dissertation was achieved by adding a delta effect to every item within the group. That is, DFFc, DFFm, and DWF were simulated by adding the differential amount to the location parameters of the items belonging with the group. For the sake of simplicity, only positive and unidirectional DIF was involved in this research; canceling effect from conflicting differential effects could be studied in the future. In reality, some items could exhibit stronger differential facet functioning than others but here only averaged effects on items were considered.

Research has demonstrated that group differences in latent trait affect DIF detection (Mazor, Clauser, & Hambleton, 1992; Narayanon & Swaminathan, 1996; Shealy & Stout, 1993). Unequal ability distributions make DIF detection more difficult than equal ability distributions. Hence, two conditions in trait distribution were constructed: a matched one where both the

reference and focal groups had the same distribution of $N(0, 1)$, and an unmatched one in which the reference group ability distribution was simulated as $N(0, 1)$, and the focal group was simulated as $N(-.7, 1)$. A difference of about 1 in group means is commonplace in application research (e.g., Donoghue, Holland, & Thayer, 1993).

For this study, the four sources of differential functioning became four levels of the key variable to manipulate. At each level, only one form of differential effect was included in constructing generation conditions and therefore, the differences in item parameters between the reference and the focal groups carried only the effect from one source of DIF.

Variables that remained constant. Other than the aforementioned four variables, several factors that could impact DIF detection efforts were assumed constant in constructing different conditions. The MIRID data structure is made of groups of items (components on one dimension and item families on the other). Although some researchers (e.g., Lee, 2010) designed their simulation with large numbers of components, in reality a test based on the MIRID with many components would be difficult to develop. The empirical data sets commonly used in MIRID-related studies were built upon one to three components (Lee, 2010). Therefore in this study only three components were designed to approximate practical settings. On the contrary, the number of item families is likely to vary in practice as evidenced by the fact that existing empirical MIRID data sets involve 5 to 12 item families. Thus a middle ground of 10 item families was used here. Because the numbers of components and families determine test length, which equals to the number of components plus one times the number of item families, there were a total of 40 items simulated for the study. Longer tests increase precision of parameter estimation at the cost of additional computing resources. The effect of varying test lengths on DIF detection in the MIRID should be investigated in future research.

The linear relationship defining the difficulty of a composite item in the MIRID requires an intercept and parameters of component weights. These parameters were simulated by emulating Wang and Jin (2010b), who, for convenience, assumed the intercept in all conditions to be zero and specified the three component weights as 0.5, 0.33, 0.17; these values were also used in this study.

In DIF research, the number of DIF-containing items is pertinent to studying differential functioning of individual items. Previous research has found that the percentage of DIF items impacts DIF detection. Too many DIF items can contaminate the conditioning variables (Gierl,

Gotzmann, & Boughton, 2004; Narayanon & Swaminathan, 1996). In practice, it is common to see 10% to 20% of items functioning differentially in conventional testing conditions (e.g., Narayanon & Swaminathan, 1996; Zhang, 2007). In this research, 20% of the items (four component items and four associated composite items) were selected as having DIF for constructing conditions. As for differential item group functioning, one component was selected for DFFc, two item families for DFFm and one component weight for DFW were used in the design. In other words, there were a total of 20 items involved in creating group level DIF for components, eight items (including two composite items) for building differential functioning item families, and 10 composite items impacted when constructing component weight DIF conditions.

Variables that varied. Since the proposed DIF models extend from the standard MIRID model, only item location parameters were necessary for data generation. Instead of fixing these parameters, the simulation added more randomness by randomly drawing values from a specified distribution, that is, a uniform distribution of $U[-2, 2]$. This method of generating true parameters was inspired by Wang and Jin (2010b) although their MIRID research was not about differential functioning.

Another important element of the MIRID is the correlation between components. With highly correlated attributes (components), the test is more unidimensional than not so that the presence of DIF may stand out and thus is likely to be detected. Conversely, low correlation may render DIF detection more difficult. The levels of component correlation specified in the original MIRID study by Butter et al. (1998) included only two levels, 0 and 0.7, although it is unlikely that the components in a test are entirely uncorrelated. Interestingly, none of the later MIRID studies considered this variable. Since all the difficulty parameters of the component items were

drawn randomly from a uniform distribution no control was applied to constrain componential correlations. As a result, these correlations varied widely across different simulation conditions.

In summary, there were in total 32 conditions constructed for the simulation study: 2 sample sizes (500, 3000) \times 2 levels of differential functioning magnitude (0.2, 0.7) \times 2 levels of population impact (0, -0.7) \times 4 levels of differential functioning (DIF, DFFc, DFFm, and DWF). A summary of these conditions is given in Table 5.

Implementation

Data Generation

Since the analysis was to be conducted using SAS, it was reasonable to employ a different computer program for data generation in order to render results more valid. WinGen 3 (Han, 2007) was selected to generate all the item parameters and response data. The following procedures were taken to create data sets under every simulation condition.

1. On WinGen, randomly drew 30 values from $U[-2,2]$ and broke them into three groups of 10 as location parameters of component items that belong with each of the three components.

Table 5.

Simulation Conditions

Cell	Sample	Impact	Type (Items involved)	Magnitude
1	500	0	DIF (8 items, 20%)	0.2
2	500	0	DIF (8 items, 20%)	0.7
3	500	-0.7	DIF (8 items, 20%)	0.2
4	500	-0.7	DIF (8 items, 20%)	0.7
5	3000	0	DIF (8 items, 20%)	0.2
6	3000	0	DIF (8 items, 20%)	0.7
7	3000	-0.7	DIF (8 items, 20%)	0.2
8	3000	-0.7	DIF (8 items, 20%)	0.7
9	500	0	DFFc (1 component, 50%)	0.2
10	500	0	DFFc (1 component, 50%)	0.7
11	500	-0.7	DFFc (1 component, 50%)	0.2
12	500	-0.7	DFFc (1 component, 50%)	0.7
13	3000	0	DFFc (1 component, 50%)	0.2
14	3000	0	DFFc (1 component, 50%)	0.7
15	3000	-0.7	DFFc (1 component, 50%)	0.2
16	3000	-0.7	DFFc (1 component, 50%)	0.7
17	500	0	DFFm (2 families, 20%)	0.2
18	500	0	DFFm (2 families, 20%)	0.7
19	500	-0.7	DFFm (2 families, 20%)	0.2
20	500	-0.7	DFFm (2 families, 20%)	0.7
21	3000	0	DFFm (2 families, 20%)	0.2
22	3000	0	DFFm (2 families, 20%)	0.7
23	3000	-0.7	DFFm (2 families, 20%)	0.2
24	3000	-0.7	DFFm (2 families, 20%)	0.7
25	500	0	DWF (10 composite, 25%)	0.2
26	500	0	DWF (10 composite, 25%)	0.7
27	500	-0.7	DWF (10 composite, 25%)	0.2
28	500	-0.7	DWF (10 composite, 25%)	0.7
29	3000	0	DWF (10 composite, 25%)	0.2
30	3000	0	DWF (10 composite, 25%)	0.7
31	3000	-0.7	DWF (10 composite, 25%)	0.2
32	3000	-0.7	DWF (10 composite, 25%)	0.7

2. Calculated location parameters of the composite items using the component item parameters drawn above and the pre-specified intercept (0) and three component weights (.5, .33, .17).
3. Imported item parameters of all 40 items into WinGen.
4. Created true theta values for the reference group according to the distribution property of $N(0,1)$.
5. Based on generated person trait scores and item parameters, simulated 500 samples of item response data for the reference group using the Rasch model because the proposed models are extensions of the standard MIRID.
6. Created another set of item parameters by randomly drawing 30 values from $U[-2,2]$ before accounting for the sources of differential functioning. For DIF, the magnitude of the current condition was added to the parameters of pre-selected four component items; for DFFc, it was added to all the items within the second component; for DFFm, the component items within two families had this effect added to their parameters; for DWF, the differential amount was added to the second of the component weight. After, the parameters of the composite items were computed accordingly. (Details of these linear relations can be found in Chapter Two.)
7. Imported this second set of 40 item parameters into WinGen.
8. Created true theta values for the focal group by following either distribution, $N(0,1)$ or $N(-.7, 1)$.
9. Based on person trait scores and item parameters for the focal group, simulated 500 item response data sets for the focal group using the Rasch model.

Research has shown that Monte Carlo simulations would have greater statistical power with more samples (e.g., Robey & Barcikowski, 1992). However, for research in social sciences, it is common to employ several hundred replications in simulation studies. Moreover, very long computing time due to complexity in modeling and analysis creates problems that force researchers to even opt for less than 100 replications (e.g., in the case of the MIRID research, Wang & Jin, 2010a; 2010b). In the DIF literature, the number of replications ranges typically from 100 to 500. For this study it was determined that 500 replications for every condition would achieve satisfactory precision albeit at a great cost of computing resources. A pilot study was conducted to test the stability of parameter estimates and found satisfactory estimation stability when there were 100 samples and that the estimates from the first 100 samples resembled closely to those from the second, third, fourth, and fifth 100 samples.

Estimation

Since much of the IRT literature that explored DIF in the GLMM framework utilized maximum likelihood estimation, the proposed models as extensions of the Rasch model also followed this method that was implemented using the SAS NLMIXED procedure.

The marginal maximum likelihood estimation (MMLE) adopted by many MIRID studies typically adopted numerical integration as the integral in this formulation does not have a closed-form solution (Tuerlinckx et al., 2004). As for the different maximization algorithms available for MMLE, the Gauss-Hermite quadrature method (Abramowitz & Stegun, 1974) was popular in earlier MIRID related research. However, analysis in this research used the non-adaptive Gaussian method instead of the more time-consuming adaptive Gaussian approach. It has been found that the two methods yielded very similar results (De Boeck & Wilson, 2004). Both Gauss-Hermite quadrature and adaptive Gaussian approaches are available through NLMIXED,

which also offers the flexibility for specifying the number of quadrature points desired. For the optimization phase, NLMIXED provides a number of options, the default of which is a dual quasi Newton-Raphson algorithm. In contrast to the original Newton-Raphson technique that calculates standard errors of the parameter estimates from the second derivative matrix of the likelihood function, the quasi-Newton approach computes only the first derivatives and thus takes much less time to run. As a note of shortcomings, mixed modeling through SAS NLMIXED is computationally intensive and requires considerable amount of memory and CPU time. In addition, the default algorithms sometimes lead to failure to converge or other estimation problems, such as the final Hessian matrix not being positive definite (Kierman, Tao, & Gibbs, 2012).

As an alternative to MMLE, Bayesian estimation was introduced in more recent MIRID studies which proposed multilevel or crossed random effect extensions of the MIRID, where a high number of parameters make it difficult to apply maximum likelihood estimation. The Bayesian Markov Chain Monte Carlo (MCMC) estimation has been offered as a valid alternative (Lee, 2010; Hung, 2011). The Bayesian method was not used in the study since the proposed models are not overly complicated. Also, the Bayesian MCMC method provides no point estimates of parameters but distributions of parameters. In addition, convergence evaluation is less straightforward with MCMC than with MMLE approach. Model comparison based on maximum likelihood is a powerful feature of this approach.

In conclusion, MMLE as implemented by SAS NLMIXED has been the engine behind much of psychometric modeling research, including earlier studies to extend and generalize the MIRID. This is logical since the MIRID and its extensions can be viewed as special cases of generalized nonlinear mixed models. Also, the well-understood properties of MMLE and the

ease in implementation and output interpretation with NLMIXED attract researchers. Because the models proposed in this dissertation are not overly complex, MMLE using NLMIXED is selected to perform estimation despite the downside that this method can be very demanding of computing resources.

Analysis Procedures

Before proceeding with the MIRID differential functioning analysis, a recovery analysis was conducted to determine the extent to which the generating parameters could be recovered from the simulated data sets simulated using the WinGen program. For this purpose, 500 data sets were simulated from the standard MIRID on the basis of the study design discussed above but minus the differential functioning effects.

In keeping with the two-fold research objective, the analysis of differential functioning consisted of two parts: to examine the efficacy of the proposed DIF models in detecting the corresponding source of differential effects and to investigate the impact from fitting the DIF models to the data with mismatched DIF source. Steps taken in the analysis under every condition are as follows, the details of which can be found in the SAS example code in Appendix A.

1. For each replication, read in the output files from WinGen and combined the reference group and focal group response data sets.
2. Converted the combined data set into “long format” by putting all item responses in one column.
3. Created indicator variables for items, item families, and components.
4. Fitted MIRID DIF models one at a time to the combined data and saved the output parameter estimates data.

Evaluation Procedures

The accuracy and precision of parameter estimates were gauged through recovery analysis. The root mean square error (RMSE) and bias were calculated for estimated parameter values to assess the deviance between the generating values and estimated ones. According to Sinharay, Grant, and Blew (2009), bias can be defined as

$$Bias = \frac{1}{S} \sum_{s=1}^S \hat{\delta}_r - \delta, \quad (39)$$

where S is the number of replications, δ refers to the true magnitude of differential functioning whereas $\hat{\delta}_r$ the estimated magnitude in the r^{th} replication. Note that the subscript for an item, or a facet, or a component weight is omitted. On the other hand, RMSEs can be expressed as follows:

$$RMSE = \sqrt{\frac{1}{S} \sum_{s=1}^S (\hat{\delta}_r - \delta)^2}, \quad (40)$$

using the aforementioned notation. Comparison of RMSE and bias reveals estimation accuracy of fitted models under a certain data construction condition.

Often when a DIF model is implemented, null hypotheses (“No DIF”) are tested against the estimates of the differential functioning parameters. In this study, SAS NLMIXED procedure produced a maximum likelihood estimate for every model parameter, including a number of zero-value DIF parameters (individual or group), followed by a p -value based on a t distribution with approximate degrees of freedom computed as the number of subjects minus the number of random effects. If the estimate was not zero with a p -value significant at .05 level, the true null hypothesis of no differential functioning was incorrectly rejected and a Type I error was committed (a “false positive”). The efficacy of the proposed models in DIF detection was assessed by computing and examining the Type I error rates and statistical power.

Testing a host of null hypotheses simultaneously gives rise to the problem of Type I error inflation. With a nominal α level that is typically pre-determined for probability, different types of Type I error rates were defined (Ryan, 1959): the error rate in a single hypothesis test (per-comparison or hypothesiswise Type I error), the average number of errors in a host (“family”) of hypotheses tests (per-experiment Type I errors), and the probability of one or more errors in a host of hypothesis tests (experimentwise or familywise Type I error). Of the three, the experimentwise error rate (α_{EW}) has often been the main concern with multiple hypotheses testing (Kromrey & La Rocca, 1995). It can be defined as $\alpha_{EW} = 1 - (1 - \alpha)^c$, where c represents of the number of tests. It is obvious that experimentwise Type I error can grow fast as the number of tests increases. In this study, the number of per-experiment Type I errors was not computed.

In DIF studies, effective control of Type I error has always been of interest. Adjustment procedures that protect hypothesis testing from inflated Type I errors include the Bonferroni correction and Hochberg’s sequential procedure (Hochberg, 1988), which is a modified Bonferroni correction procedure. The Bonferroni procedure simply calculates a new alpha to keep the experimentwise alpha value at .05 (or another specified value). The formula for calculating the adjusted significance level is $\alpha_{HW} = \frac{\alpha_{EW}}{c}$, where α_{HW} is the new alpha to evaluate each hypothesiswise significance test. The Bonferroni correction is probably the most commonly used post hoc test for its flexibility and simplicity; however, it is conservative and lacks statistical power. For the Hochberg approach, the hypotheses are first ordered according to their associated p -value in ascending order. The sequential adjustment of the experimentwise alpha calculates $\frac{\alpha_{EW}}{c}$ as the first of the set of criteria for the reject/fail-to-reject decision and

$\frac{\alpha_{EW}}{c-1}$ as the second in line, and so on, with the final one being $\frac{\alpha_{EW}}{1}$. That is, evaluation starts with

the smallest observed p -value evaluated against $\frac{\alpha_{EW}}{c}$, followed by the second smallest p -value compared with $\frac{\alpha_{EW}}{c-1}$, and proceeds up the set of p -values until a null hypothesis is rejected.

Details of the criteria calculation and evaluation can be found in Kromrey and Hogarty (2002). In addition, researchers can certainly opt for no adjustment at all when they conduct an unprotected test for every hypothesis.

For hypothesis testing, statistical power refers to the probability of rejecting a false null hypothesis. Researchers often target maximizing statistical power while maintaining the probability of a Type I error at or below the pre-determined level. Analogous to the aforementioned three types of Type I errors, three kinds of statistical power were promoted in the literature (Seaman et al., 1991). Any-pairs power is the probability of rejecting at least one false null hypothesis in the entire set. In contrast, all-pairs power is the probability of rejecting all the false null hypotheses across the tests, which never exceeds the any-pairs power. Finally, per-pair power is the probability of rejecting each false null hypothesis in the entire host of hypotheses. The all-pairs power index is naturally the lowest of the three and was not calculated for this research because the per-pair power has been the most commonly used in DIF studies and the any-pairs index is the most relevant in the context of multiple hypotheses testing.

Meaningful power comparisons across conditions depends on well controlled Type I error rates. Inflated Type I error rates result in overestimated power and deflated Type I error rates lead to underestimated power. After Type I error rates were evaluated, the ability of each model to correctly identify items with DIF (power) was examined. Specifically, the per-pair and any-pairs indices of power were calculated. Power were assessed using Cohen's (1988) standard of .8 or greater (at $\alpha = .05$) as being adequate.

To review the steps of analysis, after data sets had been simulated, bias and RMSE were computed for all model parameters as well as the differential functioning parameters for the 500 replications under each condition. For the conditions where every extended MIRID model (DIF, DFFc, DFFm, and DWF) was fitted to data generated based on themselves, different types of Type I error rates and power were calculated for each condition before being compared and evaluated. These evaluative measures were taken in order to address the following questions:

- 1.) How well are the Type I error rates controlled? When the Type I error control is maintained, what is the power of the MIRID DIF, DFFc, DFFm, and DFW models in detecting differential functioning of different sources?
- 2.) How accurate are the parameter estimates of these models, including the DIF parameters, item locations, component weights, and impact (γ_g)?
- 3.) How do the following such factors affect the performance of the proposed differential functioning approach as sample size, DIF magnitude, and group differences in trait level?

To investigate the effect of applying the unmatched model to study differential functioning in the context of the MIRID, the following questions were addressed based on the analysis results:

- 4.) How well are the model parameters estimated if the wrong models are fitted to the data? Are they more adversely impacted under some conditions than others?
- 5.) Are any of the estimates of the incorrectly specified DIF parameters statistically significant? Which differential effects in the data produce the most misleading findings when the unmatched model is fitted?

CHAPTER FOUR

RESULTS

Analysis results of this study are presented in three parts: First, outcome from the pilot study was provided. Since the proposed MIRID differential functioning approaches are model-based, their performance on detection mainly depends on how well the relevant parameters are recovered. Accordingly, a discussion on the recovery of DIF relevant and non-DIF parameters formulated in the four models, including the MIRID differential item functioning model (DIF), the model for differential facet functioning in components (DFFc), the model for differential facet functioning in item families (DFFm), and the model for differential facet functioning in component weights (DWF), is presented in the second section. In this section the generating MIRID differential functioning (“true”) models were fitted to their generated data so that the Type I error control and power in DIF testing for these models were calculated and presented. In the last part, each of the proposed models was applied to data simulated with the other (“wrong”) models. Because the DIF parameters (δ) in these scenarios all had a true value of zero, only false detection rates were discussed here before parameter recovery in these conditions was described.

Parameter Recovery of the Standard MIRID

Before the DIF simulation study, a simulation study of parameter recovery for the standard MIRID model was conducted to evaluate the accuracy and precision of the parameter estimation methods used in this research. Specifically, a standard MIRID model mimicking the

DIF simulation design which includes 10 item families and 3 components was used to generate data. Table 6 lists the 35 parameters, including 30 item location parameters ($b_1 - b_{30}$), one intercept (w_0), three component weights ($w_1 - w_3$), and the standard deviation of the sample (sd), that were produced by the WinGen program using the method described in Chapter Three. Five hundred data sets of 1,000 persons were then created based on these parameters and analyzed with SAS PROC NLMIXED. (1,000 is common as a medium-level sample size in educational DIF studies.)

Bias and RMSE results from the recovery analysis are given in Table 6. Bias values for 30 item location parameters ranged from .027 to .070, indicating slight over-estimation which is on par with the results by Butter et al. (1998) where the standard MIRID was introduced with full details. All but two (.140 and .160) of the RMSE values were low, ranging from .107 to .132, suggesting acceptable estimation. In contrast, the intercept (-.010) and three component weights parameters (-.003, -.005, and -.001), and standard deviation of the sample (-.036) were underestimated with almost negligible bias. Their associated RMSE were also small, ranging from .037 to .056. In summary, parameter recovery of the standard MIRID confirmed that the data generation and parameter estimation approaches were adequate and would be employed for study of differential functioning models.

Results for the Proposed Differential Functioning Models

For each proposed model, eight simulation conditions were constructed by crossing three factors: sample size, magnitude of differential functioning in items or item groups (DIF size), and group difference in trait level (impact). For each condition, 500 replications were simulated, resulting in 16,000 data sets in total. In the first phase of the study, the generating models were applied to the data simulated based on themselves to investigate parameter recovery and Type I

error rates and power (a.k.a. when they were the “true models”). Results from phase one are presented in Table 6.

Recovery of the True DIF (Delta) Parameters

Defined in the MIRID DIF model are thirty DIF parameters, four of which had non-zero values (d2, d14, d16, and d28) that represented the differences in item location between the focal and reference groups. The other 26 were delta parameters considered to have true value of zero in the analysis. The average bias for the four non-zero delta parameters from the 500 replications is displayed in Table 7, where every value is negative, suggesting that all the delta parameters were under-estimated in the eight conditions. In addition, there is much variation both in the estimates of each parameter and between the estimates in the same condition when true values are the same. For example, in the condition with larger delta (.7), non-zero impact (.7), and smaller sample (500), bias in estimates of d2 (-.113) and d28 (-.110) were much greater than those of d14 (-.039) and d16 (-.045). Such variation points to the fact that the quality of estimation was less than optimal.

Figure 1 graphs the average bias of the four estimated delta parameters with the left graph showing the four conditions of smaller sample size and the right one the four conditions with larger sample size. Dashed lines represent larger delta conditions and solid lines smaller delta conditions. Clearly, larger delta size resulted in significantly more bias than smaller delta for both sample size conditions (-.101 and .030 on average). Also, larger samples increased average bias but to a smaller extent (-.058 versus -.074). From these graphs, it seems that the presence of population difference (impact) decreases the bias only slightly, which is more obvious in larger delta conditions.

Table 6.

Recovery of Item Location Parameters of the Standard MIRID

Parameter No.	Parameter	True Value	Bias	RMSE
1	b1	1.366	0.057	0.110
2	b2	1.221	0.057	0.123
3	b3	-1.152	0.050	0.107
4	b4	0.750	0.061	0.111
5	b5	-1.441	0.040	0.112
6	b6	1.957	0.070	0.140
7	b7	-0.949	0.061	0.113
8	b8	-1.227	0.065	0.118
9	b9	0.299	0.060	0.113
10	b10	-0.635	0.056	0.109
11	b11	-1.700	0.028	0.124
12	b12	0.912	0.038	0.112
13	b13	-0.944	0.053	0.108
14	b14	-1.294	0.031	0.111
15	b15	0.075	0.048	0.101
16	b16	1.211	0.062	0.132
17	b17	0.519	0.062	0.112
18	b18	-0.304	0.050	0.113
19	b19	-1.893	0.062	0.160
20	b20	-1.924	0.027	0.128
21	b21	-0.085	0.044	0.110
22	b22	1.573	0.041	0.124
23	b23	-0.807	0.035	0.119
24	b24	-0.160	0.063	0.112
25	b25	-0.683	0.068	0.117
26	b26	1.155	0.055	0.123
27	b27	-0.988	0.048	0.113
28	b28	-0.677	0.043	0.120
29	b29	1.959	0.045	0.131
30	b30	-0.096	0.042	0.111
31	w0	-0.010	-0.010	0.051
32	w1	0.497	-0.003	0.045
33	w2	0.325	-0.005	0.037
34	w3	0.169	-0.001	0.056
35	sd	0.964	-0.036	0.041

Table 7

Bias of the Non-zero DIF Parameter Estimates under the MIRID DIF Model

			d2	d14	d16	d28
delta=.2	N=250*2	Impact=0	-0.013	-0.015	-0.039	-0.053
		Impact=-.7	-0.035	-0.019	-0.006	-0.038
	N=1500*2	Impact=0	-0.050	-0.031	-0.021	-0.037
		Impact=-.7	-0.038	-0.028	-0.026	-0.031
delta=.7	N=250*2	Impact=0	-0.141	-0.068	-0.072	-0.116
		Impact=-.7	-0.113	-0.039	-0.045	-0.110
	N=1500*2	Impact=0	-0.152	-0.099	-0.098	-0.135
		Impact=-.7	-0.122	-0.101	-0.100	-0.111

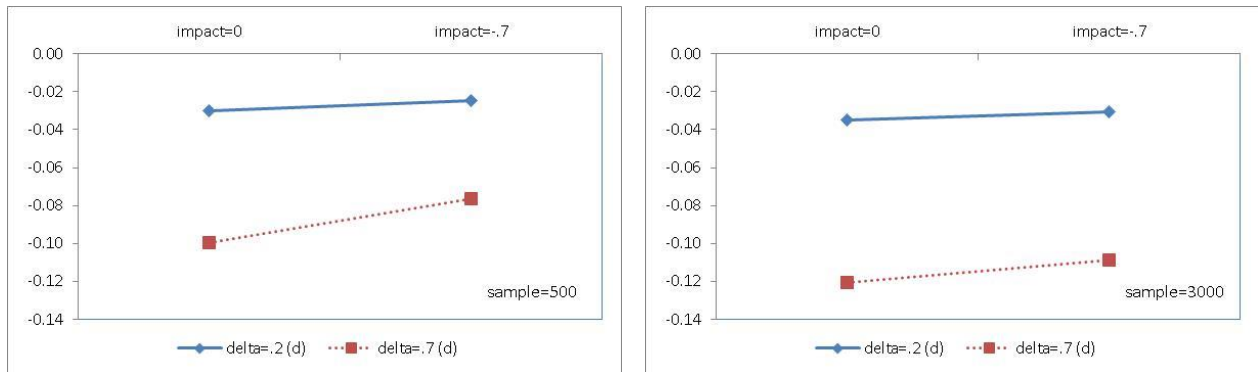


Figure 1. Average bias of the non-zero DIF parameter estimates under the MIRID DIF model by sample size

Table 8 lists the RMSEs of the estimates of the four non-zero DIF parameters and Figure 2 plots the average RMSEs of these estimates under the MIRID DIF model by sample size. As expected, larger sample sizes (sample=3,000) decreased RMSE for all parameters (<.2). Conversely, larger delta magnitude increased the sizes of RMSE in either sample size condition; specifically, a negligible increase with smaller samples but conspicuous increase for larger samples. For example, with larger samples and zero impact, RMSE went up from .091 to .169 as

delta rises from .2 to .7 for d2; with smaller samples and zero impact, however, the increase was only from .192 to .231 for the same parameter. Figure 2 confirms the prominent influence of sample size not only in reducing the RMSEs significantly but also in widening the gap in average RMSE between the two delta sizes. Also, it was obvious from Table 8 that there was very little difference in RMSEs between the two impact levels, suggesting virtually no effect from this factor. This is confirmed graphically on Figure 2, where the dotted line of larger delta runs parallel to the solid line of smaller delta in both plots and reveals a very weak relationship between group difference and DIF magnitude.

Table 8.

RMSEs of the Non-zero Delta Parameter Estimates under the MIRID DIF Model

			d2	d14	d16	d28
delta=.2	N=250*2	Impact=0	0.192	0.214	0.244	0.266
		Impact=-.7	0.227	0.225	0.226	0.262
	N=1500*2	Impact=0	0.091	0.092	0.087	0.086
		Impact=-.7	0.087	0.119	0.085	0.090
delta=.7	N=250*2	Impact=0	0.231	0.217	0.234	0.274
		Impact=-.7	0.258	0.211	0.201	0.310
	N=1500*2	Impact=0	0.169	0.130	0.127	0.160
		Impact=-.7	0.145	0.163	0.130	0.142

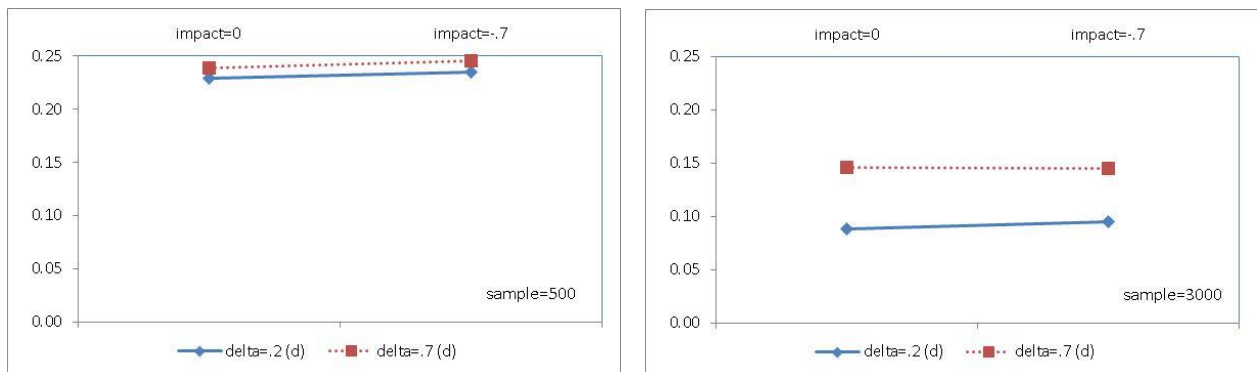


Figure 2. Average RMSEs of the non-zero DIF parameter estimates under the MIRID DIF model by sample size

The DIF detection performance of the MIRID DIF model can be evaluated from another perspective by examining the estimation of the locations of the component items that were associated with the four delta parameters. This is because the estimation process may not be able to distinguish the delta and location parameters; thus the under-estimation of the former may lead to the over-estimation of the latter, and vice versa. As seen from Table 9, the four items (b2, b14, b16, and b28) were all over-estimated but considerably so in larger delta conditions with bias ranging from .135 to .237. Such over-estimation is also shown graphically in Figure 3 where RMSEs of all 30 location parameters are plotted for all four conditions formed by sample and impact sizes. In each plot, the solid line represents RMSEs in the smaller delta (.2) condition while the dashed line connects the 30 RMSEs for the larger delta (.7) condition. Note that the spikes on the larger-delta line in all four conditions are pronounced for the four DIF-associated component items. Overall, larger delta led to more significant bias and RMSEs; for smaller delta conditions, only larger sample coupled with zero impact caused more conspicuous RMSEs for delta-associated component items. For instance, for b28, the condition of smaller delta, larger sample and zero impact had slightly larger bias (.087) than smaller delta, larger sample, and impact (.056). Table 8 also shows that larger sample size inflated bias but to a lesser degree than delta size and that the presence of group difference (impact) decreased bias slightly.

In Chapter Two, the MIRID extensions for item group level differential functioning were formulated using Equation 36, 37, and 38 as the DFFc, DFFm, and DWF models, respectively. The item-group level differential effect is involved in modeling individual item responses (see Chapter Two for details). The effectiveness of these models in DIF detection determines whether

in the context of the MIRID differential functioning can be explained effectively with common characteristics in a group of items.

In the DFFc model, three delta parameters were defined, each representing a component but only one of them (kd2) was given a non-zero value (delta) in the simulation. For the DFFm model, 10 DIF parameters were formulated to correspond to the 10 item families, two of which (fd3 and fd7) were simulated with nonzero values. Similarly, the second of the three DIF parameters (wd2) in the DWF model was simulated as the delta parameter while the other two had true values of zero. Table 10 and Table 11 list the RMSE and bias of these four item-group level delta parameters over 500 replications. Clearly, all group-level delta parameters were estimated adequately as the bias values were very close to zero (mostly between -.004 and .004) and there was sign of neither over- nor under-estimation. With sample size and impact held constant, larger DIF led to greater RMSE. The estimation appeared to be the worst in the condition of smaller sample, larger DIF, and larger impact. The mean estimate of wd2 in the condition of larger sample, no impact, and larger DIF deviated the most from the true value (bias = .021).

Table 9.

Bias of the DIF-related Item Location Parameters under the MIRID DIF

			b2	b14	b16	b28
delta=.2	N=250*2	Impact=0	0.053	0.061	0.055	0.072
		Impact=-.7	0.038	0.047	0.047	0.040
	N=1500*2	Impact=0	0.089	0.083	0.086	0.087
		Impact=-.7	0.053	0.036	0.046	0.056
delta=.7	N=250*2	Impact=0	0.196	0.192	0.193	0.206
		Impact=-.7	0.172	0.218	0.216	0.168
	N=1500*2	Impact=0	0.237	0.215	0.220	0.234
		Impact=-.7	0.211	0.135	0.223	0.172

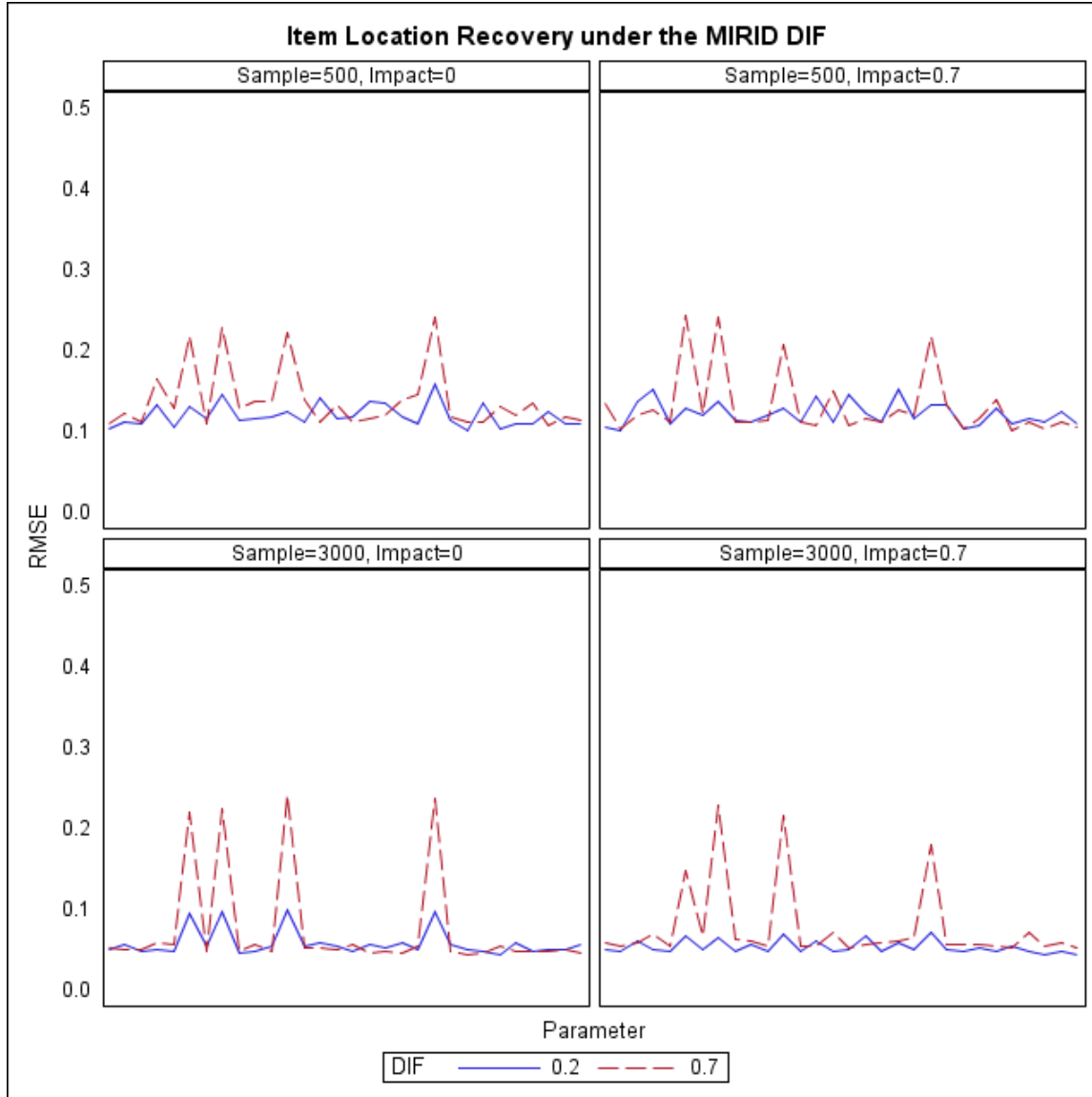


Figure 3. Recovery of the location parameter of component items under the MIRID DIF model by sample and impact

Figure 4 displays the RMSE of the DFFc model delta estimate (kd2) by sample size, where it is clear that DIF size and impact had no influence on estimation and the only influential factor was sample size. There is a difference of approximately .04 in magnitude between the RMSEs from larger sample conditions and smaller sample ones. In Figure 5, the averages

RMSEs of the two delta parameters in the DFFm model were plotted. Once again, larger sample size lead to smaller RMSEs by at least .04 when larger DIF size also reduced RMSEs, especially for non-zero impact conditions, suggesting an interaction between delta size and impact. The RMSEs of the delta parameter in the MIRID DWF model are graphed in Figure 6. Larger sample size and smaller DIF magnitude resulted in smaller RMSEs and the interaction between impact and DIF size appears to be in opposite direction to that in the DFFm model: the non-zero impact did not reduce but increased RMSEs. Across all the three group DIF models, the RMSEs for larger sample conditions were acceptable (<.05).

Table 10

RMSE of the Delta Parameters under the MIRID DFFc, DFFm, and DWF Models

			kd2	fd3	fd7	wd2
delta=.2	N=250*2	Impact=0	0.069	0.068	0.067	0.048
		Impact=-.7	0.068	0.063	0.070	0.067
	N=1500*2	Impact=0	0.027	0.026	0.025	0.027
		Impact=-.7	0.028	0.027	0.027	0.033
delta=.7	N=250*2	Impact=0	0.068	0.098	0.070	0.093
		Impact=-.7	0.070	0.121	0.114	0.089
	N=1500*2	Impact=0	0.028	0.029	0.028	0.034
		Impact=-.7	0.029	0.034	0.054	0.034

Note that because estimation of these group-level delta parameters was at an acceptable level, there was no need to examine the estimation quality of the location parameters of their associated component items.

Table 11

Bias of the Delta Parameters under the MIRID DFFc, DFFm, and DWF Models

			kd2	fd3	fd7	wd2
delta=.2	N=250*2	Impact=0	0.001	0.001	0.000	0.005
		Impact=-.7	-0.002	0.001	-0.004	0.007
	N=1500*2	Impact=0	0.001	0.000	0.002	0.002
		Impact=-.7	0.000	-0.001	-0.001	0.002
delta=.7	N=250*2	Impact=0	0.003	-0.004	0.001	0.005
		Impact=-.7	0.001	0.011	0.002	0.021
	N=1500*2	Impact=0	0.002	-0.002	-0.003	-0.002
		Impact=-.7	-0.003	0.001	-0.003	0.001

Recovery of Zero-value DIF Parameters

Estimation of the delta parameters directly affects the power of DIF detection of the proposed models. On the other hand, estimation quality of the zero-value DIF (non-delta) parameters formulated in the proposed models influences the Type I error rates in their detection. A Type I error is committed when the estimated parameter is significantly different than zero. Figure 7 presents the average estimation bias of the 26 non-delta DIF parameters in the MIRID DIF model by sample size. (Average bias and RMSE for estimates of the zero-value DIF parameters in the four models are provided in Appendix B.) Overall, bias was acceptable in smaller DIF size conditions at about -.02 for either sample condition. The levels of bias were so much higher in larger delta conditions ($< -.08$) that DIF size alone was the dominant factor for these results. Figure 7 also demonstrates very weak effects from impact on recovery of non-delta DIF parameter as it decreased the bias slightly for larger delta under both sample size conditions but had no effect when the DIF magnitude was small.

In Figure 8, the average RMSEs of the non-delta parameter estimates for smaller sample (.167) are greater than those for larger sample (.154). From another perspective, the level of

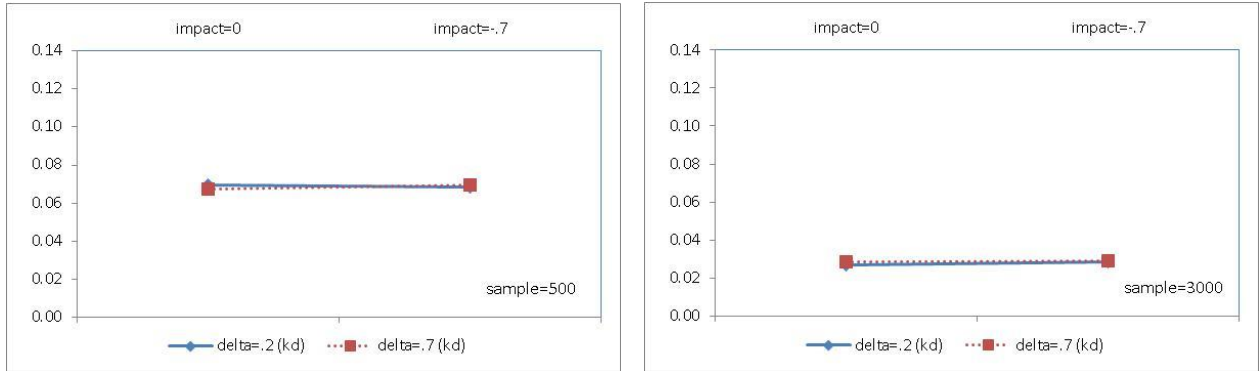


Figure 4. RMSEs of the non-zero DIF parameter estimates under the MIRID DFFc by sample size

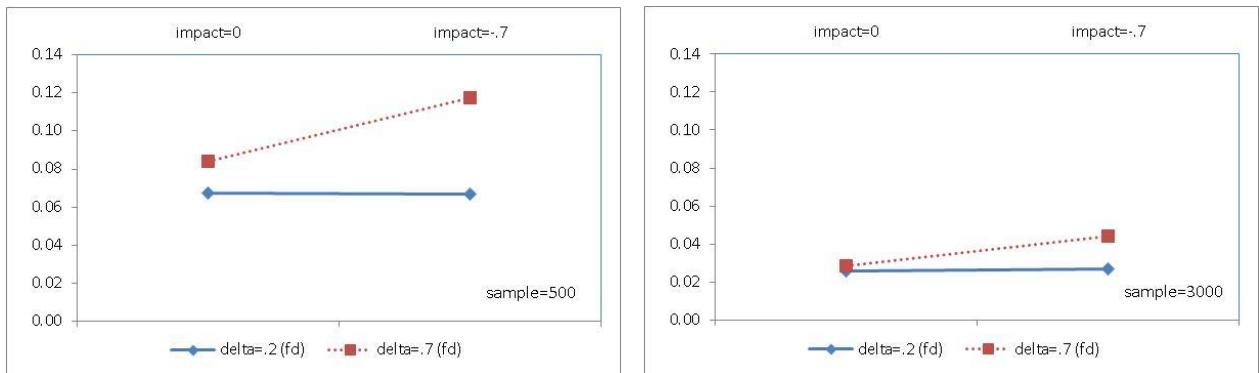


Figure 5. Average RMSEs of the non-zero DIF parameter estimates under the MIRID DFFm by sample size

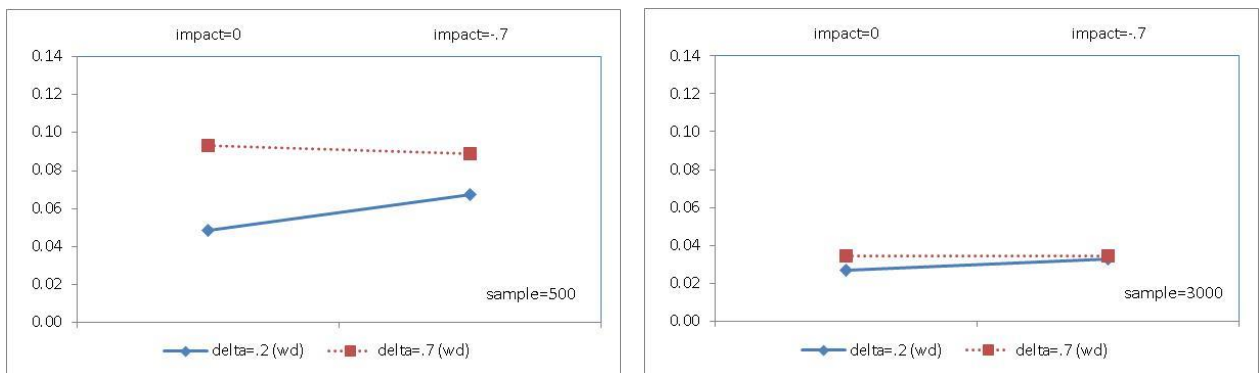


Figure 6. RMSEs of the non-zero DIF parameter estimates under the MIRID DWF by sample size

RMSEs for smaller delta was significantly lower (.114) than larger delta conditions (.226). In addition, larger sample also increased the gap in RMSE between the two delta sizes. Note that the plots in Figure 7 and 8 resemble those in Figure 1 and 2 that depict the average bias and RMSEs of the non-zero DIF parameters. Therefore, there was virtually no difference in estimation quality between the delta and zero-value DIF parameters.

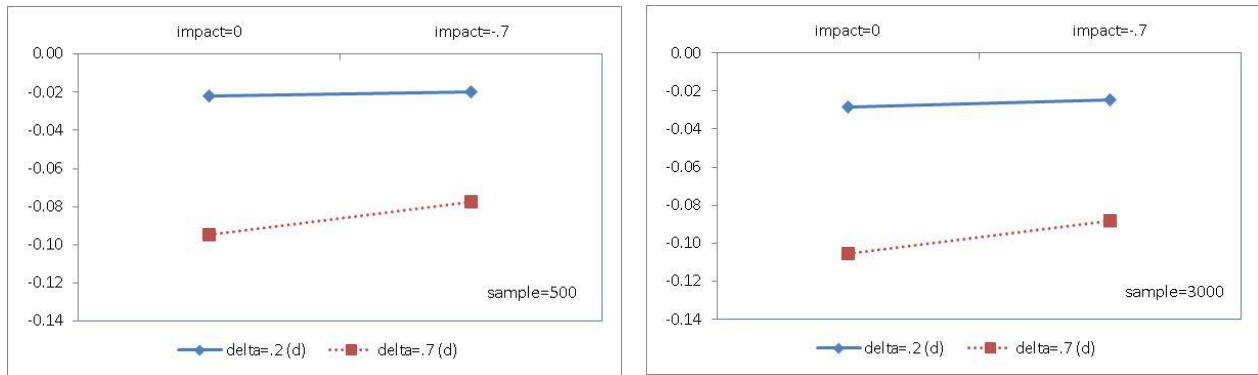


Figure 7. Average bias in estimation of zero-value DIF parameters in the MIRID DIF model

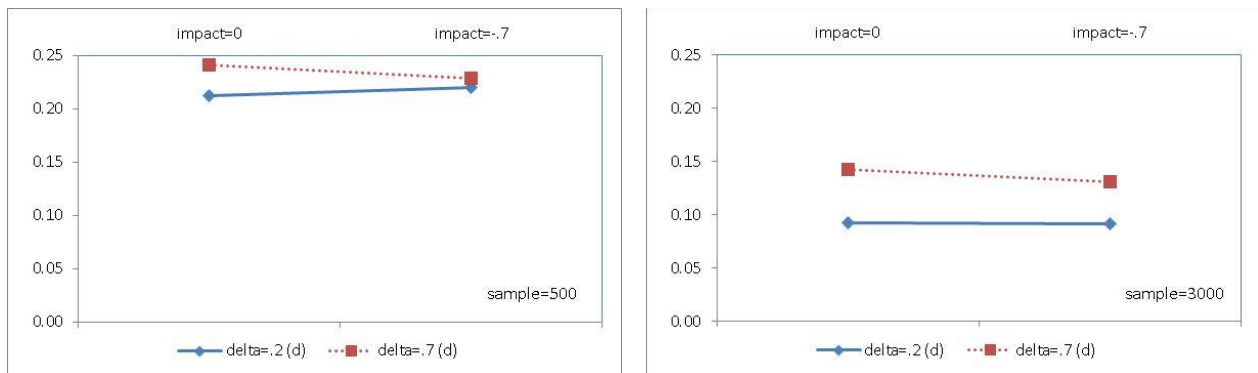


Figure 8. Average RMSEs in estimation of zero-value DIF parameters in the MIRID DIF model

As discussed previously, a number of zero-value DIF parameters are included in the formulation of every group-level differential functioning model: two component DIF parameters (kd) in the DFFc model, eight item family DIF parameters (fd) in the DFFm model, and two

component weight DIF parameters (wd) in the DWF model. Figure 9 and 10 give their average estimate bias in conditions formed by DIF magnitude and sample size. In the left graph of Figure 9 (smaller delta and sample), the zero-value DWF parameters were noticeably over estimated and such over-estimation lessened with larger sample size as shown in the right graph. In Figure 10, larger delta, smaller sample, and large impact resulted in huge bias for the DWF parameters (left) but with larger sample all the bias fell to acceptable level (right). Overall, bias of zero-value DIF parameter estimates in the DFFc and DFFm models were acceptable in all conditions and the zero-value DWF parameter estimates were the most volatile.

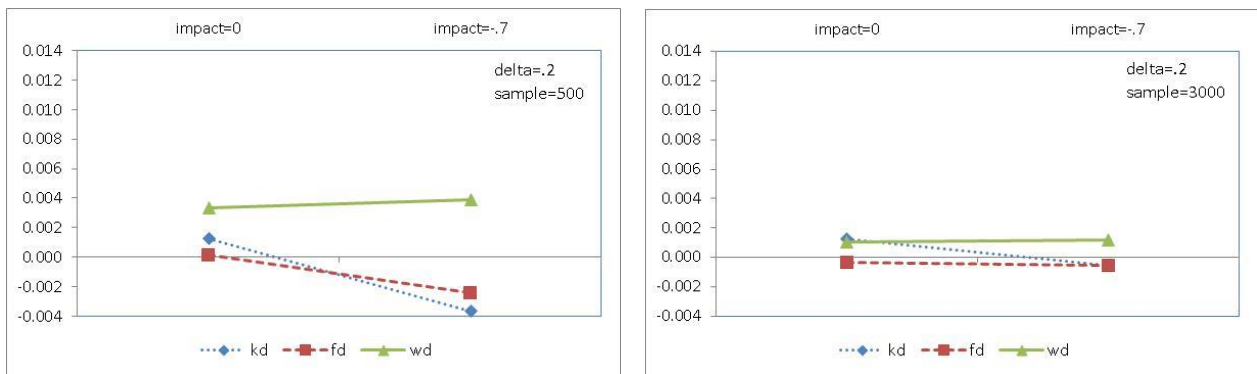


Figure 9. Average bias in estimation of zero-value DIF parameters in the MIRID DFFc, DFFm, and DWF models (smaller delta conditions)

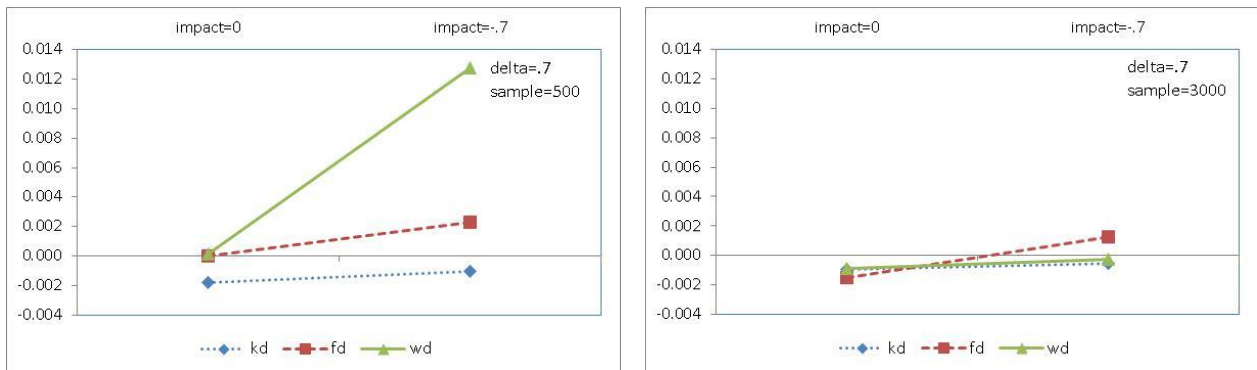


Figure 10. Average bias in estimation of zero-value DIF parameters in the MIRID DFFc, DFFm, and DWF models (larger delta conditions)

A similar scenario can be found in Figure 11 and 12 where the average RMSEs for these group-level zero-value DIF parameters are graphed. Once again, for these group-level DIF parameters, DIF size was not the prominent factor in determining their estimation quality; rather, RMSEs decreased conspicuously in large sample size conditions. Estimation quality of the zero-value DIF parameters and those of the delta parameters in the group-level DIF models is essentially the same.



Figure 11. Average RMSEs in estimation of zero-value DIF parameters in the MIRID DFFc, DFFm, and DWF models (smaller delta conditions)

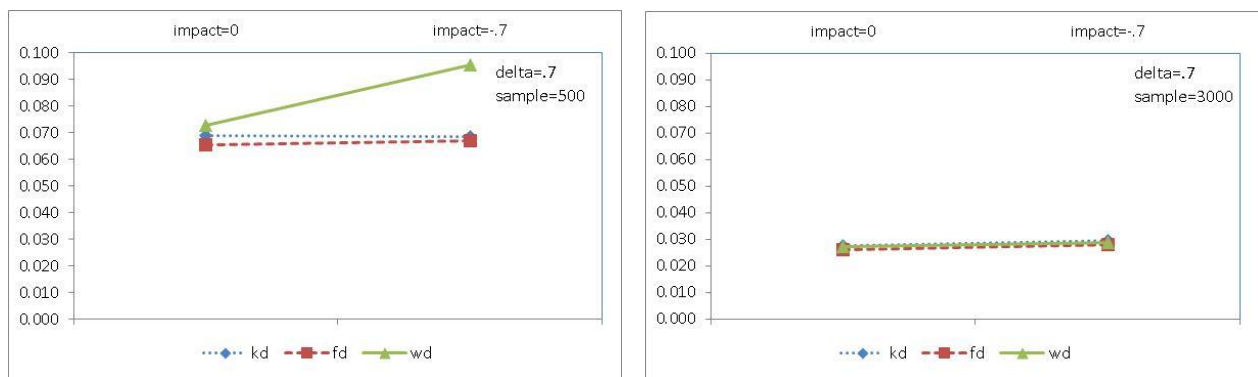


Figure 12. Average RMSEs in estimation of zero-value DIF parameters in the MIRID DFFc, DFFm, and DWF models (larger delta conditions)

In addition, recovery of the other parameters of the MIRID DIF, DFFc, DFFm, and DWF models was evaluated and found to be acceptable with no unusual patterns. Since their estimation quality does not bear direct effect on study of differential functioning of individual items or item groups, their recovery will not be discussed here but the details in the form of bias and RMSEs can be found in the Appendices C.

This section answered the second and third research questions, which focus on estimation quality of the proposed models in terms of parameter recovery and on the effects from manipulating factors. In summary, recovery of the delta and other parameters for the three item group DIF models was found to be adequate. For these models, larger samples reduce both bias and RMSE; the effect from the magnitude DIF was not strong.

For the individual item DIF model, the recovery was less than acceptable, especially in larger DIF size conditions where the magnitude of delta was shown to be the most influential factor. In particular, items associated with a real DIF parameter had their locations over-estimated consistently while the paired delta parameters were consistently under-estimated. Overall, both the four delta and 26 zero-value DIF parameters were underestimated with considerable and varying RMSEs, even under larger sample conditions. Larger sample size decreased the RMSEs of the delta and zero-value DIF parameters alike but not the bias. The effect from between-group impact was mostly weak and inconsistent.

Type I Error Control and Power of the MIRID Differential Functioning Models

A Type I error occurs when an item, item group, or an item weight was identified as functioning differentially but was not simulated with any differential effect. In the analysis, the SAS NLMIXED procedure produced a maximum likelihood estimate for every model parameter, including the zero-value DIF parameters (individual or group), followed by a *p*-value based on a

t distribution with approximate degrees of freedom computed as the number of subjects minus the number of random effects. If the estimate was not zero with the p -value significant at .05 level, the null hypothesis of no differential functioning was incorrectly rejected and a Type I error was committed.

In the MIRID DIF model, 26 of the 30 DIF parameters were simulated to have a true value of zero. In the DFFc, DFFm, and DWF models, the numbers of zero-value DIF parameters are two, eight, and two, respectively. A total of eight conditions were simulated for each of the four proposed models and 500 data sets were generated for each condition. Thus, the empirical Type I error rate was calculated for each simulation condition over the 500 replications. The liberal definition of robustness by Bradley (1978) that empirical Type I error rate should not exceed .075 at a nominal alpha of .05¹ was employed in this study as the criterion to determine whether the Type I error rate is adequately controlled or not.

Two kinds of Type I error rates were calculated: per-comparison (hypothesiswise) rates (PCER) that capture the probability of a false rejection of a single null hypothesis and experimentwise rates (EWER) as the probability of one or more Type I errors in the set of null hypothesis tests within each replication. Typically used in studies of differential item functioning, PCER is simply calculated as the number of false positives divided by the number of null hypotheses across all replications. EWER (also known as familywise Type I error rates), on the other hand, applies when a host of null hypotheses is being tested at once, or in the context of DIF studies, multiple items being tested for differential effects. Calculation of EWER involves treating each replication/sample data as an experiment and finding the probability of at least one Type I error within the experiment. Three decision criteria were used in the calculation: the

¹ Bradley (1978) also defined a lower bound for the robustness at .025 for a nominal alpha of .05. Since the harm of inflated Type I errors outweighs that of conservative ones, only the upper bound of .075 was applied in this research.

unprotected testing, the Bonferroni correction, and the Hochberg procedure, the latter two of which aim at avoiding inflation of Type I error resulting from multiple significance testing in a single replication (details can be found in Chapter Three).

Table 11 lists the empirical Type I error rates across the eight conditions under the MIRID DIF model. The error rates from the unprotected approach were calculated without adjusting for multiple tests and are thus naturally higher than those with the Bonferroni or Hochberg procedure applied. From the results for smaller delta conditions presented in the top half of the table, it is clear that the two adjustment methods led to very conservative PCER according to Bradley's criterion and that higher but still acceptable control was found in adjusted EWER. Note that the unprotected testing results under PCER are what typically get reported in DIF literature and they looked well controlled in all smaller delta conditions but with EWER this approach generated very high error rates. For instance, in the condition of no impact and smaller sample, the overall Type I error occurred to only 5.1% of the hypotheses based on PCER but they occurred in 73.8% of the generated data sets.

In the bottom half of Table 12 where results in larger delta conditions are reported, the unprotected PCER were not controlled in larger sample conditions (underlined) although the two sets of adjusted PCER were no longer conservative but still controlled. The EWER results on the right hand side were all above the Bradley upper bound of .75 even with the two adjustment procedures. Note that in all these conditions, the Bonferroni and Hochberg procedures gave very similar, if not the same, PCER and EWER.

Figure 13 depicts the influential effects from sample size and DIF magnitude by graphing the Hochberg-adjusted EWER by sample size. For smaller delta (the solid lines), the EWER were slightly higher with larger samples but still within the Bradley bound. Larger delta

conditions, however, resulted in conspicuously heightened EWER with larger samples. The presence of group difference seemed to lower the EWER somewhat for greater delta conditions.

Table 12

Type I Error Rates for MIRID DIF over 500 Replications

		# True H0	Per-comparison %			Experimentwise %			
			Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg	
delta=.2	N=250*2	Impact=0	26	0.051	0.002	0.002	<u>0.738</u>	0.040	0.040
		Impact=-.7	26	0.049	0.002	0.002	<u>0.710</u>	0.038	0.038
	N=1500*2	Impact=0	26	0.065	0.002	0.002	<u>0.846</u>	0.058	0.058
		Impact=-.7	26	0.062	0.002	0.002	<u>0.814</u>	0.050	0.050
delta=.7	N=250*2	Impact=0	26	<u>0.079</u>	0.004	0.005	<u>0.882</u>	<u>0.108</u>	<u>0.110</u>
		Impact=-.7	26	0.069	0.004	0.004	<u>0.852</u>	<u>0.088</u>	<u>0.088</u>
	N=1500*2	Impact=0	26	<u>0.253</u>	0.040	0.045	<u>1.000</u>	<u>0.686</u>	<u>0.702</u>
		Impact=-.7	26	<u>0.192</u>	0.025	0.028	<u>1.000</u>	<u>0.494</u>	<u>0.522</u>

Note: Underlined values represent inflated Type I error rates according to Bradley's liberal criterion of robustness.

These Type I error rates are consistent with the recovery results of the zero-value DIF parameters under the MIRID DIF (Figure 7 and 8), where it is clear that larger delta resulted in increased bias and RMSE. Specifically, the highest Type I error rates occurred as a consequence of larger DIF interacting with larger sample. The effect from the factor of group difference was not strong.

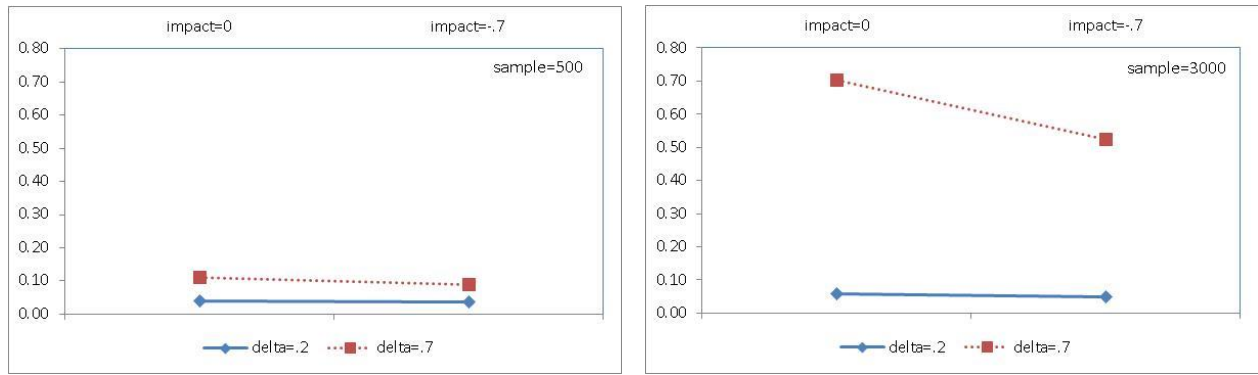


Figure 13. The MIRID DIF model experiment-wise Type I error rates after Hochberg adjustment by sample size

As discussed in Chapter Three, two kinds of power were computed for each condition. The “per-pair” index represents the power for every hypothesis test and was calculated as the percentage of the delta (non-zero DIF) items correctly detected as DIF in each condition. On the other hand, “any-pairs” index refers to the probability of identifying at least one false null hypothesis in a set of tests and was computed as percentages of the simulated data sets in which at least one true DIF item was correctly detected. Because Bonferroni and Hochberg adjustments were employed to evaluate Type I error rates, the significance levels derived from them as well as the nominal alpha, used by the unprotected approach, were adopted to calculate the two kinds of statistical power.

Statistical power of the MIRID DIF in the eight conditions is displayed in Table 13. As shown in Table 12, PCER in smaller delta conditions had adequate control and so did EWER apart from the unprotected ones. In larger delta conditions, PCER under Bonferroni or Hochberg methods had acceptable level of control. Correspondingly, the underlined values in Table 13 represent disregarded power where in conditions or methods necessary Type I error level was not maintained. Power comparisons were made only for the valid values not underlined. In the top

half of the table, larger sample increased effective power somewhat but the level remained low. For smaller sample size, the power became as weak as almost nonexistent, particularly for per-pair index. In the bottom half, the power of hypothesiswise comparisons (per-pair index) were strong in conditions of larger sample and larger delta with the Bonferroni or the Hochberg adjustment.

Table 13

Power of the MIRID DIF over 500 Replications

		# False H0	Per-pair Index			Any-Pairs Index			
			Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg	
delta=-.2	N=250*2	Impact =0	4	0.121	0.011	0.011	<u>0.424</u>	0.044	0.044
		Impact =-.7	4	0.123	0.007	0.007	<u>0.422</u>	0.026	0.026
	N=1500*2	Impact =0	4	0.505	0.119	0.120	<u>0.956</u>	0.420	0.420
		Impact =-.7	4	0.483	0.115	0.115	<u>0.944</u>	0.408	0.408
delta=.7	N=250*2	Impact =0	4	<u>0.805</u>	0.391	0.398	<u>0.880</u>	<u>0.880</u>	<u>0.904</u>
		Impact =-.7	4	0.754	0.341	0.347	<u>1.000</u>	<u>0.852</u>	<u>0.852</u>
	N=1500*2	Impact =0	4	<u>1.000</u>	1.000	1.000	<u>1.000</u>	<u>1.000</u>	<u>1.000</u>
		Impact =-.7	4	<u>1.000</u>	0.987	0.989	<u>1.000</u>	<u>1.000</u>	<u>1.000</u>

Note: Underlined values represent conditions where Type I error rates were either deflated or inflated according to Bradley's criterion.

The group DIF modeling with the MIRID presented a different outlook. The Type I error rates of the MIRID DFFc are presented in Table 14. This model is designed to capture differential functioning at the item group (component) level and only one of the three components was simulated to have DIF. The unprotected, Bonferroni-, and Hochberg-adjusted

PCER were well controlled in all eight conditions. The Hochberg procedure produced PCER at around or slightly above the lower boundary of Bradley's liberal criterion (.025), which were expectedly higher than those deflated rates from the Bonferroni correction. The unprotected EWER were high but the adjusted rates in all eight conditions were all well controlled. Overall, the MIRID DFFc demonstrated significant improvement over the MIRID DIF in this aspect.

Table 14

Type I Error Rates for the MIRID DFFc over 500 Replications

		# True H0	Per-comparison %			Experimentwise %			
			Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg	
delta=-.2	N=250*2	Impact =0	2	0.045	0.018	0.020	<u>0.090</u>	0.036	0.040
		Impact =-.7	2	0.053	0.017	0.030	<u>0.102</u>	0.034	0.056
	N=1500*2	Impact =0	2	0.060	0.020	0.026	<u>0.118</u>	0.040	0.050
		Impact =-.7	2	0.047	0.018	0.028	<u>0.094</u>	0.036	0.056
delta=.7	N=250*2	Impact =0	2	0.061	0.020	0.029	<u>0.122</u>	0.040	0.058
		Impact =-.7	2	0.049	0.014	0.026	<u>0.098</u>	0.028	0.052
	N=1500*2	Impact =0	2	0.048	0.017	0.024	<u>0.094</u>	0.034	0.046
		Impact =-.7	2	0.048	0.017	0.025	<u>0.092</u>	0.032	0.046

Note: Underlined values represent either deflated or inflated Type I error rates according to Bradley's criterion.

The Hochberg adjusted EWER by sample size are presented in Figure 14. On average, the results from both graphs are roughly equal, suggesting no obvious effects from sample size. Also, the effects from delta size were mixed since larger DIF did not necessarily lead to higher error rates. The effects from group difference were not clear either as since the interaction

between this factor and delta size were not consistent between smaller and larger sample sizes. These findings confirmed the recovery results in Figure 9 and 10 (dotted lines) where the non-zero DIF parameters in the DFFc model were shown to be well estimated with negligible bias in all conditions.

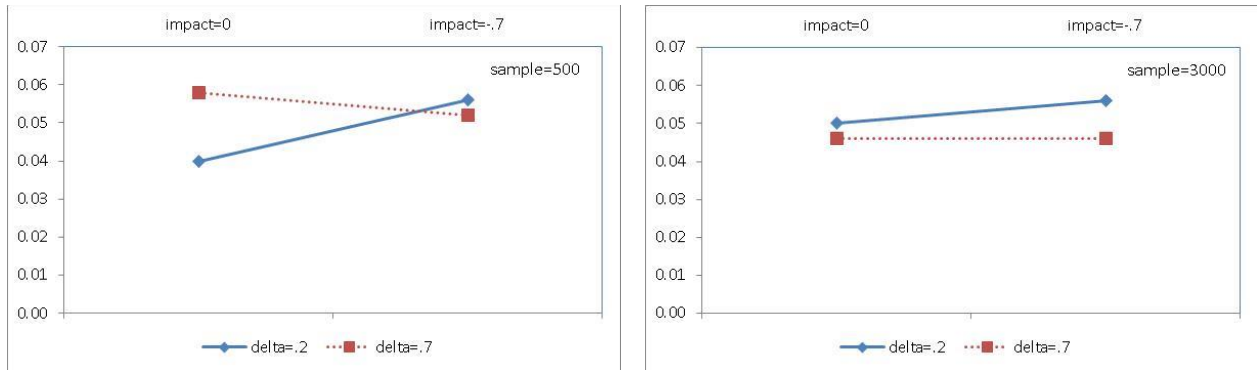


Figure 14. The MIRID DFFc model experimentwise Type I error rates after Hochberg adjustment by sample size

Statistical power of the MIRID DFFc in the eight conditions is given in Table 15. Because Table 14 suggests that, other than the unprotected EWER, there was adequate Type I error control, power comparisons were made for those valid values not associated with unprotected EWER. In smaller delta and smaller sample conditions, the per-pair power was adequate with the unprotected approach ($>.8$) and the two adjustment procedures lowered the power to slightly below the adequate level. For the remaining conditions or methods, there was 100% probability of rejecting the false hypothesis. Note there was only one false H_0 here; thus the Bonferroni and Hochberg procedures gave identical outcome.

In Table 16, Type I error rates of the MIRID DFFm model look very similar to those in Table 14, where once again unprotected EWCR were the only offenders. With two item-family

differential effects simulated, there were eight true null hypotheses to test. The PCER from the unprotected test were similarly well controlled to those under the DFFc model; however, the two adjustment procedures played a very conservative role in producing even lower error rates than those in the DFFc model in part due to more true null hypotheses to evaluate. For EWER, the unprotected approach resulted in inflated error rates expectedly. With the two adjustment approaches, the Type I rates fell within Bradley’s boundaries across all eight conditions.

Table 15
Power of the MIRID DFFc over 500 Replications

		# False H0	Per-pair Index			Any-Pair Index			
			Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg	
delta=-.2	N=250*2	Impact =0	1	0.850	0.734	0.734	<u>0.850</u>	0.734	0.734
		Impact =-.7	1	0.806	0.684	0.688	<u>0.806</u>	0.684	0.688
	N=1500*2	Impact =0	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
		Impact =-.7	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
delta=.7	N=250*2	Impact =0	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
		Impact =-.7	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
	N=1500*2	Impact =0	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
		Impact =-.7	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000

Note: Underlined values represent conditions or procedures where Type I error rates were not well maintained according to Bradley’s criterion.

The Hochberg EWCR for the DFFm model were graphed by sample size in Figure 15. On average there was almost no difference between the two levels of sample sizes, especially for

smaller delta. The condition of larger delta, smaller sample, and no impact resulted in the largest error rate, .058, well within Bradley's range. Note that Figure 15 resembles Figure 14 closely both in values and in where the interactions occurred. These findings confirmed the recovery results in Figure 9 (dashed lines) where the non-zero DIF parameters in the DFFm model were shown to be well estimated with negligible bias in all conditions.

Table 16

Type I Error Rates for the MIRID DFFm Model over 500 Replications

			# True H0	Per-comparison %			Experimentwise %		
				Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg
delta=-2	N=250*2	Impact =0	8	0.054	0.005	0.006	<u>0.318</u>	0.034	0.038
		Impact =-.7	8	0.048	0.006	0.007	<u>0.282</u>	0.044	0.050
	N=1500*2	Impact =0	8	0.050	0.005	0.005	<u>0.288</u>	0.034	0.036
		Impact =-.7	8	0.044	0.005	0.006	<u>0.278</u>	0.038	0.048
delta=.7	N=250*2	Impact =0	8	0.052	0.007	0.008	<u>0.318</u>	0.052	0.058
		Impact =-.7	8	0.049	0.006	0.007	<u>0.290</u>	0.042	0.048
	N=1500*2	Impact =0	8	0.048	0.004	0.005	<u>0.308</u>	0.028	0.036
		Impact =-.7	8	0.056	0.006	0.007	<u>0.348</u>	0.036	0.044

Note: Underlined values represent inflated Type I error rates according to Bradley's criterion.

The power of the MIRID DFFm can be found in Table 17, where values not underlined represent valid power. Given the Type I error results in Table 16, power analysis applied to the three per-pair indices and the two adjusted any-pair power indices. Similar to the DFFc model (Table 15), only smaller delta/smaller sample conditions saw less than perfect power in rejecting

the two false null hypotheses. The small number of false null hypotheses might have again contributed to the high levels of power. The unprotected approach generated adequate per-pair power ($> .8$) before any adjustments. The any-pair index was also acceptable with the Bonferroni and Hochberg adjustments. With such high power in most conditions, the effect from the impact factor was not clear. These power results were consistent with good recovery of the two item family delta parameters (Table 10 and Figure 5).

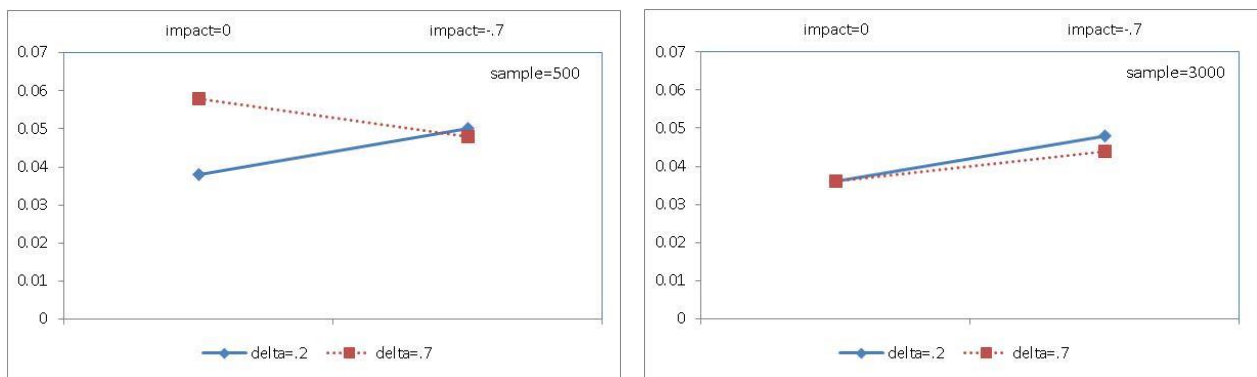


Figure 15. The MIRID DFFm model experimentwise Type I error rates after Hochberg adjustment by sample size

For the last group level DIF model, the MIRID DWF, the Type I error control is given in Table 18. Since only one of the three component weights was simulated to have differential effect, there were only two true null hypotheses. The PCEW were well controlled using the unprotected approach and were conservative with the two adjustments. For EWER, although the unprotected procedure was again not controlled, the Hochberg and Bonferroni rates were all within the healthy range. Notably, the unprotected EWER were better controlled than the DFFc (Table 14) and DFFm (Table 16) models.

These Hochberg rates are split by sample size and displayed graphically in Figure 16. On the right plot, what is remarkable was the high error rates in larger sample and larger delta conditions. On the left, larger delta did not produce higher error rates in smaller sample conditions. In fact, they were at a level very similar to the rates in the smaller delta conditions. The interaction between group difference and delta magnitude was not consistent between the two sample sizes.

Table 17
Power of the MIRID DFFm over 500 Replications

		# False H0	Per-pair Index			Any-Pair Index			
			Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg	
delta=-.2	N=250*2	Impact =0	2	0.848	0.593	0.600	<u>0.958</u>	0.784	0.784
		Impact =-.7	2	0.858	0.578	0.586	<u>0.958</u>	0.780	0.780
	N=1500*2	Impact =0	2	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
		Impact =-.7	2	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
delta=.7	N=250*2	Impact =0	2	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
		Impact =-.7	2	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
	N=1500*2	Impact =0	2	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
		Impact =-.7	2	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000

Note: Underlined values represent conditions where Type I error rates were inflated according to Bradley's criterion.

In correspondence to the obtained Type I error rates, only the unprotected any-pair power was not valid and consequently, power comparisons were made to per-pair index for all three

procedures and any-pair index for the two adjustments. In Table 19, there is very good power for all these procedures, even for smaller sample and smaller delta conditions. There was virtually no difference between the Bonferroni and Hochberg procedures in both types of power. Again, it was impossible to observe the effect of impact given the perfect power in where the alpha was maintained.

Table 18

Type I Error Rates for the MIRID DWF over 500 Replications

		# True H0	Per-comparison %			Experimentwise %			
			Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg	
delta=.2	N=250*2	Impact =0	2	0.050	0.014	0.021	<u>0.098</u>	0.028	0.040
		Impact =-.7	2	0.048	0.014	0.021	<u>0.092</u>	0.028	0.038
	N=1500*2	Impact =0	2	0.044	0.013	0.020	<u>0.084</u>	0.024	0.036
		Impact =-.7	2	0.043	0.014	0.021	<u>0.084</u>	0.028	0.040
delta=.7	N=250*2	Impact =0	2	0.039	0.013	0.017	<u>0.076</u>	0.024	0.032
		Impact =-.7	2	0.045	0.010	0.022	<u>0.084</u>	0.020	0.038
	N=1500*2	Impact =0	2	0.050	0.020	0.028	<u>0.096</u>	0.038	0.052
		Impact =-.7	2	0.048	0.020	0.029	<u>0.096</u>	0.040	0.058

Note: Underlined values represent inflated Type I error rates according to Bradley's criterion.

Additionally, influences of the design factors on the Type I Error control was investigated via a factorial ANOVA analysis with the generalized eta square used as the effect size to determine the impact. PCER and EWER were the dependent variables in separate analyses which included model, delta size, sample size, group difference, and their interactions as independent

variables. Cohen's (1988) moderate effect size of .059 was used as a cutoff value to indicate the significant level.

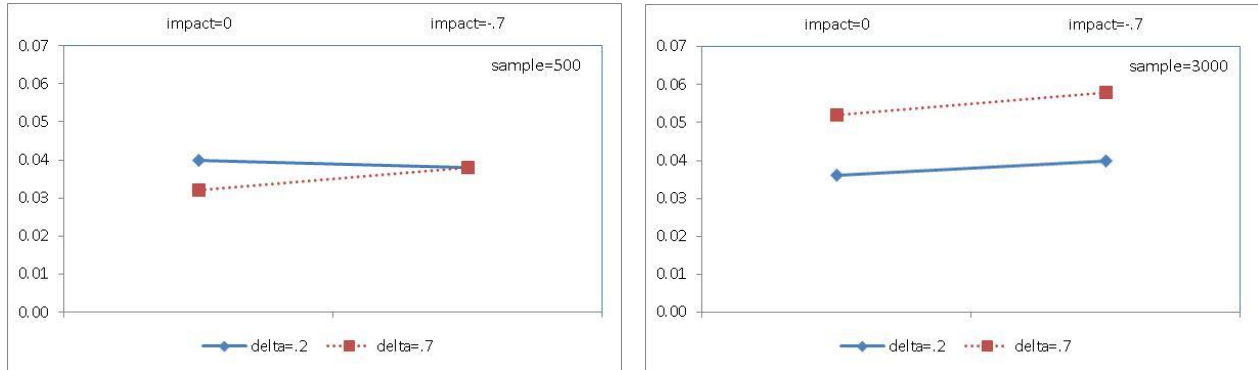


Figure 16. The MIRID DWF model experimentwise Type I error rates after Hochberg adjustment by sample size

When per-comparison Type I error was the dependent variable, among all design factors, model (.294), interaction between model and delta (.212), interaction of model by sample size (.171), and delta (DIF magnitude) (.074) had significant effects across all hypotheses tests. The factor of group difference had negligible impact. For analysis of experimentwise Type I error, the same four design factors or factor interactions were significant across all tests only with slightly different effect sizes: model (.235), interaction between model and delta (.228), interaction of model by sample size (.170), and delta (DIF magnitude) (.082).

To illustrate the significant impact from the factor of model, Table 19 and 20 list the per-comparison and experimentwise Type I error rates after Hochberg adjustments from all conditions, respectively. (In the context of multiple testing of hypotheses, experimentwise errors were more of the concern but the two kinds of Type I error rates look much alike.) These tables show that the DIF model produced tremendous error rates in larger delta conditions (underlined),

setting this model apart from the other three. On the other hand, the Type I errors for the DFFc model were greater on average than the DFFm and DWF models and the DWF model had the lowest rates.

Table 19

Power of the MIRID DWF over 500 Replications

		# False H0	Per-pair Index			Any-Pair Index			
			Un- protect.	Bonfer- roni	Hoch- berg	Un- protect.	Bonfer- roni	Hoch- berg	
delta=-.2	N=250*2	Impact =0	1	0.994	0.978	0.978	<u>0.994</u>	0.978	0.978
		Impact =-.7	1	0.892	0.788	0.790	<u>0.892</u>	0.788	0.790
	N=1500* 2	Impact =0	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
		Impact =-.7	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
delta=.7	N=250*2	Impact =0	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
		Impact =-.7	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
	N=1500* 2	Impact =0	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000
		Impact =-.7	1	1.000	1.000	1.000	<u>1.000</u>	1.000	1.000

Note: Underlined values represent conditions or procedures where Type I error rates were inflated according to Bradley's criterion.

In summary, the MIRID DFFc, DFFm, and DWF models had adequate hypothesiswise Type I error control across all data generation conditions; as for the experimentwise Type I errors, these models maintained the alpha level through the Bonferroni and Hochberg adjustments. When Type I errors were well controlled, all three models exhibited perfect or near perfect per-pair and any-pairs statistical power in larger DIF conditions and smaller DIF but

larger sample conditions. Even when sample and delta magnitude were small, there was decent per-pair and any-powers power with the former greater than .80 and the latter more than .68 for all three group-level DIF models. Such level of Type I error control and power in detection was

Table 20

Hypothesiswise Type I Error Rates for the Four Proposed MIRID Models

			MIRID DIF	MIRID DFFc	MIRID DFFm	MIRID DWF
delta=.2	N=250*2	Impact=0	0.051	0.045	0.054	0.050
		Impact=-.7	0.049	0.053	0.048	0.048
	N=1500*2	Impact=0	0.065	0.060	0.050	0.044
		Impact=-.7	0.062	0.047	0.044	0.043
delta=.7	N=250*2	Impact=0	<u>0.079</u>	0.061	0.052	0.039
		Impact=-.7	0.069	0.049	0.049	0.045
	N=1500*2	Impact=0	<u>0.253</u>	0.048	0.048	0.050
		Impact=-.7	<u>0.192</u>	0.048	0.056	0.048

Note: Underlined values represent either deflated or inflated Type I error rates according to Bradley's criterion.

Table 21

Experimentwise Type I Error Rates after Hochberg Adjustment for the Four Proposed MIRID Models

			MIRID DIF	MIRID DFFc	MIRID DFFm	MIRID DWF
delta=.2	N=250*2	Impact=0	0.040	0.040	0.038	0.040
		Impact=-.7	0.038	0.056	0.050	0.038
	N=1500*2	Impact=0	0.058	0.050	0.036	0.036
		Impact=-.7	0.050	0.056	0.048	0.040
delta=.7	N=250*2	Impact=0	<u>0.110</u>	0.058	0.058	0.032
		Impact=-.7	<u>0.088</u>	0.052	0.048	0.038
	N=1500*2	Impact=0	<u>0.702</u>	0.046	0.036	0.052
		Impact=-.7	<u>0.522</u>	0.046	0.044	0.058

Note: Underlined values represent either deflated or inflated Type I error rates according to Bradley's criterion.

related to the adequate recovery of both the delta parameters and the zero-value DIF parameters. As for the effects of the design factors on Type I error control, the models themselves, their interaction with DIF size and sample size, as well as DIF size itself, were significant. The influence of group ability difference was not apparent.

On the other hand, the MIRID DIF model had acceptable per-comparison and experimentwise Type I error rates only in smaller delta conditions whereas larger sample size increased errors. In larger delta conditions, neither PCER nor EWER was well maintained even after Hochberg and Bonferroni corrections when EWER were higher than PCER. For the DIF model, the group difference factor led to lower Type I errors in all conditions constructed. In conditions with Type I errors under control, the statistical power was weak. Such outcome was related to the underestimation of the delta and non-zero parameters of the MIRID DIF model as depicted in Figure 1, 2, 7, and 8.

Results from Fitting the Mismatched Differential Functioning Models

The second phase of the research sought to further study the characteristics of the four proposed differential functioning models by examining their performances in various scenarios; in particular, when they were fitted to data generated using other models (a.k.a. when they were the “wrong model” for the data). This part of the study was akin to research on “model misspecification” that investigates the potential impact from applying the wrong model. Performances were evaluated by examining the robustness of the estimation of the model parameters and how severe the false detection rates were associated with each data simulation condition.

False Detection Rates of the Mismatched MIRID Differential Functioning Models

When a MIRID differential functioning model is fitted to data with a different source of differential effects, it attempts to estimate its own delta parameters regardless of the true source of DIF. For example, when the MIRID DIF model is applied to the DWF data, the 30 individual item DIF parameters are estimated from the data where there is a degree of differential functioning from one of the component weights. Naturally, the true values of these DIF parameters are zero and any non-zero estimates reflect deviation from the true value whose statistical significance indicate false detection errors.

Figures 17 and 18 demonstrate the bias and RMSE of the estimated DIF parameters when the DIF model fitted to the four different differential functioning data, the DIF data serving as the reference in the leftmost column. The dashed lines connect the average bias values and RMSEs in larger delta conditions and the solid lines represent the bias and RMSEs from smaller delta conditions. For the DFFc and DFFm data, the item DIF estimates in larger delta conditions were severely inflated by the presence of item group DIF, which is depicted by the high rise or the spikes of the lines that represent those DIF parameters associated with items within the item group (component or family). Figure 17 shows that the items where no DIF effect was simulated were under-estimated. The estimation error was also present in smaller delta conditions but to a much less extent. All the three columns that show mismatched models exhibit higher bias and RMSEs than when the correct model was applied. Especially, the most inflated estimates were with the DFFm model where the highly spiked part of the dashed lines represents those DIF parameters associated with items that belong with the two item families for which differential effects were simulated. It shows that the MIRID DIF model was unable to identify the true source of DIF and “assigns” in estimation the differential effects to related individual items. In

other words, when all or the most of a group of items show significant delta when applying the DIF model it signals the presence of differential functioning effects from the item group.

However, this pattern was not found with the DWF data in the rightmost column. Although that the second component weight was simulated to have differential effects, there is no indication that the individual DIF parameters in the second component had increased bias and RMSEs. The smaller delta conditions (the solid lines) did not produce considerable level of bias or RMSEs across all conditions. On the other hand, the larger delta conditions resulted in more conspicuous bias and RMSEs for certain items but there was no clear pattern of which items had their DIF parameters estimated with huge error.

The bias and RMSEs of all three mismatched models were greater than when the correct model was applied. Of the design factors, delta magnitude increased RMSEs while sample size reduced them. The effects from group difference were not clear.

As a consequence of the conspicuous RMSEs, false detection rates became very high for the MIRID DIF model when fitted to the three mismatched models. For example, as shown in Table 21, when the DIF model was used to analyze the DFFc data, in the condition of no impact, larger sample, and larger delta, the per-experiment error rates after the Hochberg adjustment were 100%. The unprotected error rates were also high although they decreased with the help of adjustment in smaller delta and smaller sample conditions, where the Hochberg experimentwise error rates were at .15 and .16, three times the nominal level.

In Table 22 the false detection rates with the DFFm data are presented. Note that the DFFm data had lower per-comparison rates in larger delta conditions than in the DFFc data. It can be explained that fewer DIF parameters had inflated RMSEs with the DFFm data (six versus

10) although their RMSEs were greater. The experimentwise error rates from fitting the DIF model to the DFFm data were larger than those from the DFFc data in smaller delta conditions.

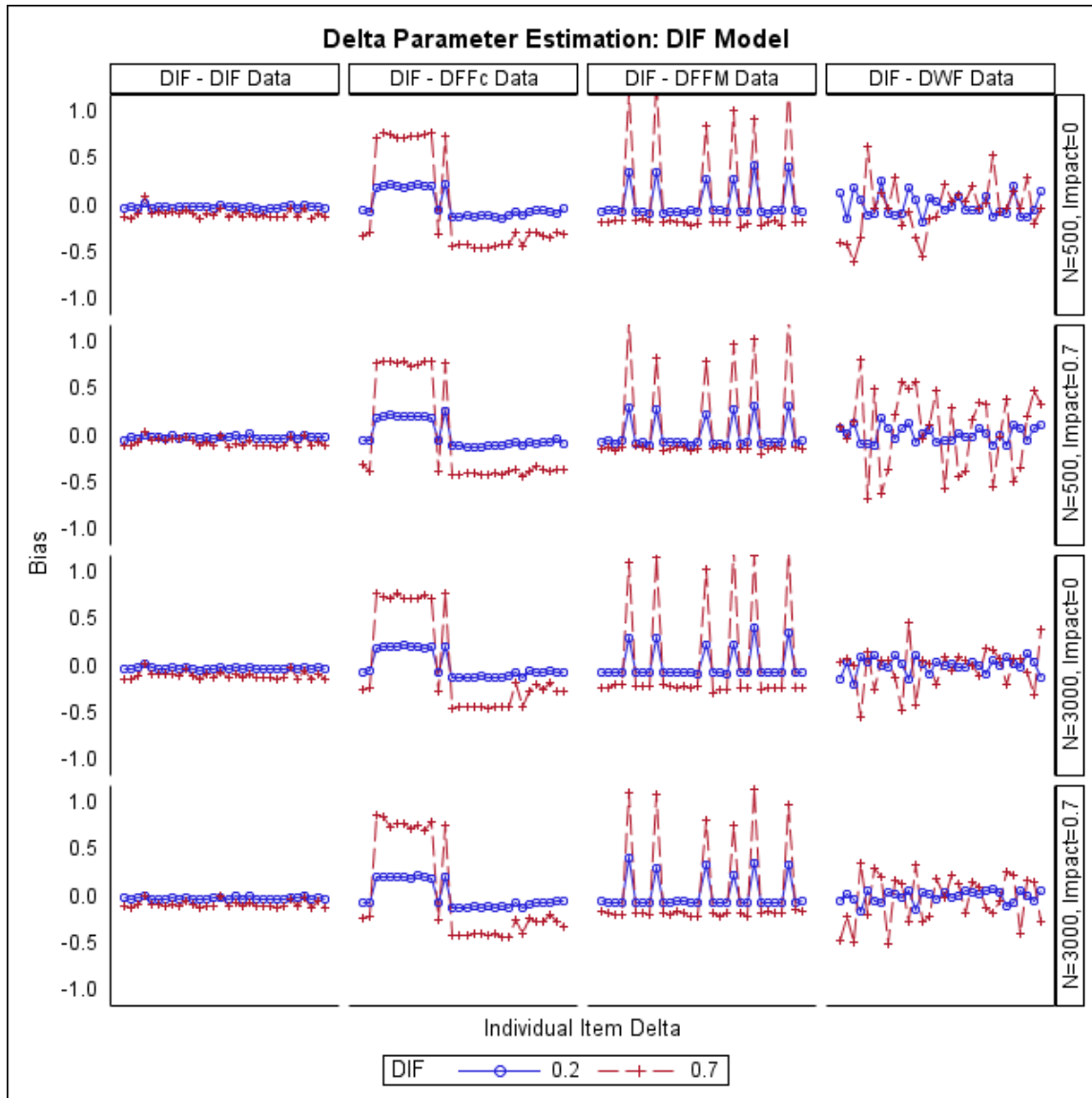


Figure 17. Bias of the 30 estimated DIF parameters of the MIRID DIF model when fitted to data with different sources of differential functioning

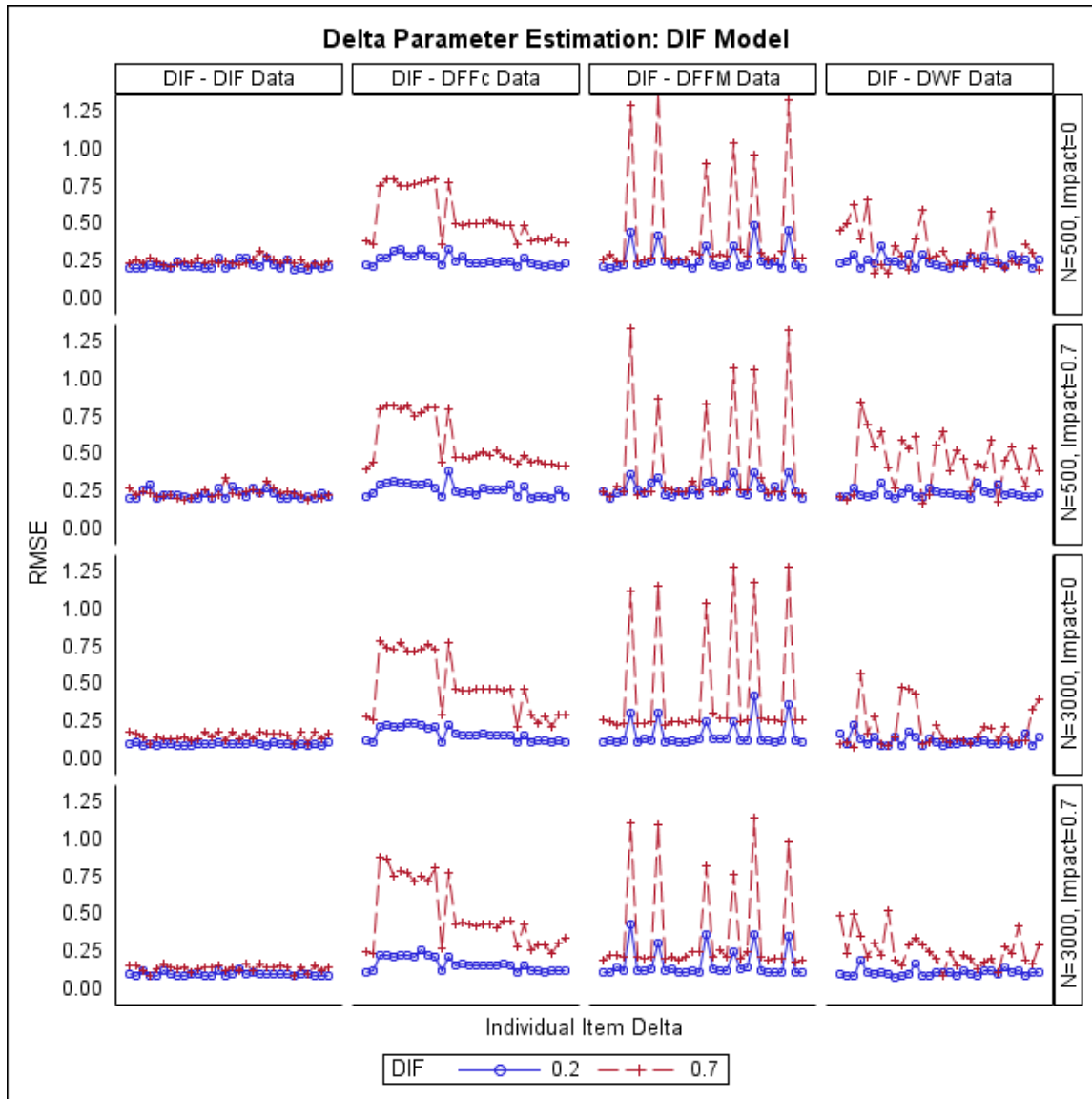


Figure 18. RMSE of the 30 estimated DIF parameters of the MIRID DIF model when fitted to data with different sources of differential functioning

For example, in smaller delta and smaller sample conditions, the rates with the DFFm data were .35 and .28 as compared with .15 and .16 from the DFFc data.

In Table 23, it can be seen that the per-comparison false detection rate with the DWF data were lower than the other two data sets on average due to fewer DIF parameters being impacted.

The Bonferroni and Hochberg procedures led to very deflated false detection rates in smaller

delta conditions. The experimentwise error rates were also lower than the other two DIF sources but still at very high levels.

Table 22

False Detection Rates when the MIRID DIF Model Was Applied to the DFFc Data

		# Parms	Per-comparison %			Experimentwise %		
			Un- protect.	Bonfer- roni	Hoch- berg	Un- protect.	Bonfer- roni	Hoch- berg
delta=.2	N=250*2	Impact =0	0.100	0.010	0.010	0.930	0.160	0.160
		Impact =-.7	0.100	0.010	0.010	0.900	0.150	0.150
	N=1500*	Impact =0	0.360	0.090	0.090	1.000	0.910	0.910
	2	Impact =-.7	0.350	0.090	0.090	1.000	0.890	0.890
delta=.7	N=250*2	Impact =0	0.610	0.310	0.330	1.000	1.000	1.000
		Impact =-.7	0.600	0.330	0.350	1.000	1.000	1.000
	N=1500*	Impact =0	0.930	0.800	0.890	1.000	1.000	1.000
	2	Impact =-.7	0.940	0.780	0.890	1.000	1.000	1.000

The outcome from fitting the MIRID DFFc model to other types of DIF data is discussed below. The plots look stable because there were only three zero-value DIF parameters to test. Compared to the recovery results from the matched data on the leftmost column of Figure 19 and 20, the mismatched data exhibited greater bias and RMSEs but all at acceptable levels in general. The most significant bias as shown in Figure 19 occurred to DFFm and DWF data in larger delta and zero impact conditions. The bias and RMSEs with the DFFm and DWF data were slightly greater than those with the DIF data. Across DIF sources and conditions, larger sample size reduced bias and RMSEs and larger delta size increased the levels of bias and RMSE. The

impact of larger delta was the most obvious with the DFFm data in the scenario of larger sample and no group difference.

Table 23

False Detection Rates when the MIRID DIF Model Was Applied to the DFFm Data

		# Parms	Per-comparison %			Experimentwise %		
			Un- protect.	Bonfer- roni	Hoch- berg	Un- protect.	Bonfer- roni	Hoch- berg
delta=.2	N=250*2	Impact =0	0.120	0.010	0.010	0.970	0.350	0.350
		Impact =-.7	0.110	0.010	0.010	0.960	0.280	0.280
	N=1500* 2	Impact =0	0.300	0.130	0.140	1.000	0.910	0.910
		Impact =-.7	0.280	0.130	0.130	1.000	0.890	0.890
delta=.7	N=250*2	Impact =0	0.310	0.180	0.180	1.000	1.000	1.000
		Impact =-.7	0.270	0.130	0.140	1.000	1.000	1.000
	N=1500* 2	Impact =0	0.840	0.500	0.580	1.000	1.000	1.000
		Impact =-.7	0.680	0.350	0.380	1.000	1.000	1.000

Table 24 shows the false detection rates when the DFFc model was fitted to the DIF data. When delta = .2, the error rates were all within the Bradley's acceptable range, including both per-comparison and experimentwise error rates. When delta = .7, only smaller samples saw acceptable per-comparison error rates after adjustments and in the other conditions they all went beyond the upper boundary (.75). These results indicate that very unlikely a significant component differential effect would be found when fitting the DFFc model to data with small individual item DIF. This finding is consistent with the plots in Figure 18.

Table 24

False Detection Rates when the MIRID DIF Model Was Applied to the DWF Data

		# Parms	Per-comparison %			Experimentwise %			
			Un- protect.	Bonfer- roni	Hoch- berg	Un- protect.	Bonfer- roni	Hoch- berg	
delta=.2	N=250*2	Impact =0	30	0.090	0.000	0.000	0.930	0.110	0.110
		Impact =-.7	30	0.070	0.000	0.000	0.860	0.050	0.050
	N=1500* 2	Impact =0	30	0.150	0.020	0.020	1.000	0.450	0.450
		Impact =-.7	30	0.110	0.010	0.010	0.980	0.260	0.260
delta=.7	N=250*2	Impact =0	30	0.260	0.080	0.080	1.000	0.920	0.920
		Impact =-.7	30	0.540	0.280	0.300	1.000	1.000	1.000
	N=1500* 2	Impact =0	30	0.420	0.250	0.260	1.000	1.000	1.000
		Impact =-.7	30	0.680	0.420	0.460	1.000	1.000	1.000

The false detection rates became much worse when the DFFc model was fitted to the DFFm data as shown in Table 25. Only the per-comparison error rates in smaller sample conditions were within the Bradley's range after either Bonferroni or Hochberg adjustment. On the other hand, the false detection rates were very high in larger sample conditions even with the adjustment. Interestingly, the non-zero impact factor decreased the error rates for larger delta conditions.

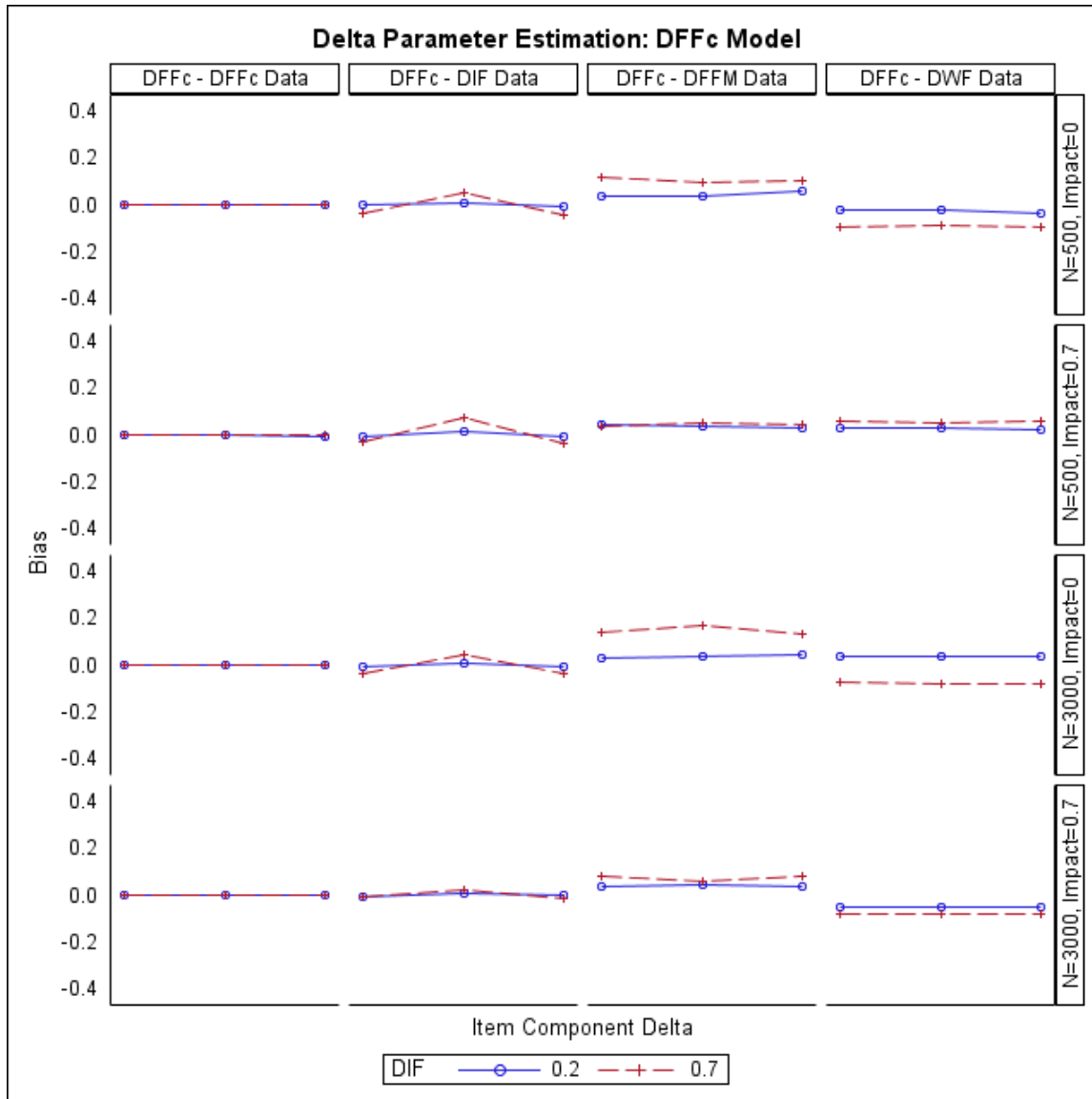


Figure 19. Bias of the three estimated DIF parameters of the MIRID DFFc model when fitted to data with different sources of differential functioning

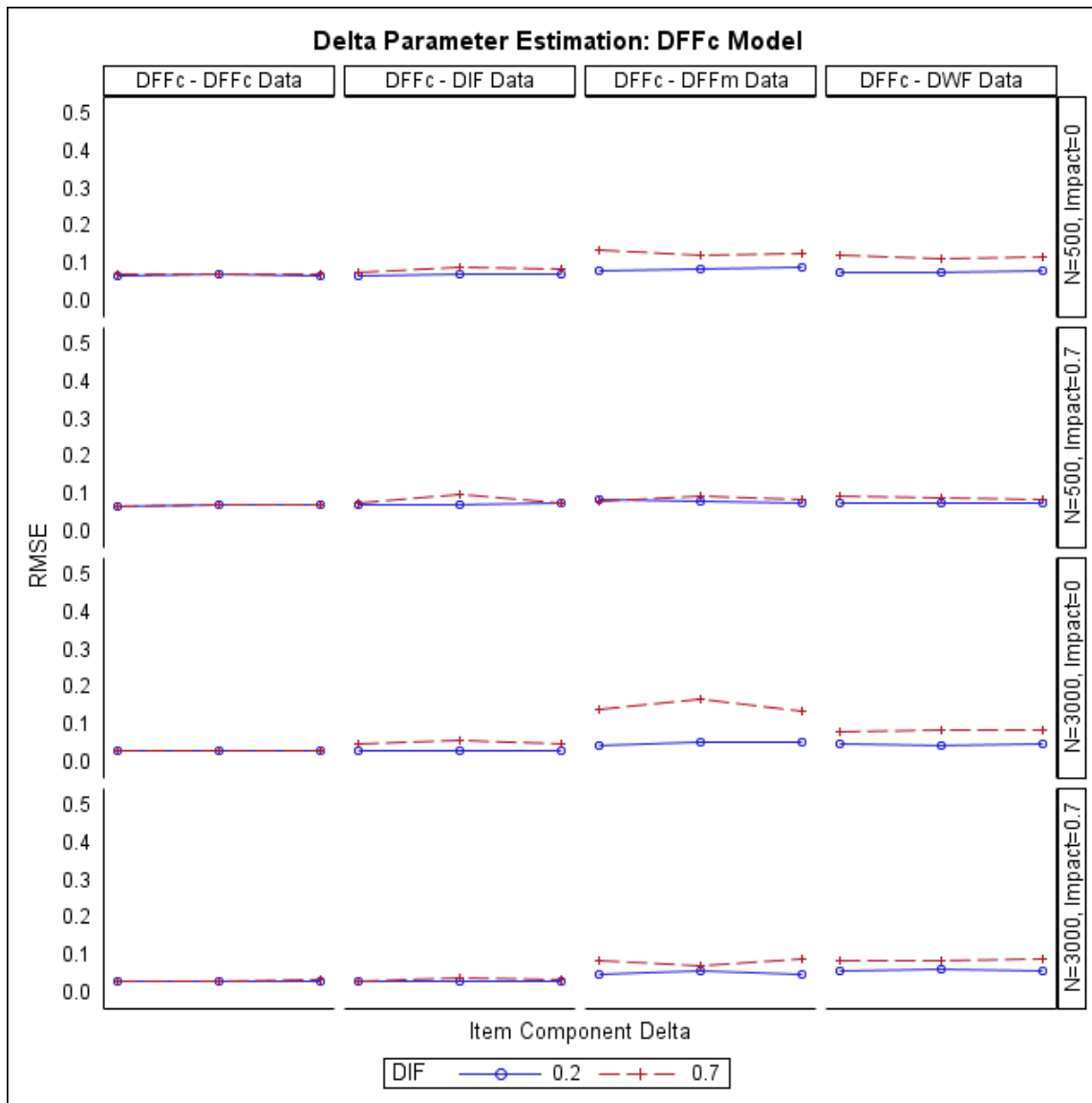


Figure 20. RMSE of the three estimated DIF parameters of the MIRID DFFc model when fitted to data with different sources of differential functioning

When fitted to DWF data simulated with larger delta size (.7), the false detection rates of the DFFc model were inflated across the conditions. Although the per-comparison rates were acceptable in lower delta and smaller sample conditions, the larger delta and larger sample conditions had very high false detection rates where all three DFFc parameters obtained significant non-zero estimates in virtually all replications.

Table 25

False Detection Rates when the MIRID DFFc Model Was Applied to the DIF Data

		# Parms	Per-comparison %			Experimentwise %			
			Un- protect.	Bonfer- roni	Hoch- berg	Un- protect.	Bonfer- roni	Hoch- berg	
delta=.2	N=250*2	Impact =0	3	0.050	0.020	0.020	0.160	0.050	0.050
		Impact =-.7	3	0.050	0.020	0.020	0.150	0.050	0.050
	N=1500*	Impact =0	3	0.060	0.020	0.020	0.160	0.060	0.060
	2	Impact =-.7	3	0.060	0.020	0.020	0.160	0.060	0.060
delta=.7	N=250*2	Impact =0	3	0.100	0.040	0.050	0.270	0.130	0.130
		Impact =-.7	3	0.100	0.040	0.040	0.270	0.120	0.120
	N=1500*	Impact =0	3	0.300	0.180	0.190	0.690	0.450	0.450
	2	Impact =-.7	3	0.080	0.040	0.040	0.230	0.110	0.110

When applied to its own data, delta parameter recovery of the DFFm model was in general acceptable except at the third and seventh parameters where delta effect was simulated. The application of the DFFm model to mismatched data is shown in the three columns on the right in Figure 21 and 22. The pattern repeated that larger samples reduced bias and RMSEs and larger delta increased them. The DIF data caused the smallest RMSEs, which were quite comparable to the matched data in smaller delta conditions (the solid lines). The fluctuating bias values in Figure 21 suggest the difficulty in estimating the family delta parameters due to the presence of individual item DIF. For the DFFc data, the bias and RMSEs were consistently higher, particularly in larger delta conditions. The component differential functioning impacted all the items in the component, which also meant one item in every item family. The estimation

Table 26

False Detection Rates when the MIRID DFFc Model Was Applied to the DFFm Data

		# Parms	Per-comparison %			Experimentwise %			
			Un- protect.	Bonfer- roni	Hoch- berg	Un- protect.	Bonfer- roni	Hoch- berg	
delta=.2	N=250*2	Impact =0	3	0.110	0.050	0.050	0.290	0.130	0.130
		Impact =-.7	3	0.080	0.030	0.030	0.220	0.090	0.090
	N=1500*2	Impact =0	3	0.300	0.170	0.190	0.670	0.420	0.430
		Impact =-.7	3	0.320	0.170	0.190	0.700	0.440	0.440
delta=.7	N=250*2	Impact =0	3	0.340	0.210	0.230	0.740	0.510	0.520
		Impact =-.7	3	0.100	0.050	0.050	0.270	0.140	0.140
	N=1500*2	Impact =0	3	1.000	1.000	1.000	1.000	1.000	1.000
		Impact =-.7	3	0.720	0.560	0.660	0.980	0.920	0.930

process attributed this effect to every item family delta parameter as a result. Regarding the DFW data, there was no easy explanation for the volatility demonstrated by the bias and RMSEs in both delta sizes.

The false detection rates from applying the DFFm model to data of three mismatched DIF sources are shown in Tables 27 to 29. When fitted to the DIF data, smaller delta and smaller sample conditions produced acceptable unprotected per-experiment and Hochberg-adjusted experimentwise error rates (Table 27). For the DFFc and DWF data, these two error rates were inflated even for the same scenarios and were highly inflated in larger delta conditions, which is consistent with the RMSEs in Figure 19.

Table 27

False Detection Rates when the MIRID DFFc Model Was Applied to the DWF Data

		# Parms	Per-comparison %			Experimentwise %			
			Un- protect.	Bonfer -roni	Hoch- berg	Un- protect.	Bonfer -roni	Hoch- berg	
delta=.2	N=250*2	Impact= 0	3	0.070	0.030	0.030	0.210	0.080	0.080
		Impact= -.7	3	0.060	0.020	0.020	0.170	0.060	0.060
	N=1500*2	Impact= 0	3	0.250	0.120	0.140	0.570	0.320	0.330
		Impact= -.7	3	0.450	0.280	0.330	0.840	0.610	0.630
delta=.7	N=250*2	Impact= 0	3	0.270	0.150	0.170	0.620	0.390	0.390
		Impact= -.7	3	0.140	0.050	0.050	0.360	0.130	0.130
	N=1500*2	Impact= 0	3	0.820	0.690	0.790	0.990	0.970	0.980
		Impact= -.7	3	0.820	0.700	0.810	1.000	0.980	0.990

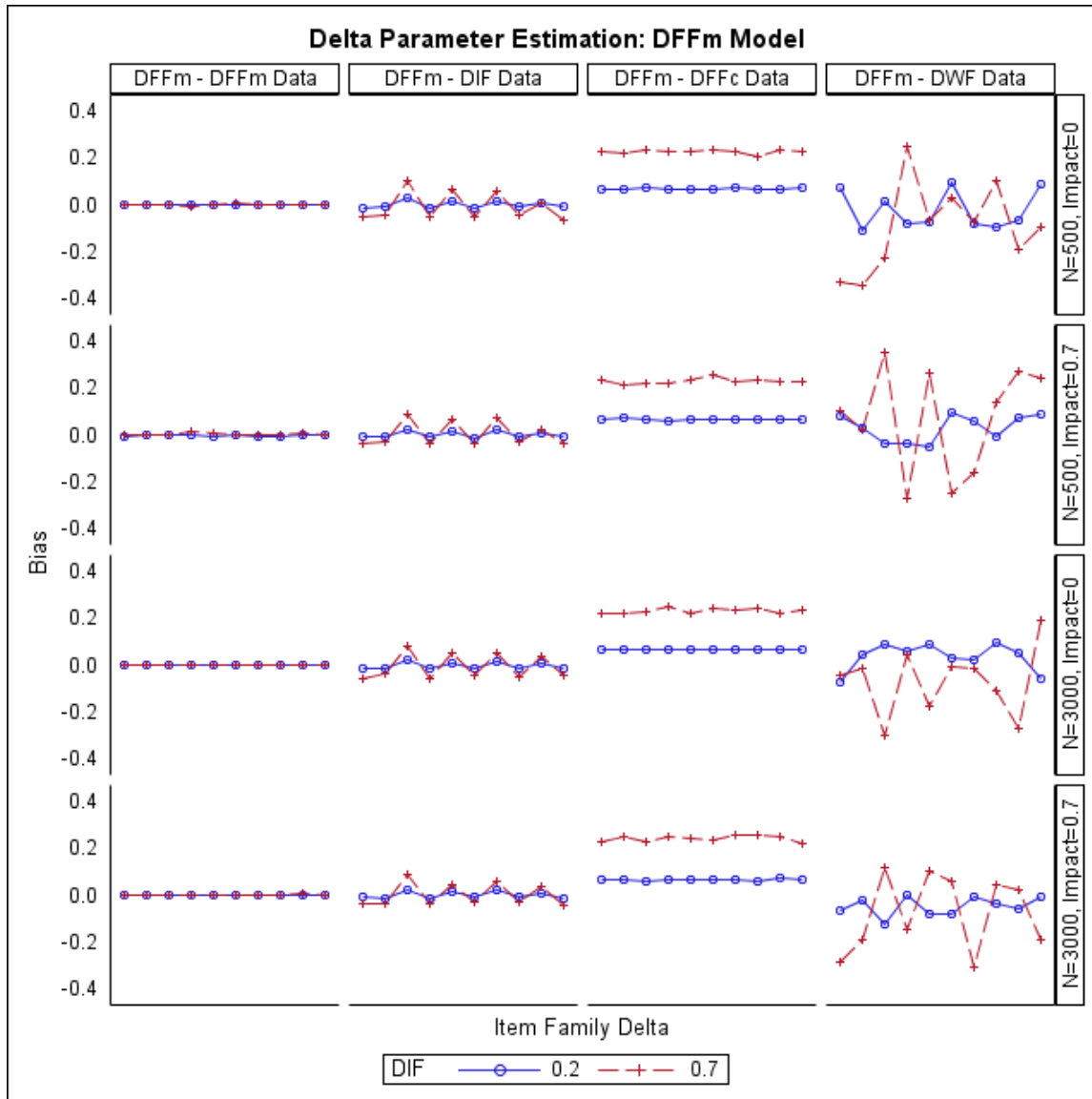


Figure 21. Bias of the ten estimated DIF parameters of the MIRID DFFm model when fitted to data with different sources of differential functioning

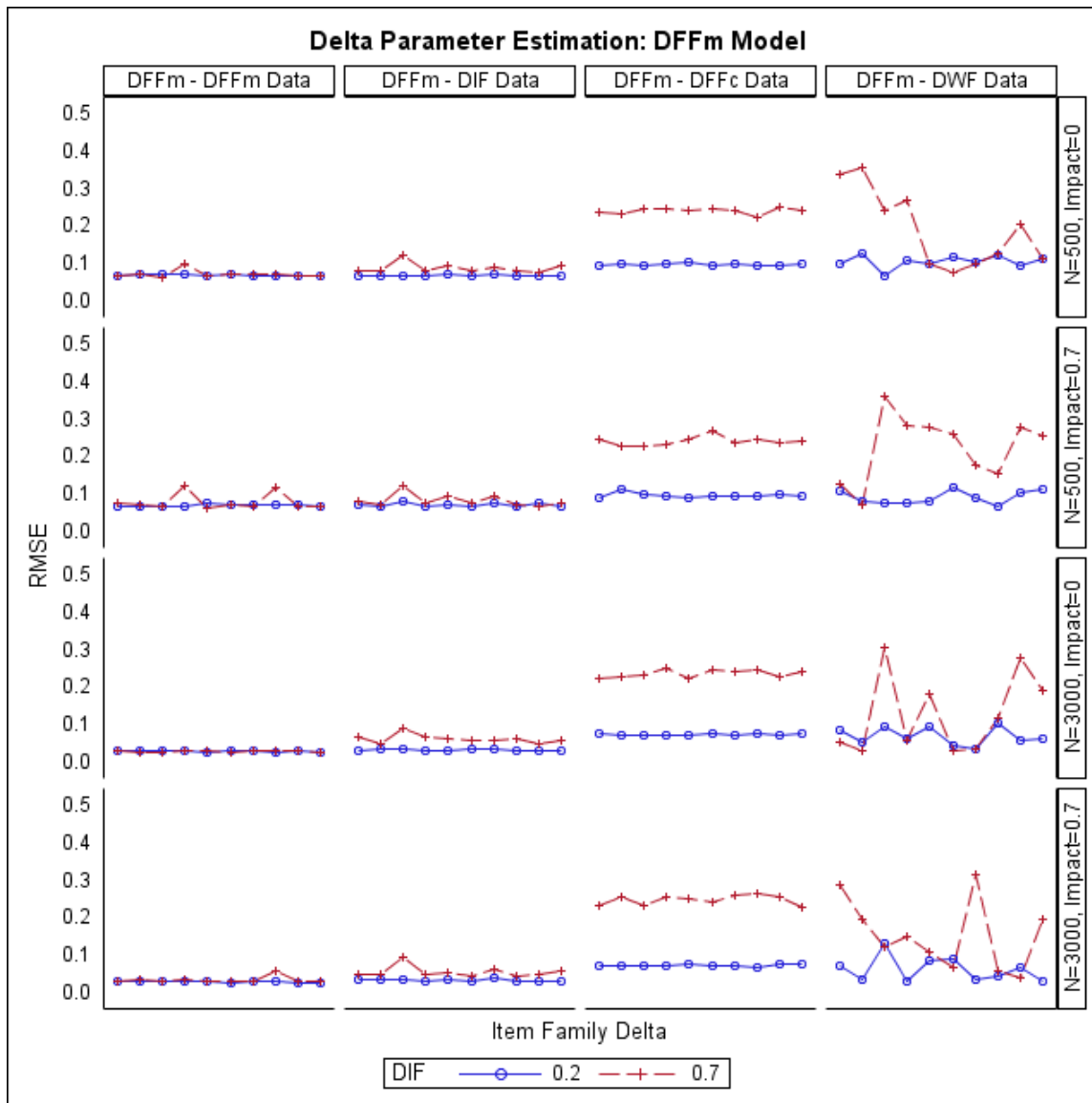


Figure 22. RMSE of the ten estimated DIF parameters of the MIRID DFFm model when fitted to data with different sources of differential functioning

DIF parameter recovery in the DWF model itself was acceptable, especially in larger sample or larger delta conditions (see also Table 10, Figures 11 and 12). The bias and RMSEs from fitting it to the DIF data were very comparable as shown in the second column of Figures 23 and 24. However, when fitted to the DFFc data and the DFFm data, the bias and RMSEs became very high for one or two but not all delta parameters and exhibited a lot of volatility.

Table 28

False Detection Rates when the MIRID DFFm Model Was Applied to the DIF Data

		# Parms	Per-comparison %			Experimentwise %			
			Un- protect.	Bonfer- roni	Hoch- berg	Un- protect.	Bonfer- roni	Hoch- berg	
delta=.2	N=250*2	Impact =0	10	0.050	0.010	0.010	0.390	0.050	0.050
		Impact =-.7	10	0.050	0.010	0.010	0.400	0.070	0.070
	N=1500*2	Impact =0	10	0.090	0.010	0.010	0.560	0.110	0.110
		Impact =-.7	10	0.090	0.010	0.010	0.580	0.110	0.110
delta=.7	N=250*2	Impact =0	10	0.130	0.030	0.030	0.750	0.260	0.260
		Impact =-.7	10	0.100	0.020	0.020	0.630	0.160	0.160
	N=1500*2	Impact =0	10	0.500	0.220	0.250	1.000	0.930	0.930
		Impact =-.7	10	0.370	0.160	0.170	1.000	0.910	0.910

Because differential functioning in a component (DFFc data) or in two item families (DFFm data) impacted items which were related different component weights, the estimation outcome became less than predictable.

When fitted to the DIF data, the DWF model found acceptable unprotected per-comparison and Hochberg-adjusted experimentwise false detection rates in smaller delta conditions (Table 30). With smaller delta, the larger sample with non-zero impact led to error rates above Bradley's range but when delta size was large, smaller sample with non-zero impact had acceptable error rates. In other conditions, fitting the DWF data meant that differential functioning in component weights was likely to be found significant.

Table 29

False Detection Rates when the MIRID DFFm Model Was Applied to the DFFc Data

		# Parms	Per-comparison %			Experiment-wise %			
			Un- protect.	Bonfer- roni	Hoch- berg	Un- protect.	Bonfer- roni	Hoch- berg	
delta=.2	N=250*2	Impact =0	10	0.180	0.040	0.050	0.690	0.280	0.280
		Impact =-.7	10	0.170	0.030	0.030	0.670	0.210	0.210
	N=1500* 2	Impact =0	10	0.720	0.390	0.500	0.990	0.920	0.920
		Impact =-.7	10	0.650	0.330	0.420	0.990	0.900	0.900
delta=.7	N=250*2	Impact =0	10	0.940	0.740	0.890	1.000	1.000	1.000
		Impact =-.7	10	0.930	0.720	0.870	1.000	0.990	0.990
	N=1500* 2	Impact =0	10	1.000	1.000	1.000	1.000	1.000	1.000
		Impact =-.7	10	1.000	1.000	1.000	1.000	1.000	1.000

When fitted to the DFFc data, only conditions with acceptable unprotected per-comparison and adjusted experimentwise false detection rates were smaller delta and smaller sample (see Table 31). Elsewhere the error rates were high, particularly in larger delta conditions. With the DFFm data, the DWF model obtained high per-comparison and adjusted experimentwise false detection rates; only adjusted per-comparison error rates were controlled in conditions of smaller delta and smaller sample. Similar results were found when the DWF model was fitted to the DWF data (Table 32). For all three type of data, larger sample aggravated false detection rates even in smaller delta conditions.

Table 30

False Detection Rates when the MIRID DFFm Model Was Applied to the DWF Data

		# Parms	Per-comparison %			Experiment-wise %			
			Un- protect.	Bonfer- roni	Hoch- berg	Un- protect.	Bonfer- roni	Hoch- berg	
delta=.2	N=250*2	Impact =0	10	0.220	0.050	0.060	0.920	0.400	0.400
		Impact =-.7	10	0.140	0.030	0.030	0.740	0.250	0.250
	N=1500*2	Impact =0	10	0.580	0.340	0.390	1.000	1.000	1.000
		Impact =-.7	10	0.460	0.280	0.310	1.000	0.990	0.990
delta=.7	N=250*2	Impact =0	10	0.560	0.390	0.420	1.000	1.000	1.000
		Impact =-.7	10	0.740	0.580	0.640	1.000	1.000	1.000
	N=1500*2	Impact =0	10	0.600	0.520	0.540	1.000	1.000	1.000
		Impact =-.7	10	0.800	0.710	0.770	1.000	1.000	1.000

Non-DIF Parameter Recovery of the Mismatched Models

In addition to the delta and non-zero DIF parameters, a MIRID differential functioning model estimates its non-DIF parameters: locations of the component items, intercept and component weights, group difference, and population variance, even when it is fitted to data of mismatched DIF source. Since the real source of differential effects remains unknown in empirical settings, understanding whether estimates of other parameters are by and large accurate despite the distortion from the DIF effects helps to evaluate the utility of the model. This section examines recovery of non-DIF parameters from fitting the proposed models to data of “incorrect”

DIF sources by using bias as the main evaluative measure because bias provides information on over- or under-estimation of the parameters. RMSEs were also provided as supplement evidence.

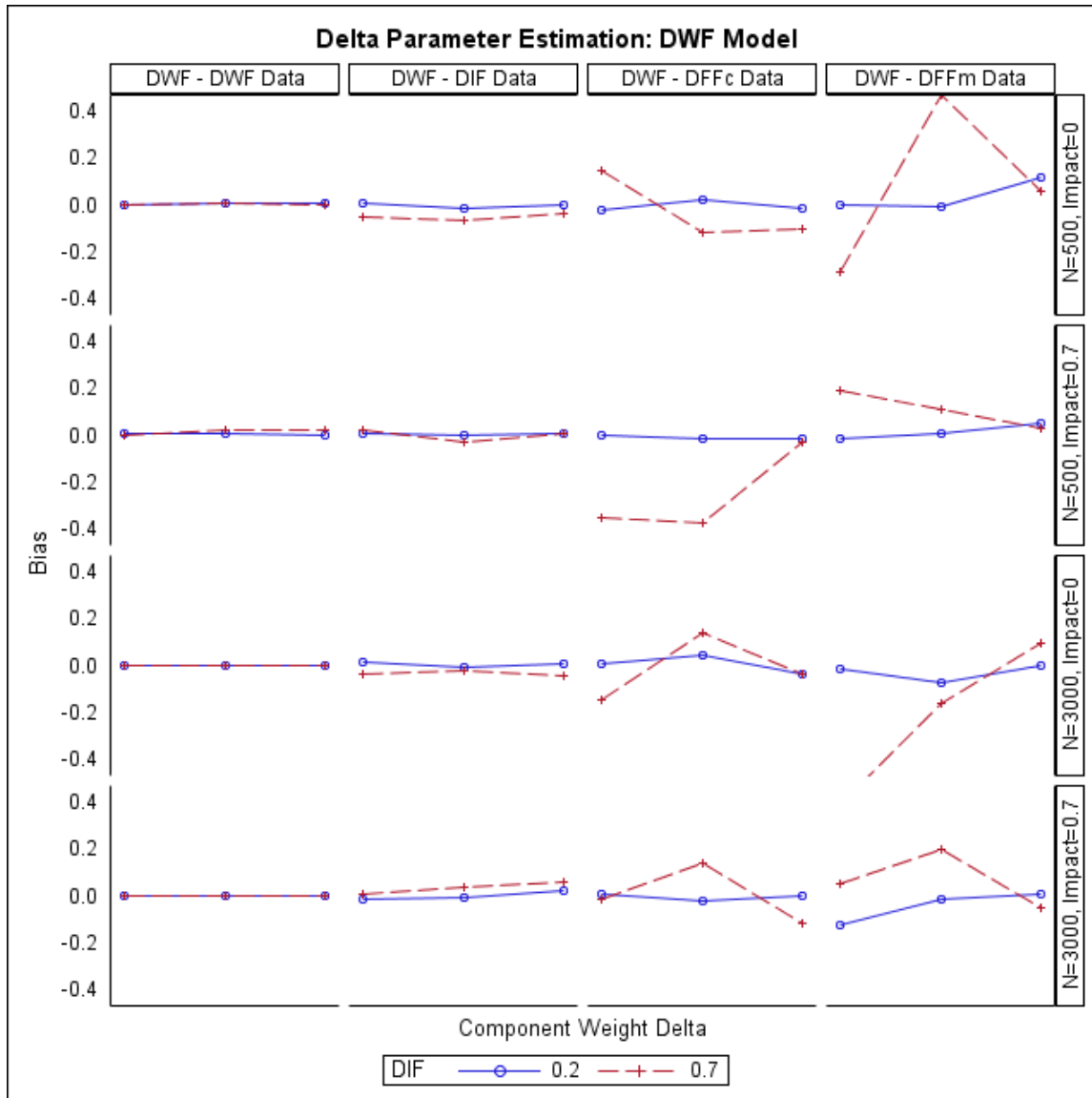


Figure 23. Bias of the three estimated DIF parameters of the MIRID DWF model when fitted to data with different sources of differential functioning

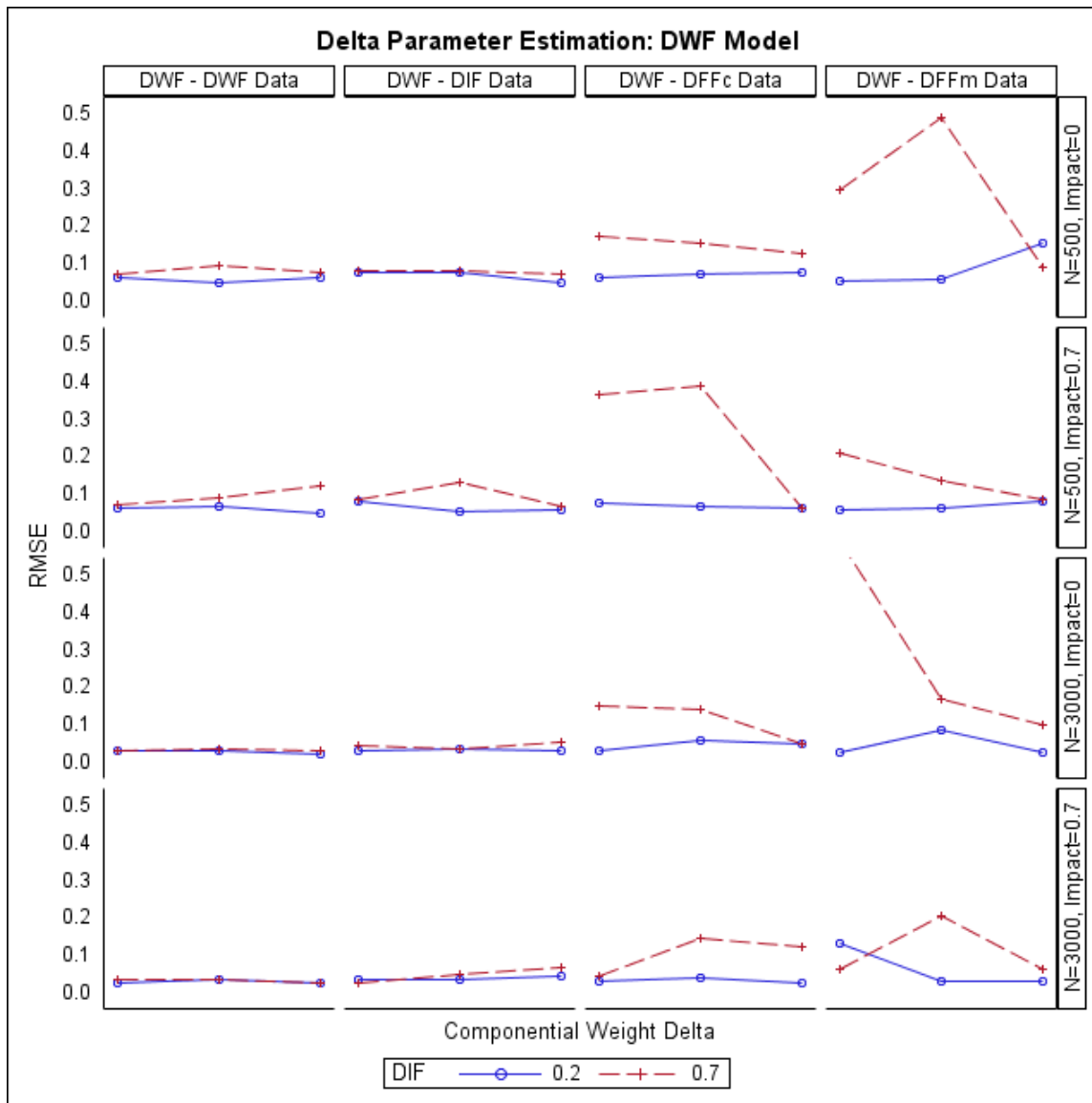


Figure 24. RMSE of the three estimated DIF parameters of the MIRID DWF model when fitted to data with different sources of differential functioning

Figure 25 shows the bias when fitting the DIF model to data of different DIF sources, including itself (individual item DIF). The dashed lines connect all bias values from the 30 estimates in larger delta conditions and the solid lines connect bias in smaller delta conditions. It is clear that over the DIF data (the leftmost column) the four items with simulated DIF effects had their locations over-estimated while the related DIF parameters were underestimated (see

Table 7). The other 26 location parameters had decent recovery outcome. Overall, the location parameters in the DIF model were under-estimated when the DIF effect in data matched and larger delta conditions led to greater bias.

Table 31

False Detection Rates when the MIRID DWF Model Was Applied to the DIF Data

		# Parms	Per-comparison %			Experimentwise %			
			Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg	
delta=-.2	N=250*2	Impact =0	3	0.050	0.020	0.020	0.140	0.050	0.050
		Impact =-.7	3	0.060	0.020	0.020	0.170	0.060	0.060
	N=1500*2	Impact =0	3	0.070	0.030	0.030	0.170	0.070	0.070
		Impact =-.7	3	0.080	0.030	0.040	0.190	0.080	0.080
delta=.7	N=250*2	Impact =0	3	0.160	0.080	0.090	0.400	0.220	0.220
		Impact =-.7	3	0.060	0.020	0.020	0.150	0.050	0.050
	N=1500*2	Impact =0	3	0.310	0.180	0.200	0.740	0.480	0.490
		Impact =-.7	3	0.290	0.180	0.190	0.610	0.400	0.400

Over the DFFc data, locations of the 10 items in the second component for which a DIF effect was simulated were slightly over-estimated whereas the other item locations were underestimated with larger delta causing greater bias. Similar story when the DIF model was applied to the DFFm model. Items with simulated DIF effect were over-estimated but on average the bias values were negative. Larger delta led to greater bias in both directions. Such distinctions were unclear with the RMSEs on Figure 26, where it is plain to see that larger sample reduced RMSEs.

The situation was different when the DIF model was fitted to the DWF model. Although larger delta still caused more fluctuations in bias, the average bias is no longer negative across all eight conditions like the other models. Smaller sample, larger delta, and no-zero impact created the greatest bias. Overall, when fitted to mismatched models, recovery of the location parameters in the DIF model was acceptable when delta magnitude was small (.2).

Table 32

False Detection Rates when the MIRID DWF Model Was Applied to the DFFc Data

		# Parms	Per-comparison %			Experimentwise %			
			Un- protect.	Bonfer- roni	Hoch- berg	Un- protect.	Bonfer- roni	Hoch- berg	
delta=.2	N=250*2	Impact =0	3	0.060	0.020	0.020	0.170	0.070	0.070
		Impact =-.7	3	0.050	0.020	0.020	0.140	0.050	0.050
	N=1500*2	Impact =0	3	0.200	0.110	0.130	0.430	0.270	0.270
		Impact =-.7	3	0.090	0.040	0.040	0.260	0.120	0.120
delta=.7	N=250*2	Impact =0	3	0.360	0.220	0.260	0.690	0.500	0.520
		Impact =-.7	3	0.660	0.590	0.630	0.980	0.950	0.960
	N=1500*2	Impact =0	3	0.770	0.720	0.770	1.000	1.000	1.000
		Impact =-.7	3	0.690	0.660	0.690	1.000	1.000	1.000

Table 33

False Detection Rates when the MIRID DWF Model Was Applied to the DFFm Data

		#	Per-comparison %			Experimentwise %			
			Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg	
delta=.2	N=250*2	Impact =0	3	0.120	0.050	0.060	0.350	0.150	0.150
		Impact =-.7	3	0.090	0.040	0.040	0.240	0.120	0.120
	N=1500*2	Impact =0	3	0.280	0.210	0.220	0.770	0.610	0.620
		Impact =-.7	3	0.380	0.340	0.350	1.000	0.990	0.990
delta=.7	N=250*2	Impact =0	3	0.720	0.680	0.710	1.000	1.000	1.000
		Impact =-.7	3	0.400	0.290	0.310	0.910	0.760	0.760
	N=1500*2	Impact =0	3	0.990	0.980	0.990	1.000	1.000	1.000
		Impact =-.7	3	0.670	0.570	0.630	1.000	1.000	1.000

As Figure 27 shows, when fitted to its own data and the DWF data, the DFFc model produced the least amount of bias with minor difference between conditions of smaller and larger delta. When applied to the DIF and DFFm data, however, significant bias occurred at items where DIF effect was simulated even with smaller delta conditions (the solid) lines. The pattern of deviance is less clear with the RMSEs on Figure 28, where the estimation quality seems worse.

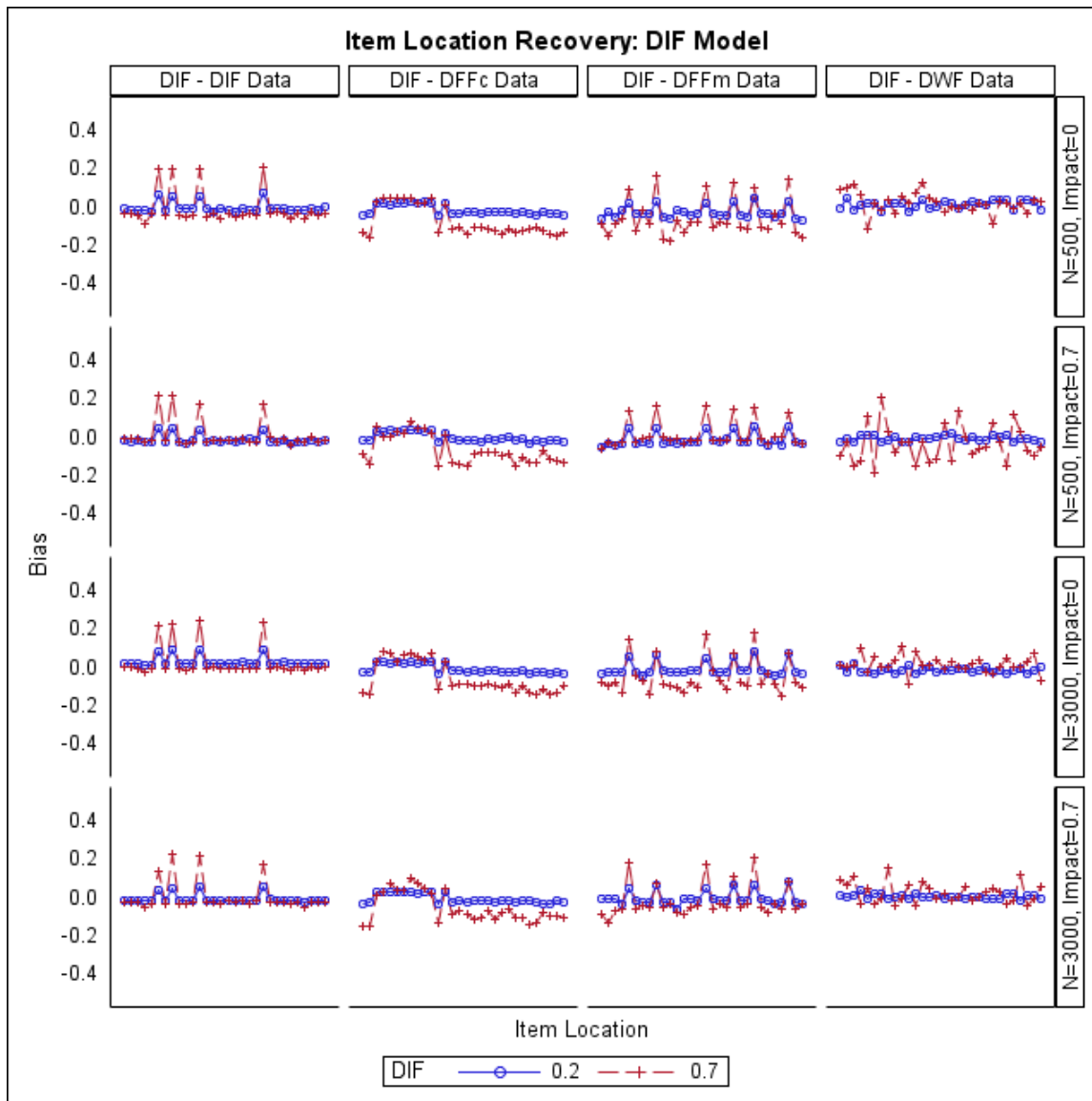


Figure 25. Bias of the Estimated Item Locations when the DIF Model Was Fitted to Different Models

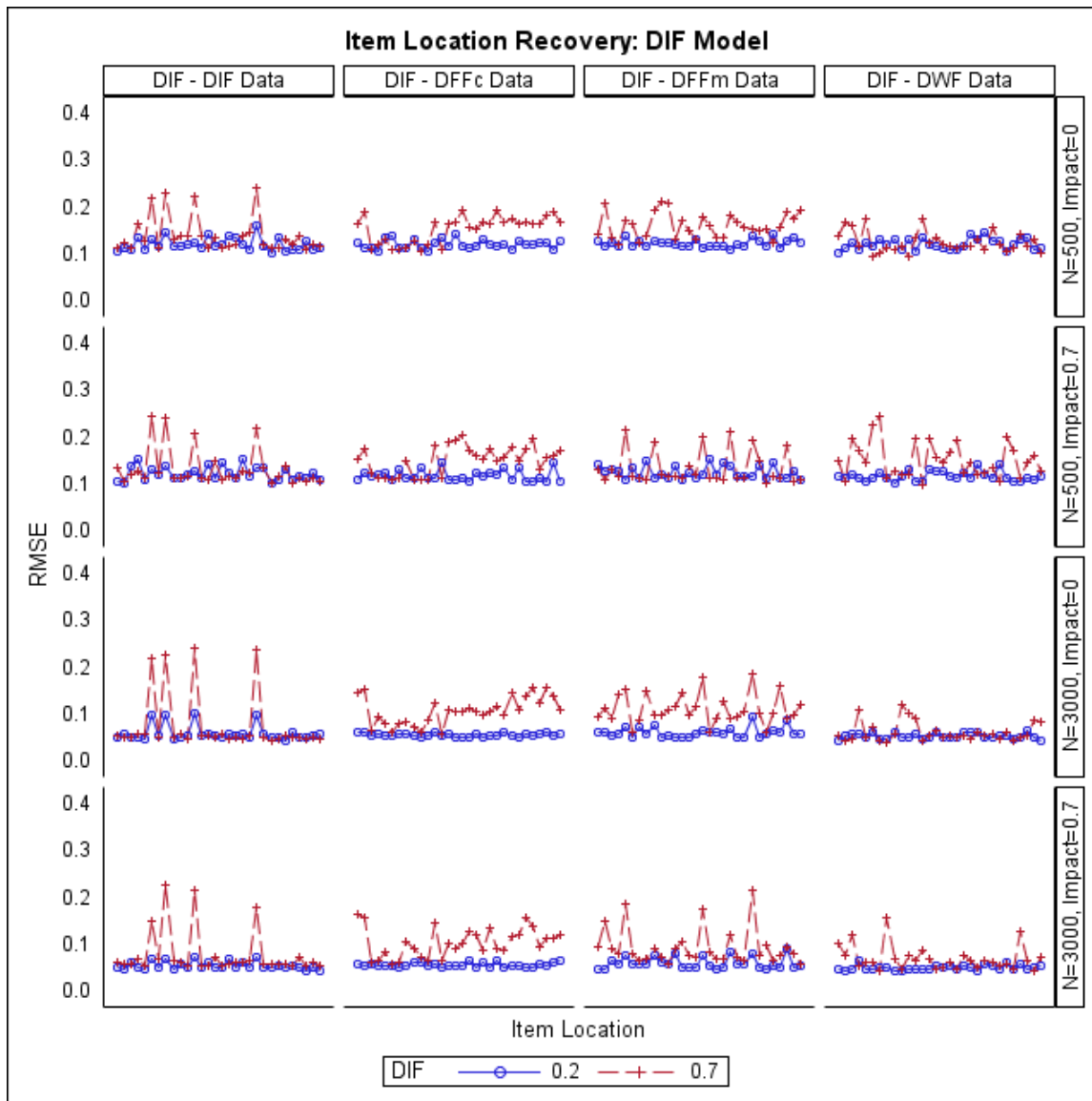


Figure 26. RMSE of the Estimated Item Locations when the DIF Model Was Fitted to Different Models

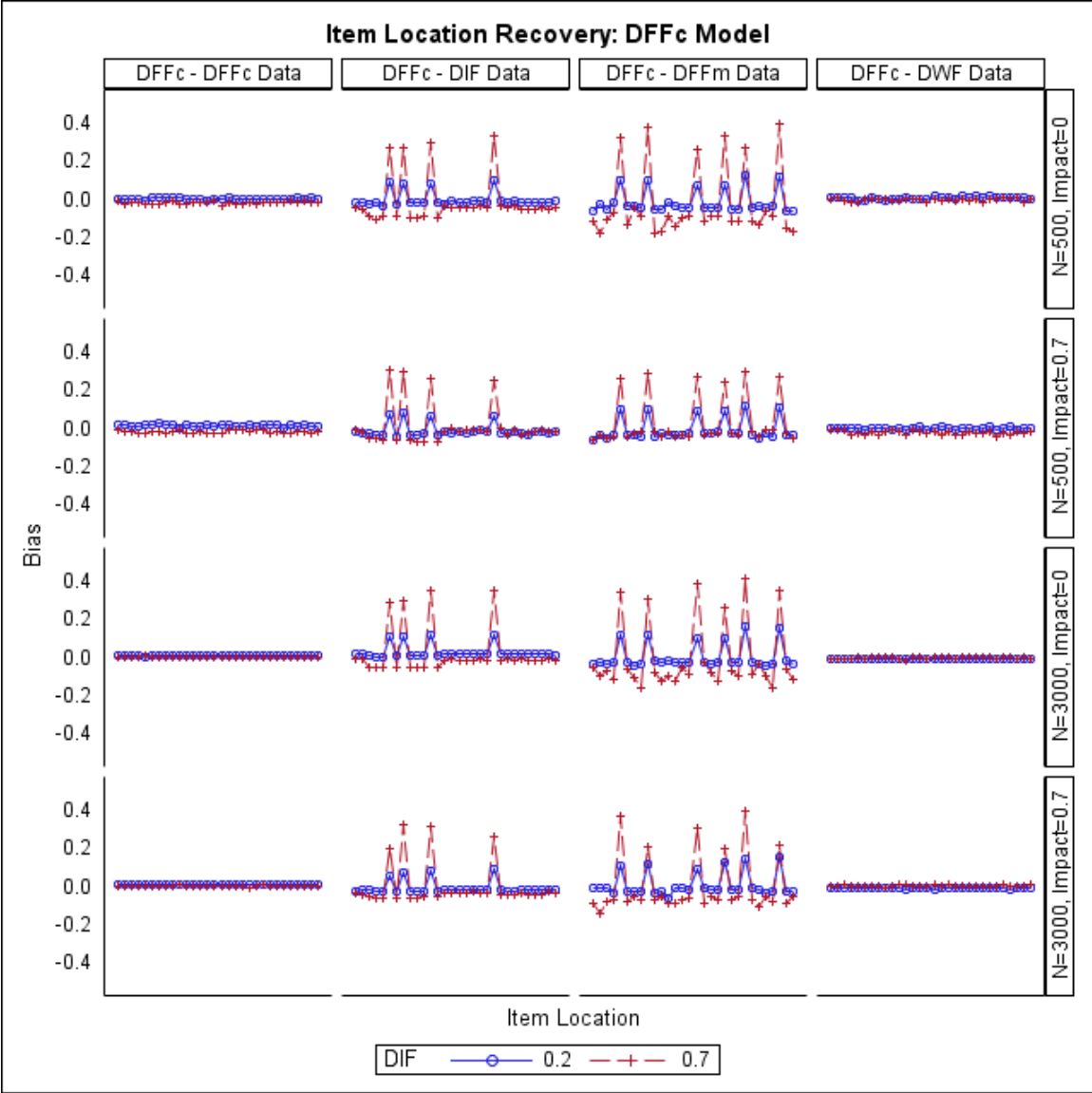


Figure 27. Bias of the Estimated Item Locations when the DFFc Model Was Fitted to Different Models

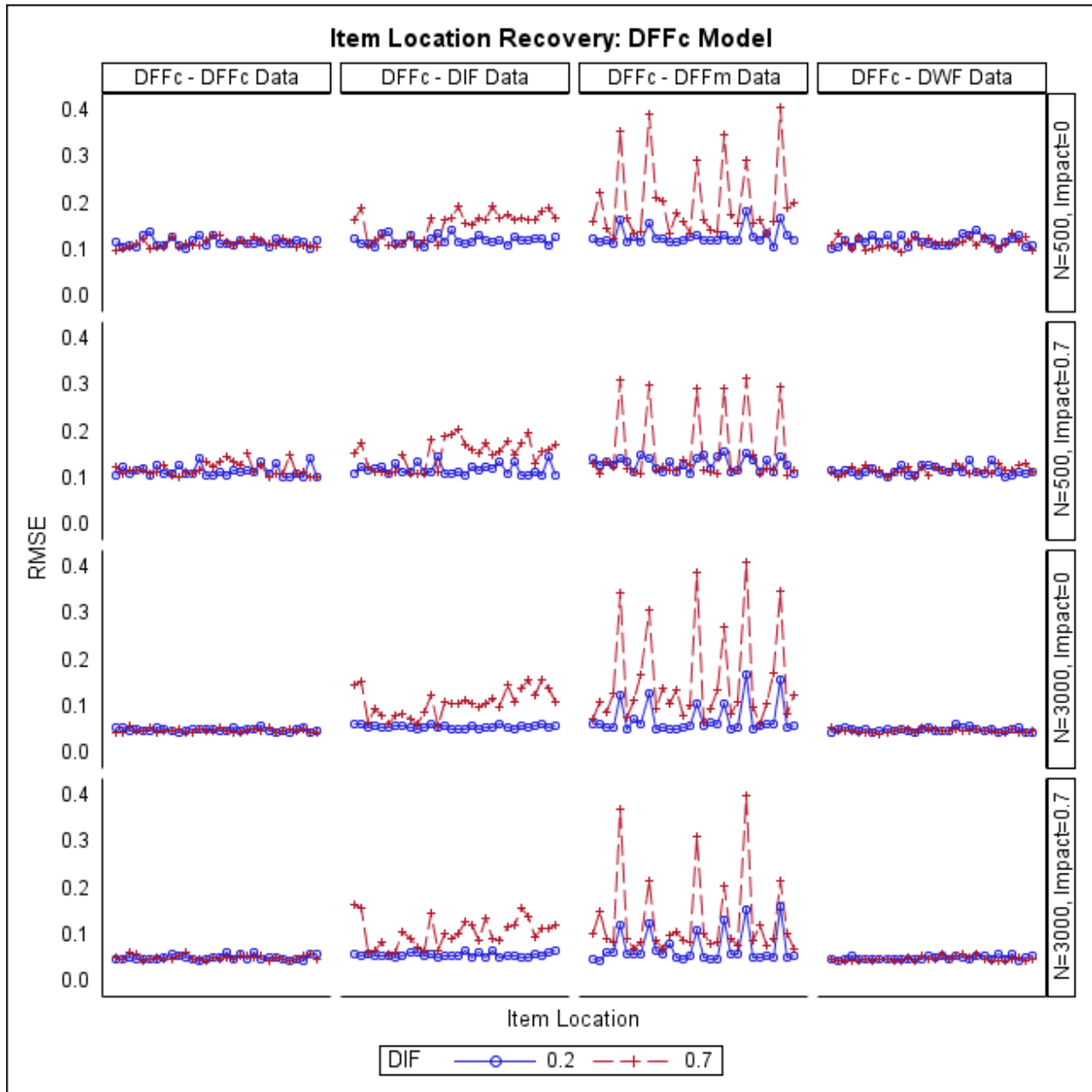


Figure 28. RMSE of the Estimated Item Locations when the DFFc Model Was Fitted to Different Models

In Figure 29, item location estimates of the DFFm model were seriously distorted by the delta size at items with non-zero DIF effects. The influence of larger delta size was evidenced by the significant fluctuations of the dashed lines. Again, items where DIF was simulated were

over-estimated whereas the rest were underestimated. The RMSEs for smaller delta and larger sample size were in general acceptable ($< .1$) (Figure 30).

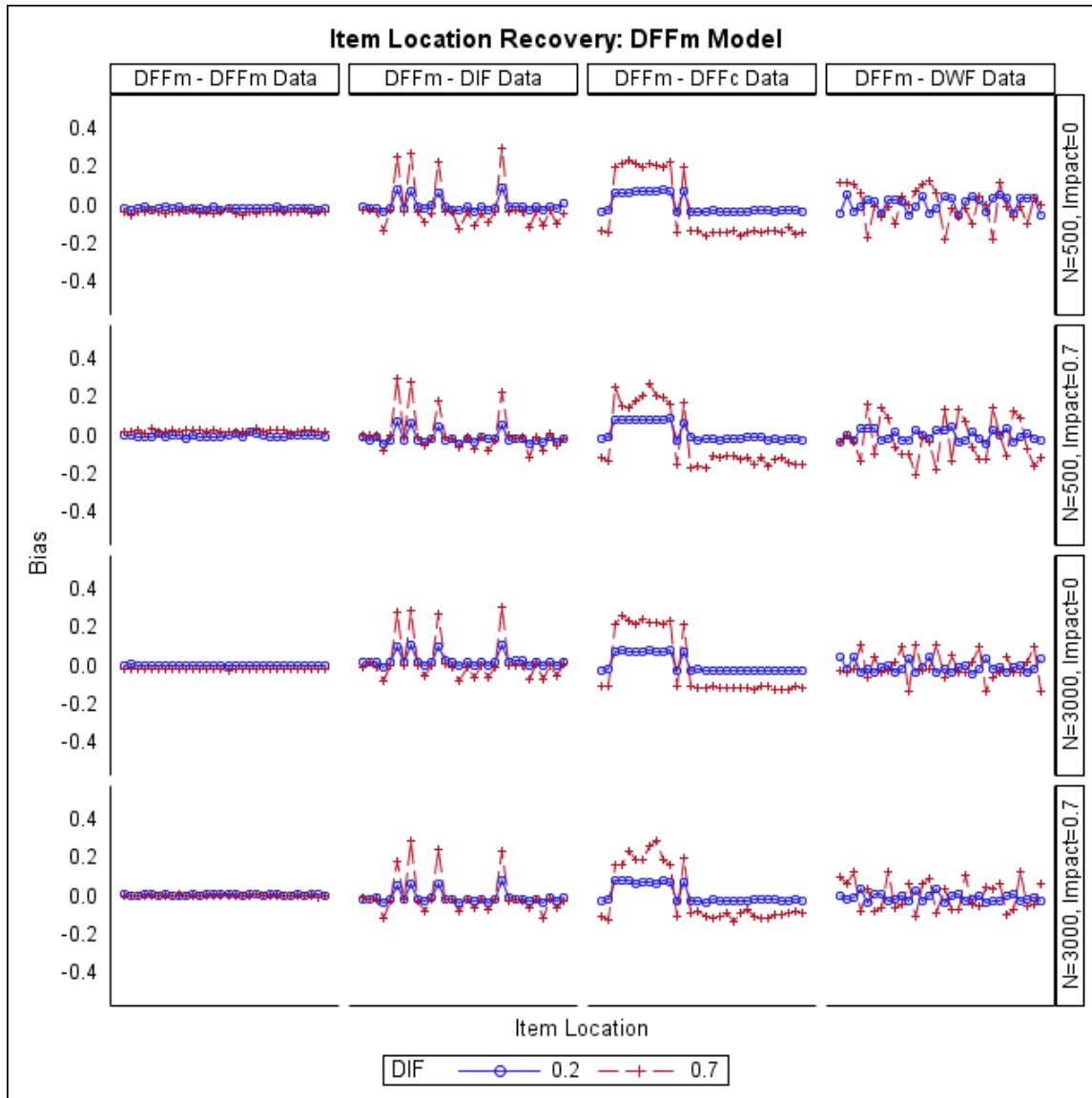


Figure 29. Bias of the Estimated Item Locations when the DFFm Model Was Fitted to Different Models

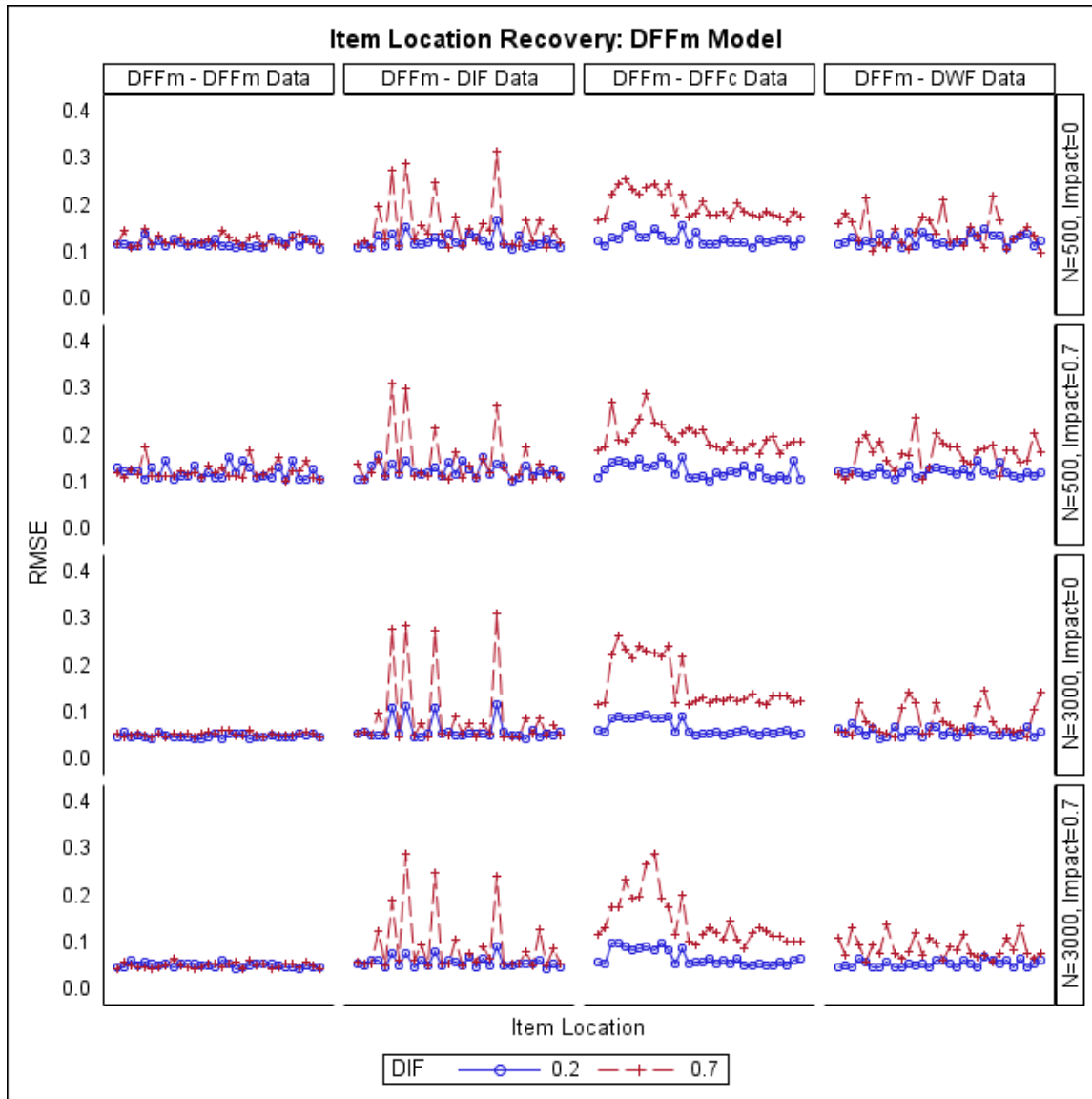


Figure 30. RMSE of the Estimated Item Locations when the DFFm Model Was Fitted to Different Models

Fitting the DWF model to mismatched data resulted in similar pattern but with greater bias (Figure 31). For example, comparing the bias over the DFFc data in Figures 25, 27 and 29 revealed that the DWF model generated the greatest amount of bias in either direction. Interestingly, the DFFm data in the conditions of no group difference and larger delta led to the

greatest under-estimation of the items without simulated delta. In Figure 32, the RMSEs look much worse in larger DIF condition for items with simulated delta effects.

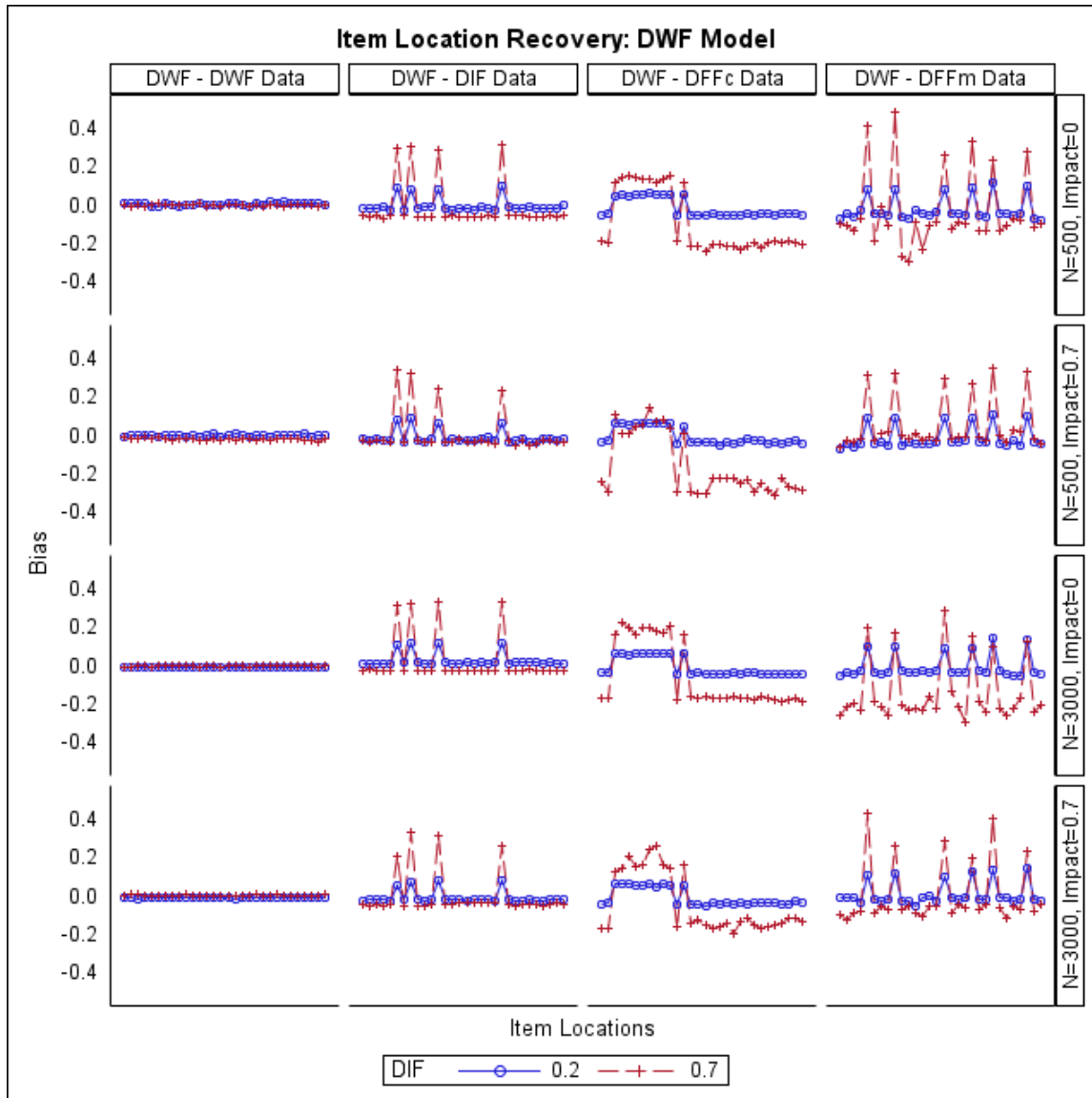


Figure 31. Bias of the Estimated Item Locations when the DWF Model Was Fitted to Different Models

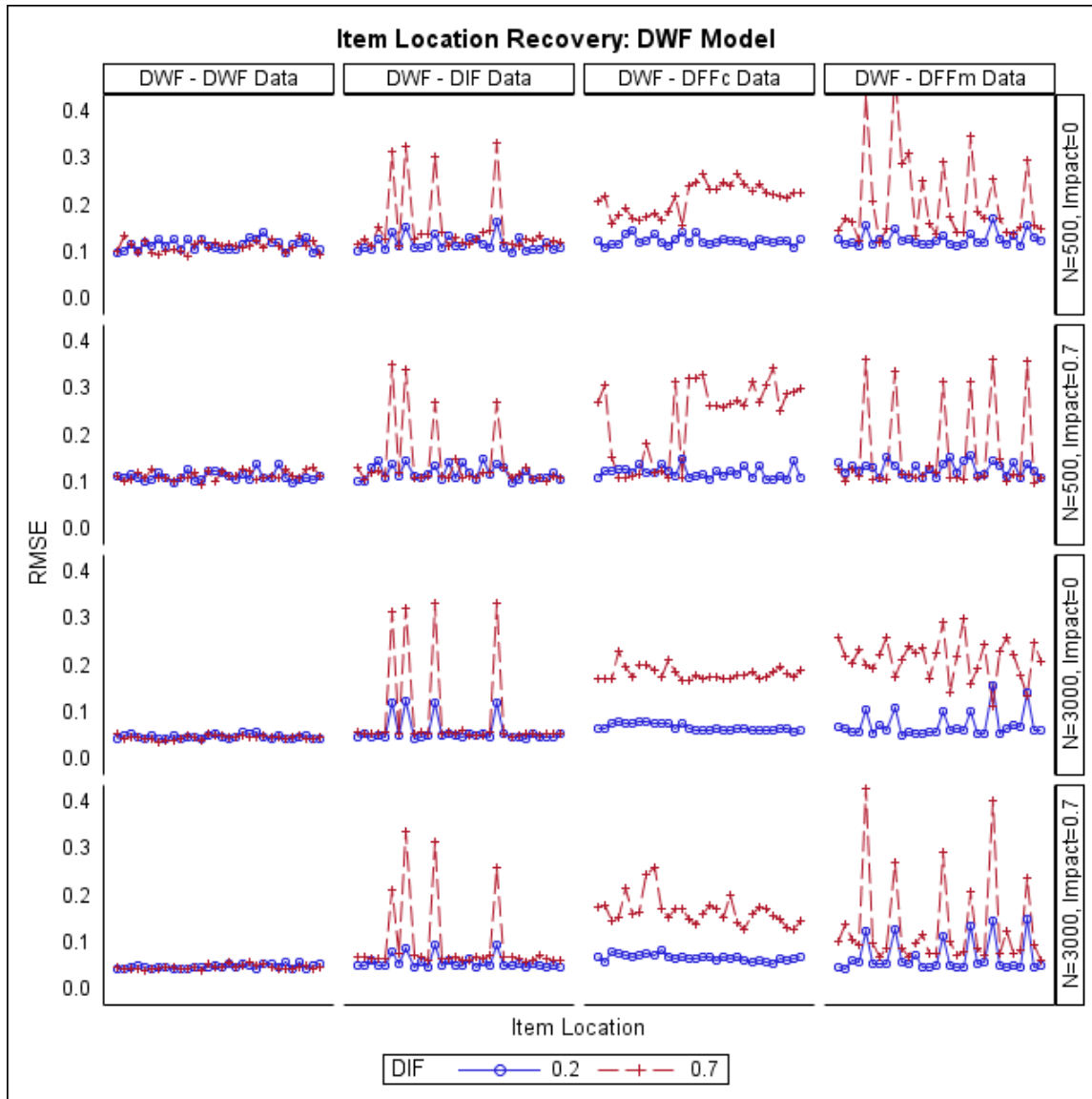


Figure 32. RMSE of the Estimated Item Locations when the DWF Model Was Fitted to Different Models

In parallel, recovery of the intercept and three component weights is presented in Figure 33 to 40, where the first data point in each plot represents the intercept and the remaining three the component weights. Recovery of these parameters by models over their own data is presented on the leftmost column of each graph as reference. Because estimation of the variance

and group difference parameters did not result in conspicuous bias and a manifest pattern, its results are not discussed here.

In Figure 33, when fitting the DIF model to the DFFc data, the intercept was over-estimated in all conditions although the component weights did not have much bias. With the DWF data, the second component weight had the worst bias due to the simulated delta effect, especially when it was larger. The corresponding RMSEs (Figure 34) are tremendous. Overall, estimation of component weights over mismatched data had poor quality even in larger-sample scenarios.

The DFFc model produced smaller bias, especially with the DIF data across all conditions (Figure 35). In fact, the recovery here was almost as good as the DFFc data as seen on the leftmost column. Its performance when fitted to the DFFm data resembled that of the DIF model with the DFFm data. Its recovery when applied to the DWF data was comparable to that of the DIF model fitted to the same data but with smaller bias and RMSEs (Figure 36) on the second component, which was associated with simulated delta, even in the condition of larger delta, smaller sample, and non-zero group difference.

Compared to the other three models when applied to mismatched data, the DFFm model produced the least amount of bias and RMSE in estimation of the component weights (Figure 37 and 38). Again the pattern resembled those of the previous models but both the dashed and solid lines were very smooth with the effect of delta size only obvious in the DWF data. Interestingly, the DFFm model had much lower bias and RMSEs when fitted to the DFFc data than when the DFFc model applied to the DFFm data. There appeared to be little difference in effects from larger and smaller delta.

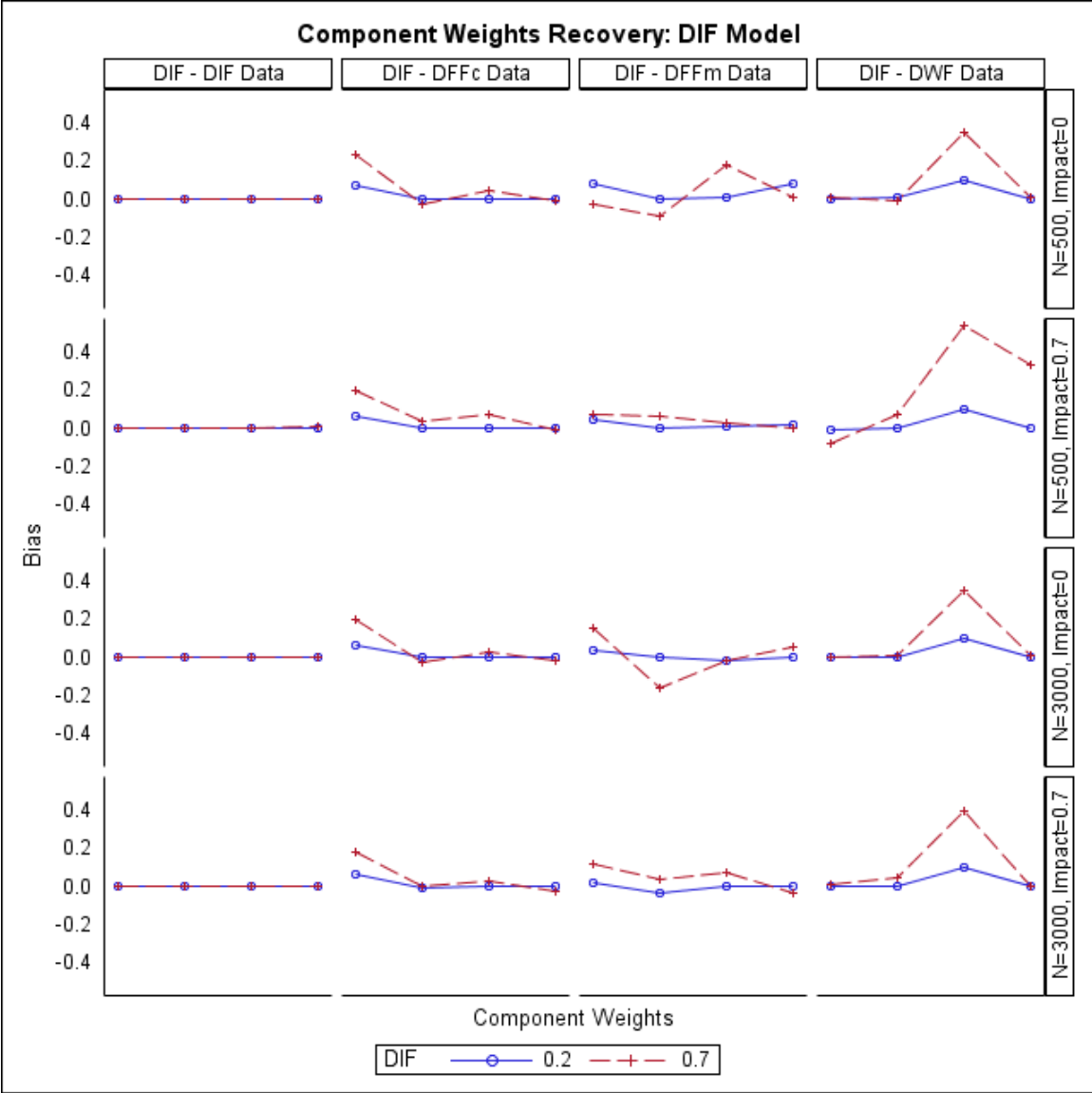


Figure 33. Bias of the Estimated Component Weights when the DIF Model Was Fitted to Different Models

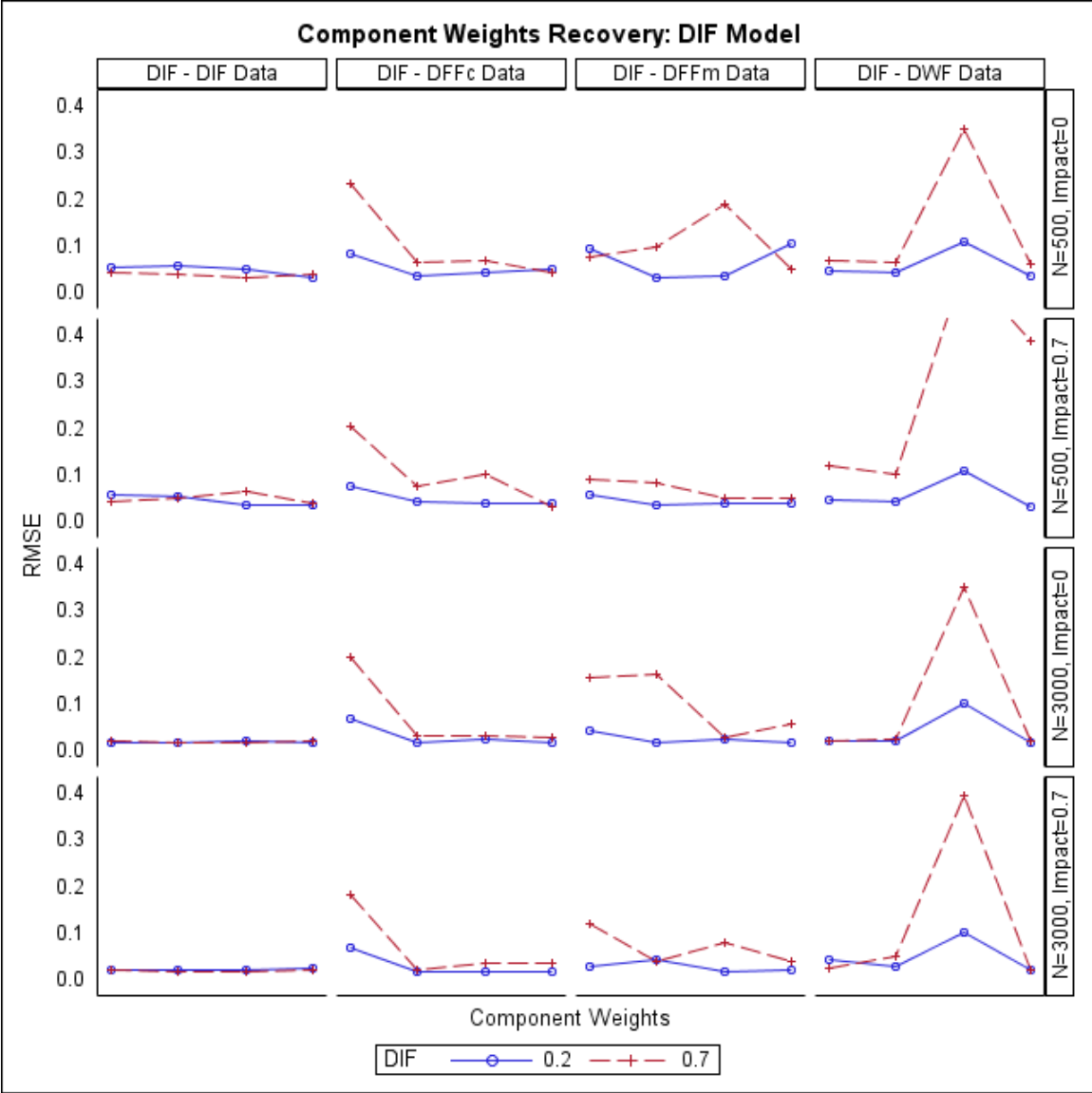


Figure 34. RMSE of the Estimated Component Weights when the DIF Model Was Fitted to Different Models

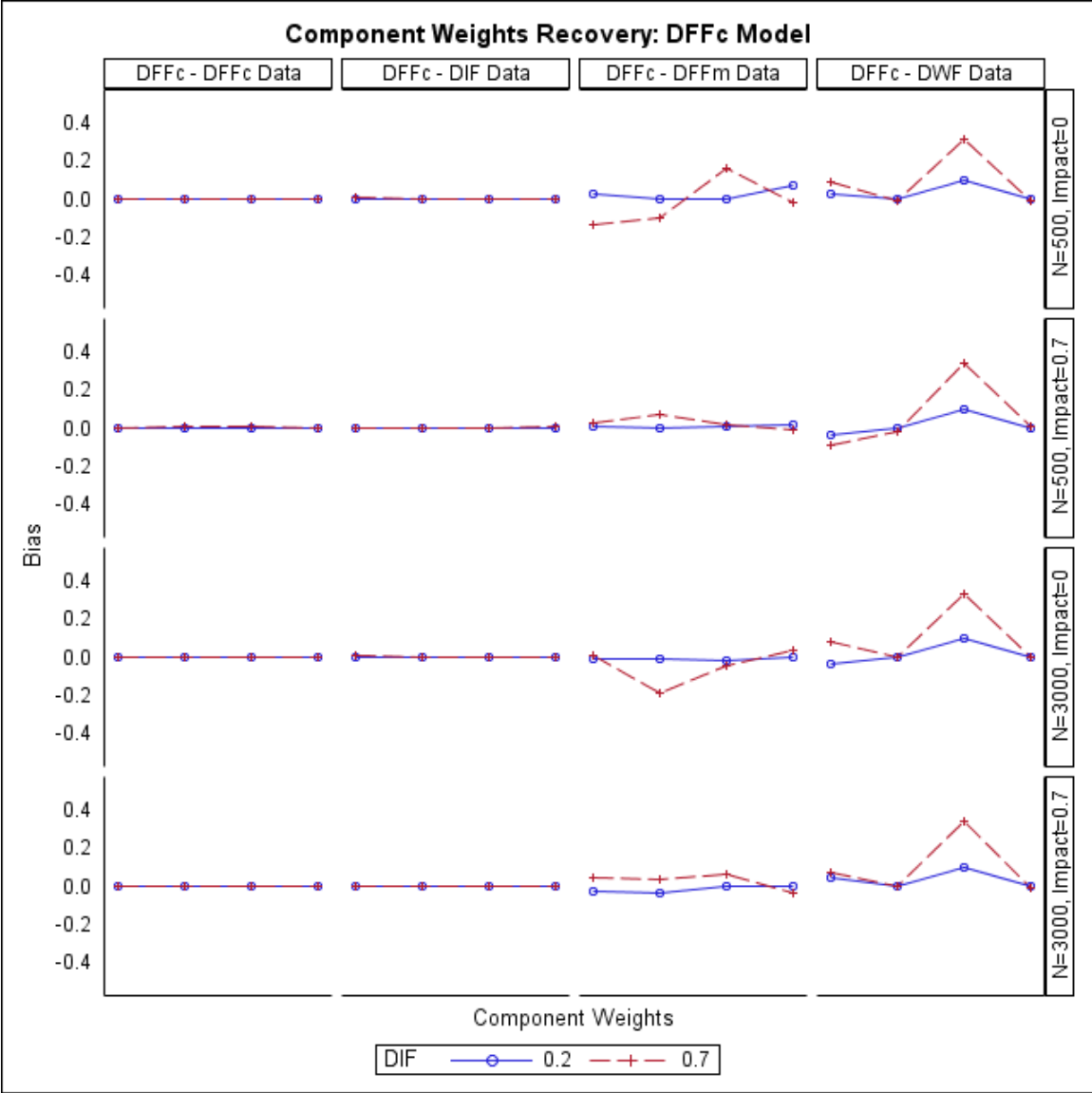


Figure 35. Bias of the Estimated Component Weights when the DFFc Model Was Fitted to Different Models

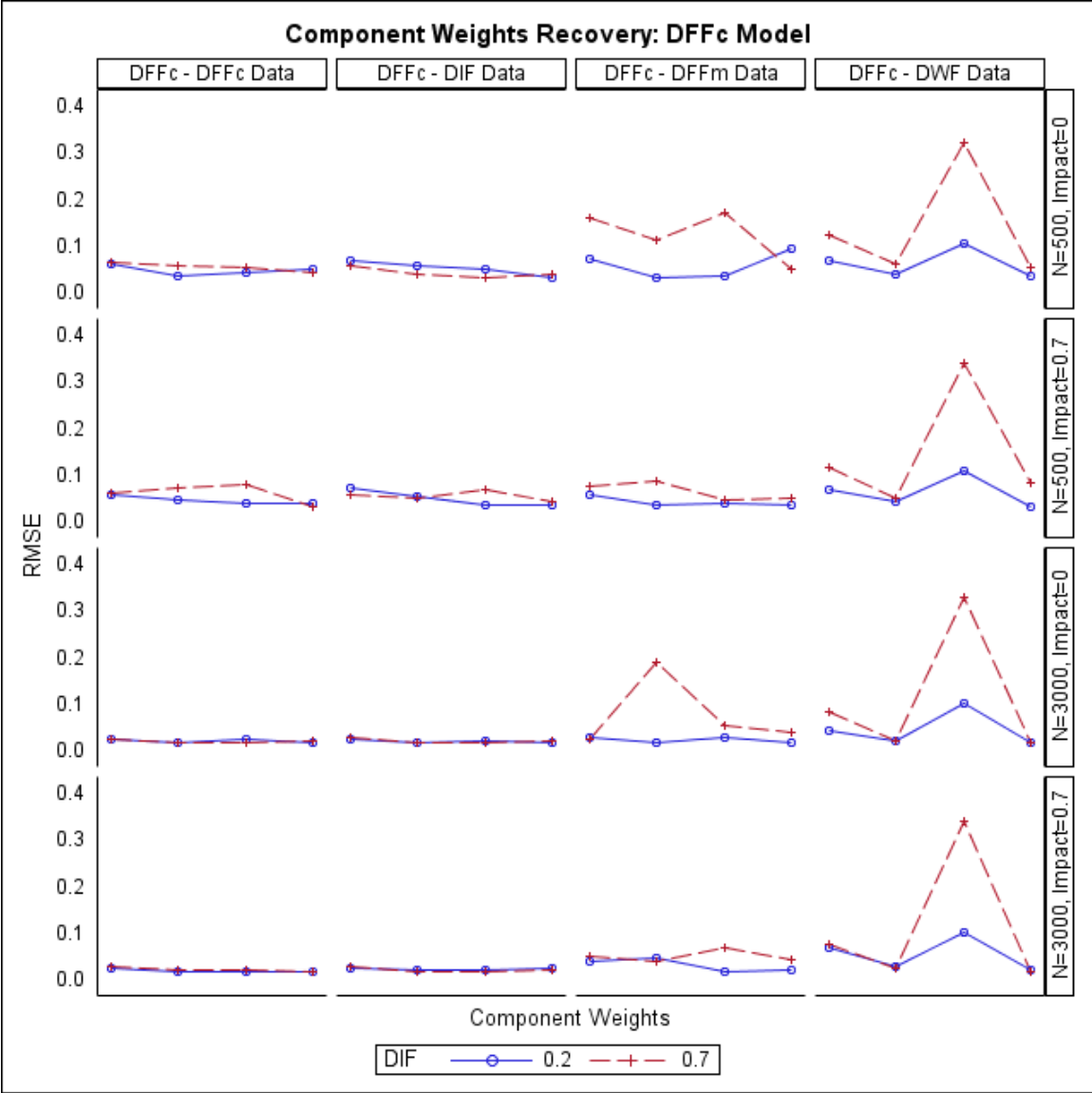


Figure 36. RMSE of the Estimated Component Weights when the DFFc Model Was Fitted to Different Models

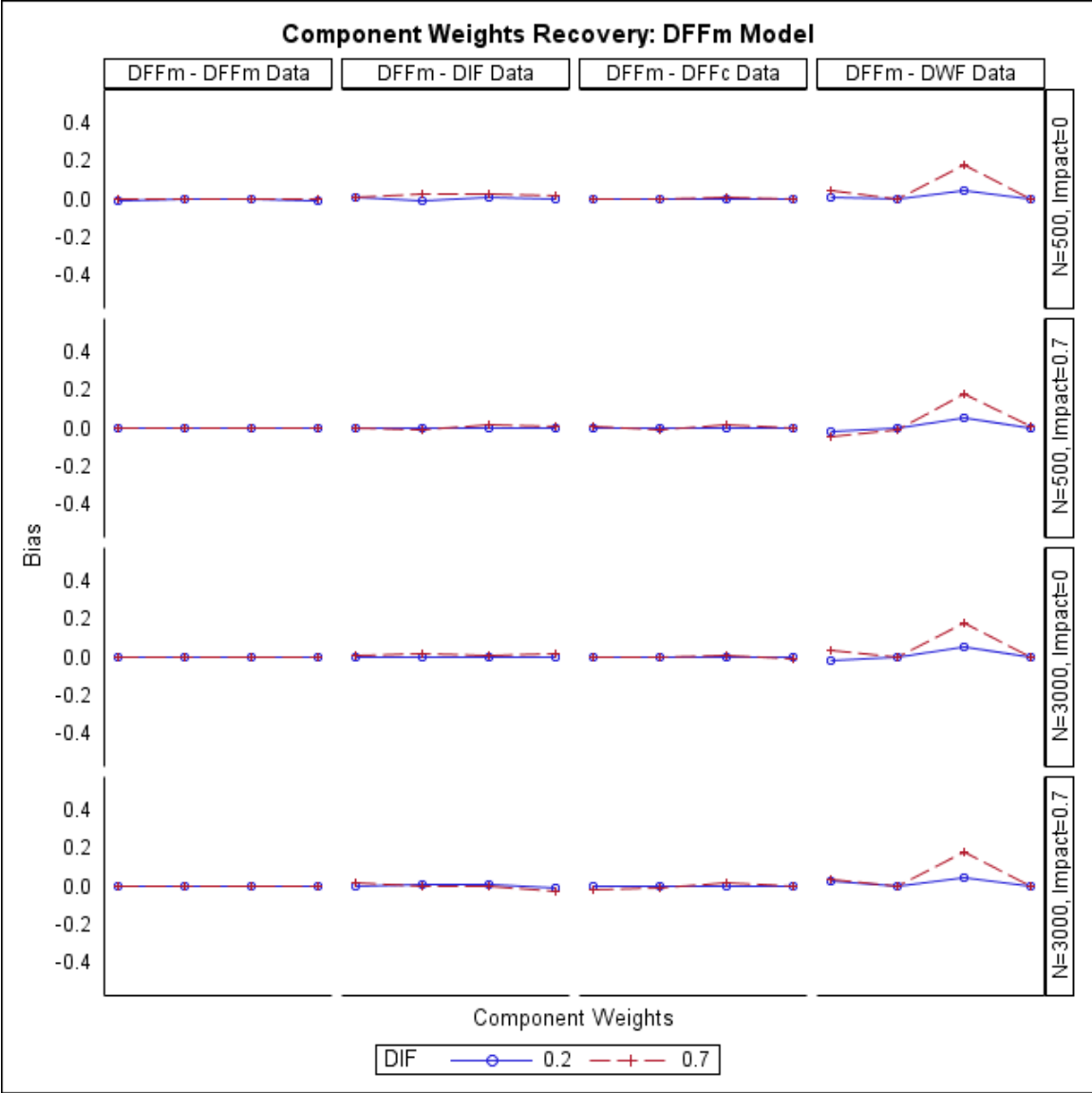


Figure 37. Bias of the Estimated Component Weights when the DFFm Model Was Fitted to Different Models

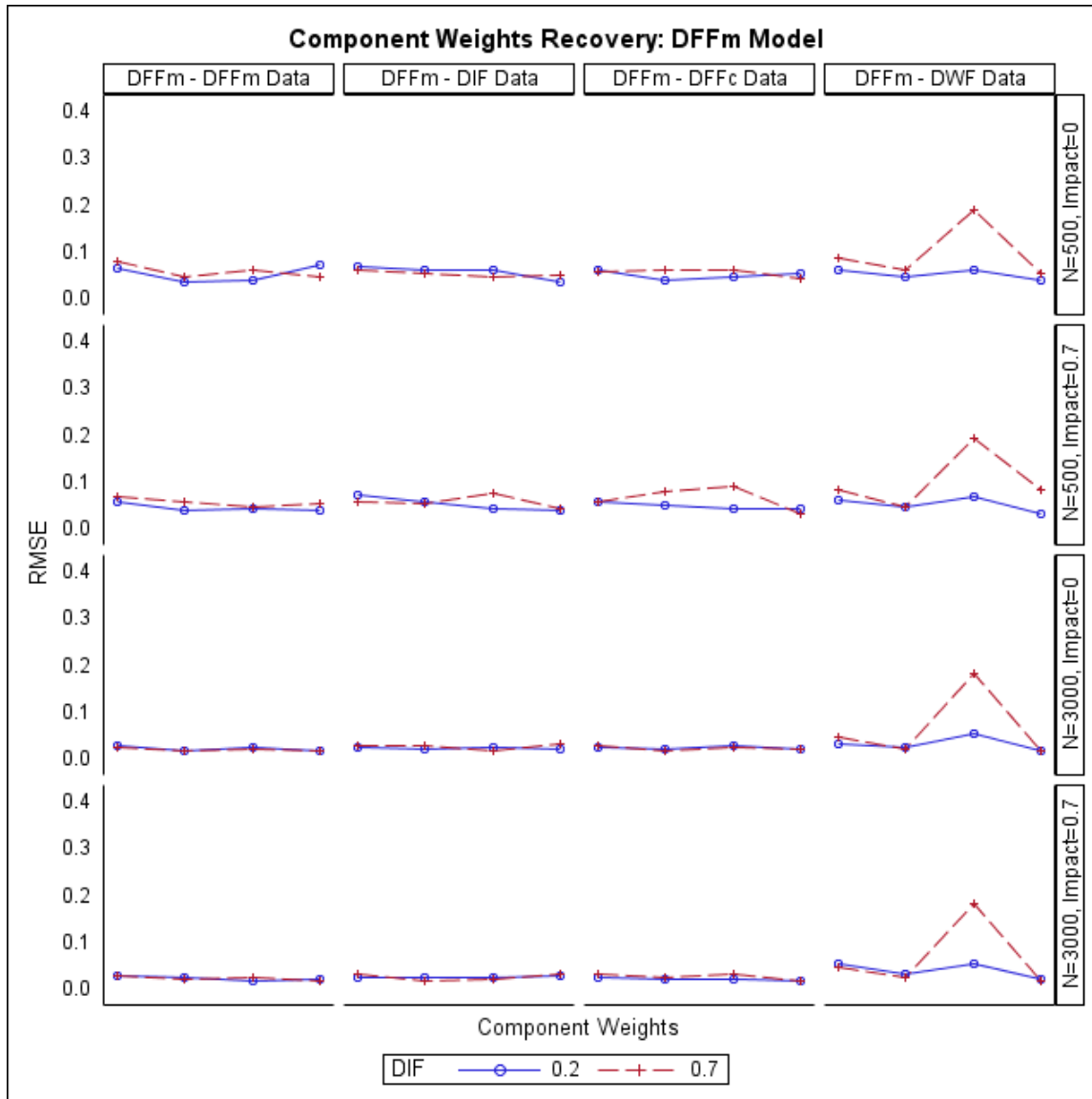


Figure 38. RMSE of the Estimated Component Weights when the DFFm Model Was Fitted to Different Models

In Figures 39 and 40, the DWF model produced negligible bias and RMSEs with the DIF data in estimation of the component weights and the intercept. For the DFFc and DFFm data, there were significant bias and RMSEs on the intercept and one or two components, especially in conditions of larger DIF. In comparison to the estimation of location parameters (Figure 31 and 32), the level of bias in this outcome was malevolent.

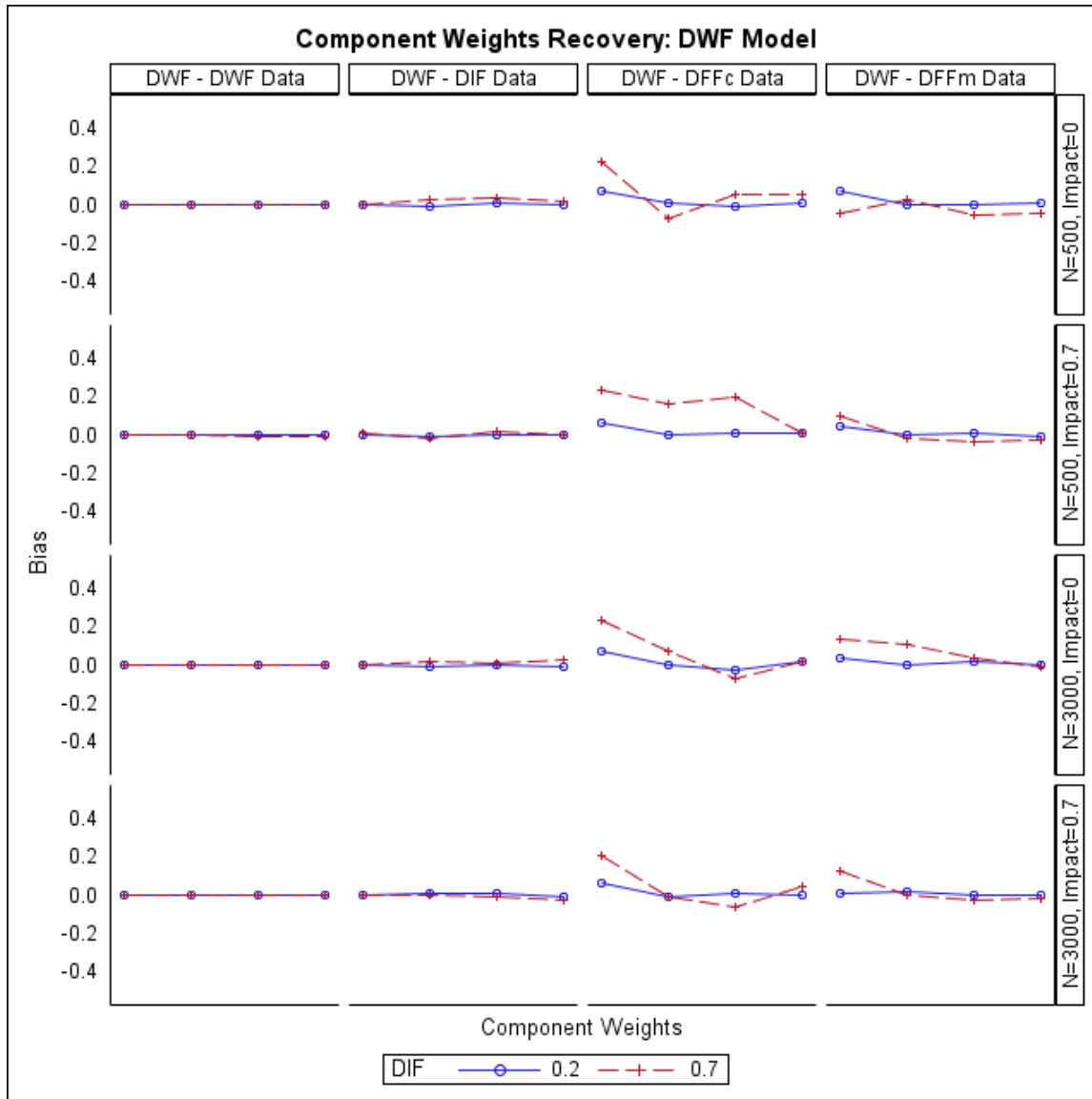


Figure 39. Bias of the Estimated Component Weights when the DWF Model Was Fitted to Different Models

To sum it up, the four proposed DIF models produced the best recovery results when fitted to data generated consistent with the model. When fitted to mismatched data, there were much greater bias and more considerable RMSEs in estimation of item locations than with component weights; noticeably, estimation quality was adversely affected by the presence of differential functioning. Although the influence of sample size and group difference on

estimation bias was not persuasive, for both individual item DIF and item group DIF, large delta magnitude resulted higher bias across most of the conditions under the mismatched models.

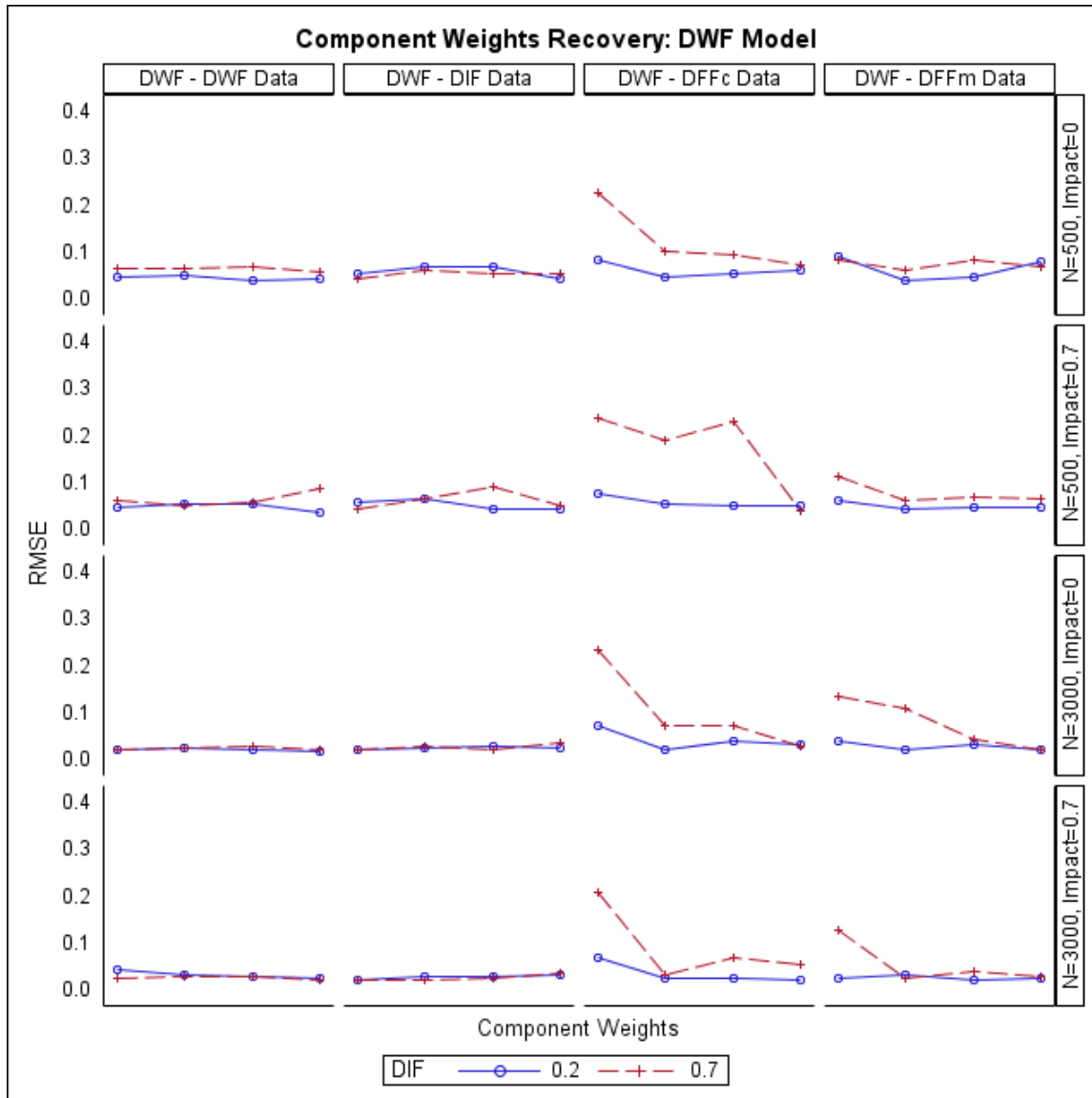


Figure 40. RMSE of the Estimated Component Weights when the DWF Model Was Fitted to Different Models

In summary, this section answered the fourth and fifth research questions. When the wrong model was fit to the data, location parameters were generally not estimated well for items

associated with generated DIF effects but component weights were estimated better. The most influential adverse design factor was delta size, which produced significant bias and RMSEs across conditions of different sample size and impact. The zero-value DIF parameters as well as other model parameters were estimated when the model was fitted to each mismatched data set. As a result of simulated differential functioning in other items or item group in the data, bias and RMSEs of these estimates were much greater than when the correct model was applied. The high level of estimation errors with DIF parameters led to high level of false detection rates as evaluated with unprotected per-comparison and Hochberg-adjusted errors. For both measures, larger delta source and larger sample always led to error rates above Bradley's upper boundary (.075) and in conditions of smaller delta and smaller sample the false detection rates were often acceptable. The model with the least estimation error of the DIF parameters and thus the lowest false detection rates was the MIRID DFFc model.

CHAPTER FIVE

DISCUSSION

This chapter presents a summary of the research, its findings, implications, and recommendations for researchers. The limitations and suggestions for future research are also discussed.

Summary

This dissertation proposed extensions to the model with internal restrictions on item difficulty (MIRID) to study differential item functioning (DIF). Each of the proposed models corresponds to a distinct potential source of differential functioning in the MIRID data: the MIRID DIF model (differential functioning in individual items), the MIRID DFFc model (differential functioning in components), the MIRID DFFm model (differential functioning in item families), and the MIRID DWF model (differential functioning in component weights). These models are designed to capture differential effects by specifying DIF parameters in their formulation in addition to the regular model parameters. As members of the Rasch family of models, estimation method of these proposed models was conceived as maximum likelihood method in keeping with the standard MIRID.

A simulation study was conducted to examine model recovery, Type I error rates, and power under practical measurement conditions as well as recovery and false positive detection rates with mismatched models. Three factors were manipulated in the simulation study: sample size (500 and 3,000), magnitude of DIF (δ) (0.2 and 0.7), and group difference (impact) (0

and -0.7). Consequently, cross-product of these factors constructed eight conditions for each of the four proposed models. For each condition, 500 data sets were generated. Because the estimation quality of the specified DIF (δ) and the non-DIF parameters affects the Type I error rates and power in DIF detection, the recovery of these parameters under each model was crucial and was evaluated by calculating bias and root mean squared error (RMSE) of the parameter estimates. Type I error rates and power were calculated to assess the effects of different testing conditions on detection of both item-level and group-level DIF for all eight conditions under each proposed model. Specifically, two types of Type I error rates were calculated for each condition: per-comparison (PCER) and experimentwise (EWER), for which, in addition to the unprotected error rates, adjusted values were computed using Bonferroni and Hochberg procedures. The Bonferroni adjustment produces Type I error rates similar to the Hochberg method but was more conservative in detection. The liberal range suggested by Bradley (1978) provided reference for assessing the Type I error control. Parallel to PCER and EWER, power in DIF detection was evaluated with per-pair and any-pair indices, each of which was also calculated three times: the unprotected and the Bonferroni and Hochber adjusted ones.

To further study the characteristics of the four proposed models, they were fitted to data generated using other models. This part of the study investigated the potential impact from applying the wrong model. The robustness of the model estimation when mis-specified and how severe the false detection rates were associated with each data simulation condition were examined.

Findings

Data generation and estimation methods used in this research were validated in a separate parameter recovery study prior to the main research. For a standard MIRID with three

components and ten item families (30 item location parameters, 3 component weight, and 1 intercept), adequate accuracy in parameter estimation was observed.

Item-Level DIF

In the DIF model, parameter recovery was less than acceptable, especially in large DIF conditions where the magnitude of delta was observed as the most influential factor. Both the four true DIF and 26 zero-value DIF parameters were underestimated with varying RMSEs. Conspicuously, both bias and RMSEs were significantly greater in conditions of large DIF magnitude ($\delta = .7$). As expected, large sample size decreased the RMSEs for all parameters but not the bias. In small delta conditions, only large sample coupled with zero impact caused great RMSEs for DIF-associated component items. The effect from group difference was mostly weak and inconsistent.

Because the MIRID DIF model was formulated with the product of the component weight and summation of the item-specific delta parameter and item location, the reason for the inadequate estimation may be that the computer program did not distinguish well the two parts within the summation in their interaction with the weight. One possible explanation is that the estimation process may give greater share of the summed value to item location, the first of the summation, than the delta parameter.

Another factor that could have possibly contributed to the less than adequate estimation of the DIF and item location parameters was the large number of parameters (66 in total) to estimate at once. To answer this question, a pilot study was carried out to estimate one DIF parameter at a time. It was found that the recovery, Type I error rates, and power were very similar to the initial analysis when the DIF parameters in one component were estimated together.

Comparison of Type I error rates and power obtained through the two methods is presented in Appendix D.

The effect of DIF size on DIF parameter recovery was inconsistent according to literature. For example, Paek and Wilson (2011) discovered that for the Rasch DIF model the large DIF size (.681) had on average slightly greater bias than with the medium DIF (.468). Similarly, Fukuhara and Kamata (2011) found that with the 2PL model and a bi-factor multidimensional IRT model bias in large DIF (.7) parameter estimates was on average greater than that in medium DIF (.5) parameter estimates. On the other hand, other IRT-based studies that modeled item-level DIF (Jeon, Rijmen, & Rabe-Hesketh, 2011) discovered minor differences in bias between small/zero and medium delta magnitude in data conditions which did not resemble those specified in this research. In the DIF literature, design variables such as model characteristics, sample size, and test length confound the effect of delta size on DIF parameter estimation.

In terms of statistical inferences, the unprotected PCER under the DIF model was controlled in all small DIF ($\delta = 0.2$) conditions but not with large DIF ($\delta = 0.7$), especially when sample size was also large ($N = 3,000$). In other words, many of the biased estimates of the non-DIF parameters were significantly different from zero, resulting in high Type I error rates. The unprotected EWER was above the nominal level in all conditions. However, the two adjustment procedures had PCER under control in all conditions and EWER under control in small delta conditions. Such findings are consistent with previous DIF studies showing that Type I error rates tended to inflate as sample size and the degree of differential functioning increased (e.g., Kim et al., 2011). When Type I errors were under control, there was good per-pair power in large sample and large delta conditions.

Because conventional DIF detection methods such as logistic regression and Mantel-Haenszel focus on one item after another, it is normal to consider only hypothesiswise Type I errors (PCER) under no adjustment conditions, which is reported most frequently in the DIF literature. Nonetheless, for model-based approaches like proposed in this study where more than one specified DIF parameters are evaluated at once, it presents a situation of multiple significance testing, for which the appropriate Type I error rates are familywise or experimentwise Type I error (EWER). To deal with the common problem of greater EWER in situations like this, adjustment procedures have been devised to correct the critical criteria such as the Bonferroni procedure (Bonferroni, 1936) and the similar but less conservative Hochberg correction (Hochberg, 1988) in order to keep the Type I error under control. The findings on the DIF model support the use of these corrections. In parallel to PCER and EWER, this study also calculated two kinds of power indices, per-pair and any-pair.

Group-Level DIF

Estimates of the DIF and non-DIF parameters of the three group-level DIF models, DFFc, DFFm, and DWF, were not biased. In large sample conditions, average bias and RMSEs of these parameter estimates were very similar, particularly in the DFFc model. There was inconsistent variation between the two levels of impact in small sample conditions and it was the most obvious for the DWF parameter. Such effect disappeared in large sample conditions. DIF parameters in the DFFc and DFFm models are only an additive component in the specification and were relatively easy to estimate. On the other hand, although the DWF parameter was similarly formulated to the DIF parameter in the item-level DIF model, their estimates were not biased. A possible explanation is that their estimation was more informative as their quantity was shared by 10 component items.

For group-level models, all three PCER indices were controlled across conditions and the EWER with the two adjustments were under control. There was perfect or near perfect per-pair and any-pairs statistical power in large DIF conditions and small DIF but large sample conditions. Even when sample and delta were small, decent per-pair and any-pair power were observed.

These findings are consistent with previous studies on group-level DIF, which provided good Type I error control and greater statistical power than individual item DIF analyses (e.g., Banks, 2013). Nevertheless, many of previous studies employed SIBTEST or similar procedures, and little research has been conducted on group-level DIF (i.e., differential bundle/facet functioning) using model-based approaches. A very pertinent example (Nixon, 2013) examined issues on a model-based DIF approach under the logistic linear test model (LLTM) and reported that lower bias occurred with large samples and small DIF size and that only small sample and medium DIF conditions led to acceptable Type I error rates.

Mismatching DIF Model and DIF Source

Studying differential item functioning means applying sometimes the “wrong” detection model to data whose real source of differential functioning remains to be discovered. The second phase of the research attempted to investigate the effects in these situations in order to understand the potential detriments. When fitting the DIF model to the other data, the false detection rates (percentages of DIF parameters in the mismatched model that gained statistically significant estimates) were mostly beyond the Bradley range. Large DIF even led to 100% false detection rates for both hypothesiswise and experimentwise errors using the Bonferroni and the Hochberge adjustment. Similarly, in their study on a Rasch DIF model, Paek and Wilson (2011) found highly inflated false detection rate for the non-DIF parameters defined in their Rasch DIF model, for which their supposition was underestimated standard deviation of the DIF parameter.

Recovery of other model parameters, including item locations, intercept, and component weights, was less than adequate in large DIF conditions. In particular, estimation of item locations was adversely affected by the differential effect in the data being fitted to.

When the DFFc model was applied to data with small DIF, the false detection rates were acceptable but the location parameter estimates were biased where the DIF effects were simulated. Interestingly, when applied to the DWF data, the recovery of item locations was acceptable but not the component weight, especially when the DIF was large. The DFFm model was not easy to use due to the high false detection rates when fitted to other DIF sources. Recovery of its item locations and component weights was unacceptable as the mismatched differential effects in the data led to more variation in their estimation. Fitting the DWF model to the DIF model resulted in acceptable false detection rates when DIF was small. Estimates of the item locations in the three mismatched data types were biased in large DIF conditions. However, the bias from component weights was less obvious than that from fitting the other models to mismatched data. It needs to be noted that recovery of the non-DIF model parameters when the model matched the data was largely acceptable.

In sum, the consequences of mismatching a proposed DIF model to a DIF source included high false detection rates and great error in model parameter estimation, which increased as the number of differential functioning items went up. Fitting a mismatched DIF model could lead to biased, misleading conclusions as to both detection and model validity.

Implications

Conventional DIF detection methods study one item at a time and can be quite laborious for a long test. More importantly, the commonly adopted DIF detection procedures may be incompatible with the unique, restrictive data structure the MIRID. The extended models as

proposed here account for these characteristics of within the framework of the generalized linear mixed models which can be conveniently implemented with general-purpose statistical packages.

The proposed models are also rooted in the framework of explanatory item response modeling, the strength of which lies with the fact it is a “one-step” approach: detecting, sizing up, and explaining differential functioning while estimating group difference all at once. By targeting different types of differential functioning, these extended MIRID models form a model-based approach to DIF investigation that is capable of separating likely construct irrelevant (i.e., adverse) DIF in individual items from probably construct pertinent (i.e., nonmalignant) DIF displayed in item groups (e.g., components, component weights, or item families). Detection of the former ensures test fairness and increases validity regarding group difference in item performance having controlled for the gap between the groups on trait level. Detection of the latter using the DFFc, DFFm, or DWF model naturally aids in its interpretation through properties shared by the group of items, and provides insight into group strength and weakness in terms of a domain of or a scenario within the trait being measured after accounting for group disparity in the primary construct. In order to utilize these advantages, this research assessed the efficacy of these models from different angles. Its results had implications for content-oriented applied researchers, who would be more interested to understand the MIRID nature of the data, and the methodologists, whose intentions are to come up with efficient ways to study DIF in the context of the MIRID.

Implications for Content Researchers

This study found that under the MIRID DIF model the unprotected hypothesiswise Type I error rate was not well maintained when DIF effect was trivial and that the unprotected experimentwise Type I error rates were always high. These results plus the detrimental outcome

from fitting a DIF model to data with a mismatched source of differential effects advised practitioners and content-oriented researchers against applying the DIF model up front in their study; rather, the first step is to have content experts conduct a substantive analysis (Paek & Fukuhara, 2014; Xie & Wilson, 2008) to identify potential items that may perform differentially on manifest groups of interest. In the simulation, when Type I errors were under control, good per-pair power in large sample and large delta conditions were observed for the DIF model. The researchers are thus encouraged, after the substantive analysis, to apply non-MIRID procedures, like the Mantel-Haenszel or the multiple-group CFA, as exploratory means in addition to the confirmatory DIF model. The target would be an agreement between the substantive analysis and the detection methods as to which items exhibit DIF. With sufficient but not massive sample sizes (1,000 ~ 3,000) a model-based but more parsimonious approach such as the Rasch DIF model may be considered so that all potential DIF items identified in the content analysis can be modeled at once.

In the simulation, detection performance of the proposed group-level DIF models was satisfactory. Therefore, if the pattern of the items exhibit DIF points to one particular fact, the MIRID DFFc, DFFm, or DWF can be fitted to the data. If the pattern does not give a clear distinction with an inclusive substantive analysis, for example, differential effects from a component and a component weight, the two models can be applied separately in order to determine the better-fitting model for explaining the extant DIF. It may be advisable to use information-based statistics such as Akaike's Information Criterion (AIC; Akaike, 1974) or Schwarz's Bayesian Information Criterion (BIC; Schwarz, 1978), both offering estimates of the relative differences between solutions. These statistics are appropriate for the extended MIRID models because of the maximum likelihood estimation they use. Small values of these indices

give indication of better model-data fit. Simulation studies on other IRT models indicated that BIC selects the correct model in general while for more complex model AIC fares better along with several other less known indices (Lee & Beretvas, 2014). However, there must be a sizeable sample (>500) before fitting an extended MIRID model according to the results here.

Implications for Methodology Researchers

Results on the DIF model suggest that the quality of model estimation was less than optimal, which had more to do with the model's inability to distinguish the two parts in the summation: the location and the DIF parameters. Methodologies could improve it as such: 1) obtain estimates of component item locations in a reduced model (the standard MIRID or the Rasch model), 2) deploy them as the starting values for the MIRID DIF model estimation, 3) even use the obtained estimates of the location and DIF parameters as the starting value of the next round of analysis, 4) stop when the estimated values stabilize.

The standard MIRID assumes strict underlying structure of the measured construct that is difficult to satisfy with empirical data. This inherent restrictiveness may be a reason for its being used less than other componential IRT models as well as the difficulty in estimation encountered in the simulation study. Various extensions have been proposed to generalize the model (see Chapter Two for details). For example, if the effect of a component is allowed to vary over people, that is, when a component weight is assumed as a random effect in order to account for individual differences in how they are affected by processes. In this case, the weight is decomposed into a mean and a variance parameter as extended from Equation 38.

$$\eta_{jmo} = \theta_j + G_j \gamma_g - \left(\sum_{k=1}^K (\omega_{jk} + G_j \delta_w) \beta_{mk} + \omega_0 \right). \quad (41)$$

where $\omega_{jk} = \theta_{jk} + \omega_k$ and the multivariate normal distribution is assumed for the vector of random effects. For example, if the first component weight is allowed to vary across persons, $\theta_j = (\theta_{j0}, \omega_{j1})$. Formulated like this, DIF investigation would consider if the mean of this random weight differs significantly between two manifest groups.

This simulation study calculated on absolute bias when evaluating parameter recovery. Very large absolute bias values were observed for the delta parameters in the MIRID DIF model. Another evaluative criterion, relative bias, was devised to answer the question “What would be an acceptable size of bias in an estimator after controlling for the magnitude of the parameter?” and has been reported by DIF studies. Hoogland and Boomsma (1998) formulated the relative bias of parameter estimators:

$$Bias(\hat{\theta}_i) = \frac{\bar{\theta}_i - \theta_i}{\theta_i}, \quad (42)$$

where $\bar{\theta}_i$ is the mean of the estimates of parameter θ for item i over all replications. They suggested an acceptability criterion for this index is its absolute value does not exceed .05.

Relative bias was reported in IRT model parameter recovery (e.g., Wang & Jin, 2010) and in DIF research (e.g., Chaimongkol, Huffer, & Kamata, 2006). Reporting relative bias in simulation results would change the outlook and interpretation of the findings. For methodology research, it may be advisable to report both absolute and relative bias.

In the findings on statistical inferences, there are quite incongruous outlooks when different Type I error rates and different adjustments were implemented in the DIF models. For example, under the DIF model, large delta conditions with the Bonferroni and Hochberg corrections resulted in PCER that were well below the nominal level but out-of-control EWER. As the most common index in DIF simulation research, the unprotected PCER is appropriate for

detection methods that test items individually. For other methods that assess multiple items at once, such as the DIF approach proposed in here, EWER is the more appropriate Type I errors as an evaluative criterion in a setting of multiple significance testing. Given the fact that the two correction procedures in this research were successful in avoiding Type I error inflation in a number of conditions, methodologists would benefit from incorporating them in their research. Or, they could explore other ways to adjust the critical value such as the Benjamini and Hochberg False Discovery Rate method (Benjamini & Hochberg, 1995; Kromrey & Hogarty, 2002) or the adjustment suggested by Oort (1992, 1998).

Lastly, for simulation research on complex models sample size matters: large sample sizes are necessary for reliable and accurate parameter estimation. For example, in a recent study of explanatory IRT model involving multiple predictors (Tay, Huang, & Vermunt, 2016), sample sizes were set at 5,000, 10,000, 20,000, and 40,000.

Limitations and Future Studies

Due to limited resources, performance of the proposed DIF models was not assessed in light of the easy-to-use, conventional methods such as the Mantel-Haenszel procedure, logistic regression, SIBTEST, etc. Future research may include such comparisons. Another direction is to experiment with some of the newly tested methods. For example, the forward procedure in the multiple-group categorical CFA approach (Kim & Yoon, 2011) could prove to be a useful alternative.

This dissertation was not able to look into the issue of non-uniform DIF and considered only single-source differential functioning. Future studies on the former can be conducted from the platform of the two-parameter MIRID (Wang & Jin, 2010). The latter, concurrent DIF sources, however, could be investigated from different angles. Xie and Wilson (2008)

demonstrated with the LLTM that in educational assessments more than one content domain can be incorporated in the same DIF model, which in the MIRID context would be analogous to modeling DFFc and DFFm at once, for instance. Obviously, there needs to be more simulation-based studies featuring model-comparison to evaluate this approach for its efficacy in various conditions.

Through a DIF decomposition perspective, Paek and Fukuhara (2014) showed that item-level and group-level (i.e., testlet) level DIF can be modeled simultaneously; but in the MIRID context modeling the two levels of differential functioning at once has yet to be experimented. Despite different modeling assumptions and purpose, cognitive diagnostic modeling (CDM) could provide another alternative to study DIF because the data structure in the MIRID could be seamlessly converted to a CDM Q-matrix where each attribute would encompass all items within one component and the “composite” attribute would have all items indexed as 1. Unfortunately, there is no parallel in CDM to an item family. The various techniques developed in the CDM DIF literature could be useful such as modeling attribute DIF with a higher order model.

The maximum likelihood estimation method as implemented with PROC NLMIXED was time consuming. Plus, a sizable sample and a large number of replications were necessary for consistent and trustworthy outcome in Monte Carlo simulation studies. As a consequence, the research design was constrained by time such that only three factors and two levels of variation were possible. Specifically, the design in this study evaluated only the large and small DIF levels but left unconsidered the medium DIF level, which arguably can be more relevant in DIF research. For example, a DIF magnitude around .4, a group difference in mean trait level of .4, and a sample size of 1,500. A possible follow-up of the research may involve more levels in

design factors and factors deployed as fixed in this research, such as test length in terms of numbers of item families and components, correlation between components, direction of DIF, etc.

Conclusions

A model-based approach to studying differential functioning of individual items and item bundles in the context of the model with internal restrictions on item difficulty (MIRID) was proposed and evaluated in this dissertation. In the simulation study, the group-level DIF models had good Type I error control and overall achieved excellent detection power across the conditions while the item-level DIF model maintained Type I error control in conditions of small DIF but failed to gain considerable power. Research on this topic should be continued, especially on detecting item-level DIF. For content-oriented practitioners, DIF study in the context of the MIRID must begin with a substantive analysis of potential DIF source; without it, misleading outcome may arise from applying a DIF model up front. It is more important to be able to interpret discovered group-level differential functioning, which through substantive analysis can be determined as either a nuisance or complementary dimension, secondary to the primary construct.

REFERENCES

- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.
- Abramowitz, M. & Stegun, I. (1974). *Handbook of Mathematical Functions*, New York: Dover Publications.
- Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.
- Agresti, A., Booth, J.G., Hobart, J.P. & Caffo, B. (2000). Random-effects modeling of categorical response data, *Sociological Methodology, 30*, 27–80.
- Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*, 251–269.
- Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications*, Amsterdam: North Holland Publishing Co.
- Banks, K. (2013). A Synthesis of the Peer-Reviewed Differential Bundle Functioning Research. *Educational Measurement: Issues and Practice, 32*(1), 43-55.
- Barrett, K. C. (1995). A functionalist approach to shame and guilt. In J. P. Tangney & K. W. Fischer (Eds.), *Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride* (pp. 25-63). New York: Guilford Press.
- Bechger, T., Verstralen, H., & Verhelst, N. (2002). Equivalent linear logistic test models. *Psychometrika, 67*, 123-136.
- Beguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some model fit analysis of multidimensional IRT models. *Psychometrika, 66*, 541–562.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.
- Bolt, D., & Stout, W. F. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23*, 67-95.
- Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

- Breslow, N.E. & Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 88 (421): 9–25.
- Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, 44, 131-155.
- Butter, R. P. (1994). *Item response models with internal restrictions on item difficulty*. Unpublished doctoral dissertation, Catholic University of Leuven, Belgium.
- Butter, R., De Boeck, P., & Verhelst, N. D. (1998). An item response model with internal restrictions on item difficulty. *Psychometrika*, 63, 1-17.
- Chaimongkol, S., Huffer, F. W., & Kamata, A. (2006). A Bayesian approach for fitting a random effect differential item functioning across group units. *Thailand Statistician*, 4, 27-41.
- Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, 33(3), 333-353.
- Christensen KB (2013). Conditional Maximum Likelihood Estimation in Polytomous Rasch Models using SAS." *ISRN Computational Mathematics*, Article ID 617475, 8 pages.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ.
- Congdon, P. (2003). *Applied Bayesian Modeling*. New York, NY: Wiley.
- Erlbaum. De Boeck. P. (1991). *Componential IRT models*. Unpublished manuscript, University of Leuven, Belgium.
- De Boeck, P., & Wilson, M. (Eds) (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer.
- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73, 533-559.
- De la Torre, J., Stark, S., & Chernyshenko, O. (2006). Markov Chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement*, 30, 216–232.
- Donoghue, J., Holland, P., & Thayer, D. (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: Erlbaum.
- Douglas, J., Roussos, L. A., & Stout, W. F. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement*, 33, 465-485.

- Embretson, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*, 479-494.
- Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and IRT likelihood ratio. *Applied Psychological Measurement*, *29*, 278-295.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.
- Fox, J. P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software*, *20*(5), 1-16.
- Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. Springer.
- Fox, J.-P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*(2), 271-288.
- Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement*, *35*(8), 604-622.
- Gilbert, P., Pehl, J., & Allan, S. (1994). The phenomenology of shame and guilt: An empirical investigation. *British Journal of Medical Psychology*, *67*, 23-36.
- Gierl, M.J., Gotzmann, A., & Boughton, K.A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education*, *17*, 241-264.
- Green, K. E., & Smith, R. S. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics*, *12*, 369-381.
- Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457-459.
- Han, K. T., & Hambleton, R. K. (2007). User's Manual: WinGen (*Center for Educational Assessment Report No. 642*). Amherst, MA: University of Massachusetts, School of Education.
- Hedeker, D. (1999). MIXNO: A computer program for mixed-effects nominal logistic regression. *Journal of Statistical Software*, *4*, 1-92.
- Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, *22*, 1433-1446.
- Hedeker, D., & Gibbons, R. D. (1996). MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, *49*, 157-176.

- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800–803.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.
- Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling an overview and a meta-analysis. *Sociological Methods & Research*, 26(3), 329-367.
- Hung, L. F. (2011). Formulation and Application of the Hierarchical Generalized Random-Situation Random-Weight MIRID. *Multivariate Behavioral Research*, 46(4), 643-668.
- Janssen, R., Hoskens, M., & De Boeck, P. (1993). An application of Embretson's multicomponent latent trait model to synonym tests. In R. Steyer, K.F. Wender, & K. F. Widaman (Eds.) *Psychometric methodology, proceedings of the 7th European meeting of the Psychometric Society in Trier* (pp. 87-190). Stuttgart: Gustaf Fischer Verlag.
- Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). New York: Springer.
- Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, 38(1), 32-60.
- Kachman, S. D. (2000). *An introduction to generalized linear mixed models*. In Proceedings of a symposium at the organizational meeting for a NCR coordinating committee on "Implementation Strategies for National Beef Cattle Evaluation," Athens (pp. 59-73).
- Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38(1), 79-93.
- Kamata, A. (April 2002). *Procedure to perform item response analysis by hierarchical generalized linear model*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- Kamata, A., & Cheong, Y. F. (2007). *Multilevel Rasch models*. In Multivariate and mixture distribution Rasch models (pp. 217-232). Springer: New York.
- Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, 31, 331-358.

- Kim, J. S., & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, 26(4), 38-51.
- Kim, D. H., & Huynh, H. (2010). Equivalence of paper-and-pencil and online administration modes of the statewide English test for students with and without disabilities. *Educational Assessment*, 15(2), 107–121.
- Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, 18(2), 212-228.
- Kim, E. S., Yoon, M., & Lee, T. (2011). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, 72(3), 469-492.
- Kromrey, J.D., & La Rocca, M.A. (1995). Power and Type I error rates of new pairwise multiple comparison procedures under heterogeneous variances. *Journal of Experimental Education*, 63, 343-362.
- Kromrey J. D., Hogarty K. Y. (2002). *FDR_TEST: A SAS Macro for Calculating New Methods of Error Control in Multiple Hypothesis Testing*. Savannah, GA: Southeast SAS Users Group.
- Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* 38, 963–974.
- Lee, H., & Beretvas, S. N. (2014). Evaluation of two types of Differential Item Functioning in factor mixture models with binary outcomes. *Educational and Psychological Measurement*, 74(5), 831-858.
- Lee, Y., & Wilson, M. (2009). *An extension of the MIRID model for polytomous responses and random effects*. Paper presented at the annual meeting of American Educational Research Association, San Diego.
- Lee, Y. (2010). *Random Item Modeling: An Extension and Generalization of MIRID models*. Unpublished doctoral dissertation, University of California, Berkeley.
- Li, F., Cohen, A. S., Kim, S., & Cho, S. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, 33, 353-373.
- Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–687.
- Maris, G., & Bechger, T. M. (2004). Equivalent MIRID models. *Psychometrika*, 69(4), 627-639.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

- Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, 52, 443-451.
- McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*, 2nd edition. Boca Raton: Chapman and Hall/CRC.
- McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.
- Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, 115(2), 300.
- Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65(3), 391-411.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM-algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus. The comprehensive modeling program for applied researchers: User's guide*, 5.
- Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement*, 30, 293-311.
- Narayanan, P., & Swaminathan H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, 20, 257 - 274.
- Natesan, P., Limbers, C., & Varni, J. W. (2010). Bayesian estimation of graded response multilevel models using Gibbs sampling: formulation and illustration. *Educational and Psychological Measurement* 70(3),420-439.
- Naylor, J. C. & Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, 31, 214-225.
- Nelder, J. & Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General) (Blackwell Publishing)* 135 (3): 370-384.
- Nixon, C. B. (2013). *Issues inherent in detecting attribute-level DIF using item-level methods*. Unpublished doctoral dissertation, University of Georgia..
- Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, 6, 150-166.
- Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, 5, 107-124.
- Paek, I. (2002). *Investigations of differential item functioning: comparisons among approaches, and extension to a multidimensional context*. Unpublished doctoral dissertation. University of California, Berkeley.

- Paek, I., & Wilson, M. (2011). Formulating the rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*(6), 1023-1046.
- Paek, I., & Fukuhara, H. (2014). Estimating a DIF decomposition model using a random-weights linear logistic test model approach. *Behavior research methods, 47*(3), 890-901.
- Pinheiro, P. C. & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics, 4*, 12-35.
- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*, 167-190.
- Rasbash, J., Browne, W. J., Goldstein, H., Yang, M., Plewis, I., Healy, M., et al. (2000). *A User's Guide to MLwiN (Version 2.1)* [Computer software and manual]. London: Institute of Education, University of London.
- Raudenbush, S. W. & A. S. Bryk. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods. 2nd edition*. Thousand Oaks CA: Sage.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y.-F., & Congdon, R. (2004). *HLM 6: Hierarchical Linear and Nonlinear Modeling* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.
- Rijmen, F., Tuerlinckx, F. De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*, 185-205.
- Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283-288.
- Rogers, S. J., & Swaminathan H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.
- Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.
- Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin, 56*, 26-A1.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17*, 1-100.

- SAS Institute Inc. (2011). *SAS/STAT® 9.3 User's Guide*. Cary, NC: SAS Institute Inc.
- Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin, 110*, 577-586.
- Shealy, R., & Stout, W. F. (1993a). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Erlbaum.
- Shealy, R., & Stout, W. F. (1993b). 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.
- Singer, J. D. (1998). "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." *Journal of Educational and Behavioral Statistics, 24*:323—355.
- Sinharay, S., Dorans, N. J., Grant, M. C., & Blew, E. O. (2009). Using past data to enhance small sample DIF estimation: A Bayesian approach. *Journal of Educational and Behavioral Statistics, 34*, 74-96.
- Smits, D. J. M., & De Boeck, P. (2003). A componential IRT model for guilt. *Multivariate Behavioral Research, 38*, 161-188.
- Smits, D. J. M., De Boeck, P., Verhelst, N. D., & Butter, R. (2001). *The MIRID program (version 1.0)* [Computer program manual]. , K. U. Leuven, Belgium.
- Smits, D. J. M., De Boeck, P., & Verhelst, N. D. (2003). Estimation of the MIRID: A program and a SAS based approach. *Behavior Research Methods, Instruments, and Computers, 35*, 537-549.
- Smits, D. J. M., De Boeck, P., & Vansteelandt, K. (2004). The inhibition of verbally aggressive behaviour. *European Journal of Personality, 18*, 537-555.
- Smits, D. J. M., & Moore, S. (2004). Latent item predictors with fixed effects. In P. De Boeck & M. R. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 267-287). New York: Springer-Verlag.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, 64*, 583-616.
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS version 1.4* [computer program]. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health.

- Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. Boca Raton: CRC Press.
- Tangney, J. P. (1995). Shame and guilt in interpersonal relationships. In J. P. Tangney & K. W. Fischer (Eds.), *Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride* (pp. 114-139). New York: Guilford Press.
- Tay, L., Huang, Q., & Vermunt, J. K. (2016). Item response theory with covariates (IRT-C): Assessing item recovery and differential item functioning for the three-parameter logistic model. *Educational and Psychological Measurement, 76*, 22-42.
- Tuerlinckx, F. & Wang, W. C. (2004). Models for polytomous data. In P. De Boeck & M. R. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 75-109). New York: Springer-Verlag
- Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Van den Noortgate, W., Meulders, M., & De Boeck, P. (2004). Estimation and software. In P. De Boeck & M. R. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 343-373). New York: Springer-Verlag.
- Vansteelandt, K. (2000). *Formal models for contextualized personality psychology*. Unpublished doctoral dissertation, K. U. Leuven, Belgium.
- Verhelst, N. D, Glas, C. A. W., & Verstralen, H. H. F. M. (1994). *One parameter logistic model* [Computer program and manual]. Arnhem, The Netherlands: CITO.
- Wainer, H. (1993). Measuring differential impact of items. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.3-23). Hillsdale, NJ: Erlbaum.
- Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*, 197-219.
- Wang, W. C., Wilson, M., & Adams, R.J. (1997). Rasch models for multidimensionality between and within items. In M. Wilson and G. Engelhard, (Eds.), *Objective measurement: Theory into Practice. Vol IV*. Norwood, NJ: Ablex.
- Wang, W. C., & Jin, K. Y. (2010a). A generalized model with internal restrictions on item difficulty for polytomous items. *Educational and Psychological Measurement, 70*(2), 181-198.
- Wang, W. C., & Jin, K. Y. (2010b). Multilevel, two-parameter, and random-weights generalizations of a model with internal restrictions on item difficulty. *Applied Psychological Measurement, 34*(1), 46-65.
- Wang, W. C., Wilson, M. R., & Shih, C.-L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement, 43*, 335-353.

- Wilson, M. R. (2003). On choosing a model for measuring. *Methods of Psychological Research*, 8(3), 1-22.
- Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: an international testing context. *Psychology Science*, 50(3), 403.
- Zhang, W. (2007). *Detecting differential item functioning using the DINA model*. Unpublished doctoral dissertation, University of North Carolina at Greensboro.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). Hillsdale, NJ: Erlbaum.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
- 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.

APPENDIX A:

EXAMPLES OF ANALYSIS CODE IN SAS

```
title 'DIF model';
proc nlmixed data=dif method=gauss qpoints=15 noad
  technique=quanew maxfunc=5000 ;
  parms b1-b30=0 w0-w3=.3 d11-d20=0 gamma=0 sd=1;
  beta1=b1*x1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6+b7*x7+b8*x8+b9*x9+b10*x10;
beta2=b11*x11+b12*x12+b13*x13+b14*x14+b15*x15+b16*x16+b17*x17+b18*x18+b19*x19
+b20*x20;
beta3=b21*x21+b22*x22+b23*x23+b24*x24+b25*x25+b26*x26+b27*x27+b28*x28+b29*x29
+b30*x30;
delta2=d11*x11+d12*x12+d13*x13+d14*x14+d15*x15+d16*x16+d17*x17+d18*x18+d19*x1
9+d20*x20;
  ex=exp(theta+gamma*grp-(1-co)*(beta1+beta2+beta3+delta2*grp)
-co*(w0+w2*(beta2+delta2*grp)+w1*beta1+w3*beta3));
  p=ex/(1+ex);
  model y ~ binary(p);
  random theta ~ normal(0,sd*sd) subject=person;
  estimate 'sd**2' sd*sd;
run;
```

```
title 'DFFc model ';
proc nlmixed data=dffc method=gauss qpoints=15 noad
  technique=quanew maxfunc=5000 ;
  parms b1-b30=0 w0-w3=.3 kd1-kd3=0 gamma=0 sd=1;
  beta1=b1*x1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6+b7*x7+b8*x8+b9*x9+b10*x10;
beta2=b11*x11+b12*x12+b13*x13+b14*x14+b15*x15+b16*x16+b17*x17+b18*x18+b19*x19
+b20*x20;
beta3=b21*x21+b22*x22+b23*x23+b24*x24+b25*x25+b26*x26+b27*x27+b28*x28+b29*x29
+b30*x30;
  kd=kd1*k1+kd2*k2+kd3*k3;
  ex=exp(theta+gamma*grp-(1-co)*(beta1+beta2+beta3+kd*grp)
-co*(w0+w1*beta1+w2*beta2+w3*beta3+kd*grp));
  p=ex/(1+ex);
  model y ~ binary(p);
  random theta ~ normal(0,sd*sd) subject=person;
```

```
estimate 'sd**2' sd*sd;
run;
```

```
title 'DFFm model ';
proc nlmixed data=dffm method=gauss qpoints=15 noad
  technique=quanew maxfunc=5000 ;
  parms b1-b30=0 w0-w3=.3 fd1-fd10=0 gamma=0 sd=1;
  beta1=b1*x1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6+b7*x7+b8*x8+b9*x9+b10*x10;
  beta2=b11*x11+b12*x12+b13*x13+b14*x14+b15*x15+b16*x16+b17*x17+b18*x18+b19*x
19+b20*x20;
  beta3=b21*x21+b22*x22+b23*x23+b24*x24+b25*x25+b26*x26+b27*x27+b28*x28+b29*x
29+b30*x30;
  fd=fd1*m1+fd2*m2+fd3*m3+fd4*m4+fd5*m5+fd6*m6+fd7*m7+fd8*m8+fd9*m9+fd10*m
10;
  ex=exp(theta+gamma*grp-(1-co)*(beta1+beta2+beta3+fd*grp)
-co*(w0+w1*beta1+w2*beta2+w3*beta3+3*fd*grp));
  p=ex/(1+ex);
  model y ~ binary(p);
  random theta ~ normal(0,sd*sd) subject=person;
  estimate 'sd**2' sd*sd;
run;
```

```
title 'DWF model ';
proc nlmixed data=dwf method=gauss qpoints=15 noad
  technique=quanew maxfunc=5000 ;
  parms b1-b30=0 w0-w3=.3 wd1-wd3=0 gamma=0 sd=1;
  beta1=b1*x1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6+b7*x7+b8*x8+b9*x9+b10*x10;
  beta2=b11*x11+b12*x12+b13*x13+b14*x14+b15*x15+b16*x16+b17*x17+b18*x18+b19*x
19+b20*x20;
  beta3=b21*x21+b22*x22+b23*x23+b24*x24+b25*x25+b26*x26+b27*x27+b28*x28+b29*x
29+b30*x30;
  ex=exp(theta+gamma*grp-(1-co)*(beta1+beta2+beta3)
-co*(w0+(w1+wd1*grp)*beta1+(w2+wd2*grp)*beta2+(w3+wd3*grp)*beta3));
  p=ex/(1+ex);
  model y ~ binary(p);
  random theta ~ normal(0,sd*sd) subject=person;
  estimate 'sd**2' sd*sd;
run;
```


APPENDIX B:

**ESTIMATION BIAS AND RMSES OF THE ZERO-VALUE DIF PARAMETERS OF
THE MIRID DIF, DFFC, DFFM, AND DWF MODELS**

Table B1: Average Bias and RMSEs of the Zero-value Delta Parameter Estimates under the
MIRID DIF Model

	#	Bias	Bias_sd	RMSE	RMSE_sd		
delta=.2 N=250*2	Impact=0	26	-0.022	0.013	0.213	0.023	
	Impact=-.7	26	-0.020	0.016	0.220	0.030	
	N=1500*2	Impact=0	26	-0.029	0.013	0.093	0.008
		Impact=-.7	26	-0.025	0.012	0.092	0.011
delta=.7 N=250*2	Impact=0	26	-0.095	0.051	0.242	0.020	
	Impact=-.7	26	-0.078	0.047	0.229	0.029	
	N=1500*2	Impact=0	26	-0.106	0.042	0.142	0.027
		Impact=-.7	26	-0.088	0.038	0.131	0.022

Table B2: Average Bias and RMSEs of the Zero-value Delta Parameter Estimates under the
MIRID DFFc Model

	#	Bias	Bias_sd	RMSE	RMSE_sd		
delta=.2 N=250*2	Impact=0	2	0.001	0.002	0.066	0.000	
	Impact=-.7	2	-0.004	0.001	0.068	0.004	
	N=1500*2	Impact=0	2	0.001	0.001	0.028	0.000
		Impact=-.7	2	-0.001	0.000	0.028	0.000
delta=.7 N=250*2	Impact=0	2	-0.002	0.001	0.069	0.000	
	Impact=-.7	2	-0.001	0.000	0.068	0.002	
	N=1500*2	Impact=0	2	-0.001	0.001	0.028	0.001
		Impact=-.7	2	-0.001	0.001	0.029	0.002

Table B3: Average Bias and RMSEs of the Zero-value Delta Parameter Estimates under the MIRID DFFm Model

	#	Bias	Bias_sd	RMSE	RMSE_sd		
delta=.2 N=250*2	Impact=0	8	0.000	0.002	0.067	0.002	
	Impact=-.7	8	-0.002	0.003	0.068	0.003	
	N=1500*2	Impact=0	8	0.000	0.001	0.026	0.001
		Impact=-.7	8	-0.001	0.001	0.027	0.002
delta=.7 N=250*2	Impact=0	8	0.000	0.003	0.066	0.004	
	Impact=-.7	8	0.002	0.002	0.067	0.004	
	N=1500*2	Impact=0	8	-0.002	0.001	0.026	0.002
		Impact=-.7	8	0.001	0.001	0.028	0.002

Table B4: Average Bias and RMSEs of the Zero-value Delta Parameter Estimates under the MIRID DWF Model

	#	Bias	Bias_sd	RMSE	RMSE_sd		
delta=.2 N=250*2	Impact=0	2	0.003	0.001	0.059	0.000	
	Impact=-.7	2	0.004	0.001	0.055	0.011	
	N=1500*2	Impact=0	2	0.001	0.000	0.025	0.006
		Impact=-.7	2	0.001	0.003	0.025	0.000
delta=.7 N=250*2	Impact=0	2	0.000	0.003	0.073	0.003	
	Impact=-.7	2	0.013	0.014	0.095	0.034	
	N=1500*2	Impact=0	2	-0.001	0.000	0.027	0.001
		Impact=-.7	2	0.000	0.000	0.029	0.006

APPENDIX C:

**ESTIMATION BIAS AND RMSES OF THE MODEL PARAMETERS OF THE MIRID
DFFC, DFFM, AND DWF MODELS**

Table C1: Average Bias and RMSEs of Item Location Parameter Estimates under the MIRID

DIF Model

			#	Bias	Bias_sd	RMSE	RMSE_sd
delta=.2	N=250*2	Impact=0	30	-0.004	0.026	0.119	0.014
		Impact=-.7	30	-0.014	0.024	0.121	0.015
	N=1500*2	Impact=0	30	0.028	0.024	0.057	0.016
		Impact=-.7	30	-0.008	0.023	0.053	0.008
delta=.7	N=250*2	Impact=0	30	-0.012	0.084	0.138	0.038
		Impact=-.7	30	0.009	0.075	0.131	0.041
	N=1500*2	Impact=0	30	0.023	0.082	0.073	0.063
		Impact=-.7	30	-0.001	0.076	0.076	0.048

Table C2: Average Bias and RMSEs of the Intercept and Component Weights Estimates under the MIRID DIF Model

			#	Bias	Bias_sd	RMSE	RMSE_sd
delta=.2	N=250*2	Impact=0	4	0.000	0.001	0.047	0.011
		Impact=-.7	4	-0.001	0.001	0.044	0.011
	N=1500*2	Impact=0	4	0.000	0.000	0.017	0.002
		Impact=-.7	4	0.000	0.001	0.019	0.002
delta=.7	N=250*2	Impact=0	4	0.002	0.002	0.036	0.004
		Impact=-.7	4	0.001	0.004	0.048	0.012
	N=1500*2	Impact=0	4	0.000	0.000	0.016	0.002
		Impact=-.7	4	0.001	0.001	0.017	0.003

Table C3: Average Bias and RMSEs of Item Location Parameter Estimates under the MIRID

DFFc Model

			#	Bias	Bias_sd	RMSE	RMSE_sd
delta=.2	N=250*2	Impact=0	30	0.002	0.004	0.114	0.009
		Impact=-.7	30	0.013	0.006	0.113	0.012
	N=1500*2	Impact=0	30	0.009	0.002	0.047	0.004
		Impact=-.7	30	0.009	0.002	0.048	0.005
delta=.7	N=250*2	Impact=0	30	-0.020	0.007	0.112	0.009
		Impact=-.7	30	-0.019	0.006	0.117	0.014
	N=1500*2	Impact=0	30	0.001	0.002	0.045	0.004
		Impact=-.7	30	0.000	0.003	0.048	0.005

Table C4: Average Bias and RMSEs of the Intercept and Component Weights Estimates under

the MIRID DFFc Model

			#	Bias	Bias_sd	RMSE	RMSE_sd
delta=.2	N=250*2	Impact=0	4	0.000	0.002	0.045	0.011
		Impact=-.7	4	0.001	0.002	0.043	0.009
	N=1500*2	Impact=0	4	0.000	0.001	0.019	0.004
		Impact=-.7	4	0.000	0.001	0.017	0.004
delta=.7	N=250*2	Impact=0	4	-0.001	0.002	0.052	0.009
		Impact=-.7	4	0.004	0.003	0.060	0.022
	N=1500*2	Impact=0	4	0.000	0.001	0.018	0.004
		Impact=-.7	4	0.000	0.000	0.020	0.004

Table C5: Average Bias and RMSEs of Item Location Parameter Estimates under the MIRID

DFFm Model

			#	Bias	Bias_sd	RMSE	RMSE_sd
delta=.2	N=250*2	Impact=0	30	-0.019	0.005	0.117	0.008
		Impact=-.7	30	-0.003	0.006	0.119	0.014
	N=1500*2	Impact=0	30	0.001	0.002	0.047	0.004
		Impact=-.7	30	0.005	0.002	0.049	0.005
delta=.7	N=250*2	Impact=0	30	-0.038	0.007	0.123	0.011
		Impact=-.7	30	0.021	0.006	0.121	0.018
	N=1500*2	Impact=0	30	-0.018	0.002	0.050	0.004
		Impact=-.7	30	0.000	0.003	0.048	0.006

Table C6: Average Bias and RMSEs of the Intercept and Component Weights Estimates under the MIRID DFFm Model

			#	Bias	Bias_sd	RMSE	RMSE_sd
delta=.2	N=250*2	Impact=0	4	-0.002	0.003	0.051	0.018
		Impact=-.7	4	0.001	0.001	0.042	0.009
	N=1500*2	Impact=0	4	0.000	0.000	0.020	0.005
		Impact=-.7	4	0.001	0.001	0.021	0.005
delta=.7	N=250*2	Impact=0	4	0.001	0.003	0.057	0.015
		Impact=-.7	4	0.000	0.002	0.055	0.008
	N=1500*2	Impact=0	4	0.001	0.002	0.018	0.004
		Impact=-.7	4	0.000	0.001	0.021	0.004

Table C7: Average Bias and RMSEs of Item Location Parameter Estimates under the MIRID

DWF Model

			#	Bias	Bias_sd	RMSE	RMSE_sd
delta=.2	N=250*2	Impact=0	30	0.005	0.007	0.113	0.011
		Impact=-.7	30	-0.001	0.005	0.111	0.010
	N=1500*2	Impact=0	30	-0.008	0.001	0.045	0.005
		Impact=-.7	30	-0.010	0.002	0.046	0.004
delta=.7	N=250*2	Impact=0	30	-0.001	0.005	0.110	0.012
		Impact=-.7	30	-0.020	0.006	0.112	0.009
	N=1500*2	Impact=0	30	-0.003	0.003	0.044	0.004
		Impact=-.7	30	0.003	0.002	0.044	0.005

Table C8: Average Bias and RMSEs of the Intercept and Component Weights Estimates under the MIRID DWF Model

			#	Bias	Bias_sd	RMSE	RMSE_sd
delta=.2	N=250*2	Impact=0	4	-0.001	0.003	0.043	0.005
		Impact=-.7	4	-0.001	0.002	0.045	0.008
	N=1500*2	Impact=0	4	0.000	0.001	0.020	0.003
		Impact=-.7	4	0.000	0.001	0.029	0.008
delta=.7	N=250*2	Impact=0	4	-0.001	0.003	0.062	0.005
		Impact=-.7	4	-0.002	0.004	0.063	0.015
	N=1500*2	Impact=0	4	0.000	0.001	0.021	0.003
		Impact=-.7	4	-0.001	0.001	0.022	0.003

APPENDIX D:

**TYPE I ERROR RATES AND POWER OBTAINED FROM ESTIMATING ITEM DIF
PARAMETERS BY COMPONENT AND BY ITEM**

Table D1: Type I Error Rates for the MIRID DIF Model Obtained from Estimating by
Component and by Item

			Per-comparison %			Experimentwise %			
			Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg	
		by							
delta=.2	N=250*2	Component	0.051	0.002	0.002	0.738	0.040	0.040	
		Impact=0	Item	0.053	0.001	0.001	0.782	0.036	0.036
		Component	0.049	0.002	0.002	0.710	0.038	0.038	
		Impact=-.7	Item	0.048	0.002	0.002	0.728	0.040	0.040
	N=1500*2	Component	0.065	0.002	0.002	0.846	0.058	0.058	
		Impact=0	Item	0.067	0.002	0.002	0.852	0.058	0.060
		Component	0.062	0.002	0.002	0.814	0.050	0.050	
		Impact=-.7	Item	0.062	0.002	0.002	0.818	0.058	0.060
delta=.7	N=250*2	Component	0.079	0.004	0.005	0.882	0.108	0.110	
		Impact=0	Item	0.081	0.004	0.004	0.924	0.104	0.106
		Component	0.069	0.004	0.004	0.852	0.088	0.088	
	N=1500*2	Impact=-.7	Item	0.071	0.004	0.004	0.872	0.092	0.094
		Impact=0	Component	0.253	0.040	0.045	1.000	0.686	0.702
		Component							

		Per-comparison %			Experimentwise %		
by		Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg
	Item	0.264	0.040	0.044	1.000	0.666	0.712
	Component	0.192	0.025	0.028	1.000	0.494	0.522
Impact=-.7	Item	0.211	0.029	0.033	0.094	0.218	5.492

Table D2: Power for the MIRID DIF Model Obtained from Estimating by Component and by

Item

			Per-comparison %			Experimentwise %			
			Un-protect.	Bonfer-roni	Hoch-berg	Un-protect.	Bonfer-roni	Hoch-berg	
			by						
delta=.2	N=250*2	Impact=0	Comp onent	0.121	0.011	0.011	0.012	0.061	0.000
			Item Comp onent	0.128	0.013	0.013	0.013	0.071	0.000
		Impact=-.7	Comp onent	0.123	0.007	0.007	0.007	0.067	0.000
			Item Comp onent	0.128	0.007	0.007	0.008	0.069	0.000
	N=1500* 2	Impact=0	Comp onent	0.505	0.119	0.120	0.145	0.347	0.036
			Item Comp onent	0.551	0.144	0.144	0.178	0.386	0.070
		Impact=-.7	Comp onent	0.483	0.115	0.115	0.144	0.328	0.052
			Item	0.523	0.137	0.139	0.172	0.359	0.046
delta=.7	N=250*2	Impact=0	Comp onent	0.805	0.391	0.398	0.520	0.692	0.424
			Item Comp onent	0.831	0.446	0.452	0.584	0.756	0.490
		Impact=-.7	Comp onent	0.754	0.341	0.347	0.437	0.638	0.268
			Item Comp onent	0.788	0.359	0.364	0.468	0.655	0.344
	N=1500* 2	Impact=0	Comp onent	1.000	1.000	1.000	1.000	1.000	1.000
			Item Comp onent	1.000	1.000	1.000	1.000	1.000	1.000
		Impact=-.7	Comp onent	1.000	0.987	0.989	0.995	0.998	0.998
			Item	1.000	0.990	0.991	0.998	0.999	0.998