ABSTRACT

Title of Documents	DIFFERENT APPROACHES TO COVARIATE INCLUSION IN THE MIXTURE RASCH MODEL
	Tongyun Li, Doctor of Philosophy, 2014
Directed By:	Dr. Hong Jiao Measurement, Statistics and Evaluation Department of Human Development and Quantitative Methodology

The present dissertation project investigates different approaches to adding covariates and the impact in fitting mixture item response theory (IRT) models. Mixture IRT models serve as an important methodology for tackling several important psychometric issues in test development, including detecting latent differential item functioning (DIF). A Monte Carlo simulation study is conducted in which data generated according to a two-class mixture Rasch model (MRM) with both dichotomous and continuous covariates are fitted to several MRMs with misspecified covariates to examine the effects of covariate inclusion on model parameter estimation. In addition, both complete response data and incomplete response data with different types of missingness are considered in the present study in order to simulate practical assessment settings. Parameter estimation is carried out within a Bayesian framework vis-à-vis Markov chain Monte Carlo (MCMC) algorithms. Two empirical examples using the Programme for International Student Assessment (PISA) 2009 U.S. reading assessment data are presented to demonstrate the impact of different specifications of covariate effects for an MRM in real applications.

DIFFERENT APPROACHES TO COVARIATE INCLUSION
IN THE MIXTURE RASCH MODEL

By

Tongyun Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:

Professor Hong Jiao, Chair
Professor George B. Macready
Professor Jeffrey R. Harring
Professor Xin He
Professor Matthias von Davier
Professor Margaret J. McLaughlin, Dean’s Representative

Dedication

To my beloved parents,

Yikang Li and Yan Shen,

and my grandparents,

who love and support me every step of the way.

And also to my husband,

for your love and understanding.

Acknowledgements

First I would like to express my sincere gratitude to Dr. Hong Jiao, my academic advisor and dissertation chair, for your support and guidance throughout my doctoral study. Without your thoughtful insight and valuable advice, I would not be able to complete my coursework and dissertation research in a timely manner. Your vision and wisdom have been and will continue to be the guiding light in my future professional development.

I also would like to thank my other committee members, Dr. George B. Macready, Dr. Jeffrey Harring, Dr. Xin He, Dr. Matthias von Davier and Dr. Margaret J. McLaughlin, for your support, encouragement and mentorship. Each of you has played an important role in my academic development. My thanks especially go to Dr. George B. Macready, my former advisor in the master's study. Your insightful mentorship and warm encouragement helped me find a clear track early on for my later graduate study and research. My special gratitude also goes to Dr. Matthias von Davier, my internship mentor at ETS, for giving me guidance and hands-on experience of large-scale assessment. These enabled me to gain a deeper understanding of psychometric research in real assessment settings.

Finally, I am grateful for all the other faculty members and my graduate colleagues in the EDMS program. Thank you for providing me support during my years of graduate study at the University of Maryland.

Table of Contents

Dedication……………………………………………………………………………………ii

Acknowledgements…………………………………………………………………………iii

Table of Contents……………………………………………………………………………iv

List of Tables………………………………………………………………………………..vi

List of Figures………………………………………………………………………………viii

Chapter 1 Introduction…………………………………………………………………….……1

1.1 Statement of the Problem…………………………………………………….……..1

1.2 The Purpose of the Study…………………………………………………….……..3

1.3 Significance of the Study…………………………………………………….……..4

1.4 Overview of the Chapters………………………………………………………..….6

Chapter 2 Literature Review………………………………………………………………….8

2.1 Latent Variable Modeling Framework………………………………..………………9

2.1.1 Item response theory (IRT) and latent class analysis (LCA)………………10
2.1.2 General diagnostic model (GDM)…………………………………………15
2.1.3 Mixture IRT models………………………………………………………..17

2.2 The Rationale of Including Covariates………………………………………………21

2.2.1 Covariate inclusion in LCA………………………………………………..23
2.2.2 Covariate inclusion in non-mixture IRT models…………………………..25
2.2.3 Covariate inclusion in mixture IRT models………………………………..30
2.2.4 Other relevant modeling approaches………………………………………31

2.3 Model Estimation……………………………………………………………………..32

2.3.1 Introduction to Bayesian inference………………………………………...35
2.3.2 Convergence diagnosis……………………………………………………..38

2.4 Summary……………………………………………………………………………..41

Chapter 3 Methodology……………………………………………………………………….43

3.1 Different Approaches to Covariate Inclusion in the Mixture Rasch Model (MRM)...43

3.2 WinBUGS Implementation of Model Parameter Estimation……………………......47

3.3 Simulation Design……………………………………………………………………51

3.3.1 Fixed factors………………………………………………………………..51
3.3.2 Manipulated factors………………………………………………………..53
3.3.3 Evaluation criteria…………………………………………………………..60

3.4 Real Data Examples…………………………………………………………………..63

Chapter 4 Results…………………….……………………………………………………..65

4.1 Results of the Simulation Study……………………………………………………65

4.1.1 Latent group classification…………………………………………………67
4.1.2 Model parameter recovery…………………………………………………75
4.1.3 Overall model fit indices…………………………………………………115

4.2 Empirical Examples: PISA 2009 U.S. Reading……………………………………125

4.2.1 Sample 1…………………………………………………………………..126
4.2.2 Sample 2…………………………………………………………………..129

Chapter 5 Discussion……………………………………………………………………...132

5.1 Discussion of the Simulation Results………………………………………………132

5.1.1 Different approaches to covariate inclusion……………………………133
5.1.2 Effects of the other manipulated factors…………………………………135
5.1.3 Model selection……………………………………………………………137

5.2 Applications of Covariate Inclusion…………………………………………………139

5.3 Limitations and Future Direction……………………………………………………141

Appendix A…………………………………………………………………………………..143

Appendix B…………………………………………………………………………………..162

Appendix C…………………………………………………………………………………..163

References……………………………………………………………………………………164

List of Tables

Table 2.1. Explanatory Rasch models…………………………….……………………...27

Table 3.1. True data-generating model and alternative models…………………………………47

Table 3.2. The hierarchy of sampling methods used by WinBUGS……………………………..49

Table 3.3. Factors that are fixed at a single level in the simulation……………………………...52

Table 3.4. Manipulated factors in the simulation………………………………………………..53

Table 4.1. Variable and model name abbreviations……………………………………………...65

Table 4.2. The descriptive statistics of correct classification rate by manipulated factors………68

Table 4.3. ANOVA results of manipulated factors on the correct classification rate…………...70

Table 4.4. The descriptive statistics of item parameter by manipulated factors…………………77

Table 4.5a. Statistically significant ANOVA results of manipulated factors on the SE of item parameters………………………………………………………………………………………78

Table 4.5b. Statistically significant ANOVA results of manipulated factors on the RMSE of item parameters……………………………………………………………………………………..79

Table 4.5c. Statistically significant ANOVA results of manipulated factors on the 95% coverage of item parameters……………………………………………………………………………80

Table 4.6. The descriptive statistics of person parameter by manipulated factors...…………….85

Table 4.7a. Statistically significant ANOVA results of manipulated factors on the bias of person parameters……………………………………………………………………………………..86

Table 4.7b. Statistically significant ANOVA results of manipulated factors on the SE of person parameters……………………………………………………………………………………..88

Table 4.7c. Statistically significant ANOVA results of manipulated factors on the RMSE of person parameters………………………………………………………………………………88

Table 4.7d. Statistically significant ANOVA results of manipulated factors on the 95% Coverage of person parameters……………………………………………………………………………88

Table 4.8. The descriptive statistics of linear regression parameter (linking C_j with θ_{jg}) by manipulated factors……………………………………………………………………………92

Table 4.9a. Statistically significant ANOVA results of manipulated factors on the bias of $\alpha_{0.}$ ...93

Table 4.9b. Statistically significant ANOVA results of manipulated factors on the SE of $\alpha_{0.}$ 95

Table 4.9c. Statistically significant ANOVA results of manipulated factors on the RMSE of $\alpha_{0.}$ ...……………………………………………………………………………………………95

Table 4.10a. Statistically significant ANOVA results of manipulated factors on the bias of $\alpha_{1.}$ .97

Table 4.10b. Statistically significant ANOVA results of manipulated factors on the SE of $\alpha_{1.}$ ...97

Table 4.10c. Statistically significant ANOVA results of manipulated factors on the RMSE of $\alpha_{1\cdot}$……………………………………………………………………………………...98

Table 4.11. The descriptive statistics of logistic regression parameter by manipulated factors..102

Table 4.12a. Statistically significant ANOVA results of manipulated factors on the bias of β_{02}……………………………………………………………………………………103

Table 4.12b. Statistically significant ANOVA results of manipulated factors on the SE of β_{02}……………………………………………………………………………………105

Table 4.12c. Statistically significant ANOVA results of manipulated factors on the RMSE of β_{02}…………………………………………………………………………..…………106

Table 4.13a. Statistically significant ANOVA results of manipulated factors on the bias of β_{12}……………………………………………………………………………………108

Table 4.13b. Statistically significant ANOVA results of manipulated factors on the SE of β_{12}……………………………………………………………………………………109

Table 4.13c. Statistically significant ANOVA results of manipulated factors on the RMSE of β_{12}……………………………………………………………………………………110

Table 4.14a. Model selection frequency for simulation cells…………………….…………….116

Table 4.14b. Model selection decision by simulation condition and its frequency……………118

Table 4.15. Model fit Indices based on Sample 1……………….……………………………...127

Table 4.16. Sample 1 characteristics based on parameter estimates…………………….……...127

Table 4.17. Model fit Indices based on sample 2………………….…………………………...130

Table 4.18. Sample 2 characteristics based on parameter estimates…………………….……...130

List of Figures

Figure 2.1. The person explanatory model……………………………………………………28

Figure 2.2. The item explanatory model………………………………………………………...29

Figure 2.3. The doubly explanatory model…………………………………………………30

Figure 2.4. An example of within-chain label-switching………………………………………..40

Figure 2.5. An example of no mixting between chains…………………………………………40

Figure 2.6. An example of converged Markov chains…………………………………………41

Figure 3.1. The mixture Rasch model…………………………………………………………44

Figure 3.2. The data generating model………………………………………………………..45

Figure 3.3. The standard errors for the item parameters by the number of replications (LC1)…57

Figure 3.4. The standard errors for the item parameters by the number of replications (LC2)…58

Figure 3.5. The standard errors for the person parameters by the number of replications………58

Figure 3.6. The bias for the person parameters by the number of replications………………….59

Figure 3.7. The standard errors for the mixing proportion by the number of replications………59

Figure 4.1a. Statistically significant two-way interactions among between-replication variables on the correct classification rate…………………………………………………………………70

Figure 4.1b. Statistically significant two-way interactions between Model and other between-replication variables on the correct classification rate………………………………………...71

Figure 4.1c. Statistically significant three-way interactions among between-replication variables on the correct classification rate…………………………………………………………………72

Figure 4.2a. Statistically significant two-way interactions among between-replication variables on the SE of item parameters……………………………………………………………………….78

Figure 4.2b. Statistically significant two-way interactions among between-replication variables on the RMSE of item parameters…………………………………………………………………...79

Figure 4.2c. Statistically significant two-way interactions among between-replication variables on the 95% coverage of item parameters…………………………………………………………80

Figure 4.2d. Statistically significant three-way interactions among between-replication variables on the 95% coverage of item parameters…………………………………………………………81

Figure 4.3a. Statistically significant two-way interactions among between-replication variables on the bias of person parameters…………………………………………………………………86

Figure 4.3b. Statistically significant three-way interactions among between-replication variables on the bias of person parameters…………………………………………………………………87

Figure 4.4a. Statistically significant two-way interactions among between-replication variables on the bias of α_{0}. ……………………………………………………………………….......94

Figure 4.4b. Statistically significant three-way interactions among between-replication variables on the bias of $\alpha_{0.}$ ……………………………………………………………………….......94

Figure 4.4c. Statistically significant two-way interactions among between-replication variables on the RMSE of $\alpha_{0.}$ ………………………………………………………………………...95

Figure 4.4d. Statistically significant three-way interactions among between-replication variables on the RMSE of $\alpha_{0.}$ ………………………………………………………………………...96

Figure 4.5a. Statistically significant two-way interactions among between-replication variables on the bias of $\alpha_{1.}$ ……………………………………………………………………….......97

Figure 4.5b. Statistically significant two-way interactions among between-replication variables on the RMSE of $\alpha_{1.}$ ………………………………………………………………………...98

Figure 4.6a. Statistically significant two-way interactions among between-replication variables on the bias of β_{02} ……………………………………………………………………….....103

Figure 4.6b. Statistically significant three-way interactions among between-replication variables on the bias of β_{02} ……………………………………………………………………….....104

Figure 4.6c. Statistically significant two-way interactions among between-replication variables on the SE of β_{02} ……………………………………………………………………….......105

Figure 4.6d. Statistically significant two-way interactions among between-replication variables on the RMSE of β_{02} …………………………………………………………………………106

Figure 4.6e. Statistically significant three-way interactions among between-replication variables on the RMSE of β_{02} …………………………………………………………………………107

Figure 4.7a. Statistically significant two-way interactions among between-replication variables on the bias of β_{12} ……………………………………………………………………….....108

Figure 4.7b. Statistically significant two-way interactions between model and between-replication variables on the SE of β_{12} …………………………………………………………109

Figure 4.7c. Statistically significant two-way interactions between model and data completeness on the RMSE of β_{12} …………………………………………………………………………110

Figure 4.7d. Statistically significant two-way interactions between model and between-replication variables on the RMSE of β_{12} ………………………………………………….....111

Figure 4.8a. Overall model selection percentage across simulation conditions…………..……120

Figure 4.8b. Model selection percentage by data completeness across other manipulated factors…………..………………..………………..………………..………………..…………121

Figure 4.8c. Model selection percentage by the strength of relation between the dichotomous covariate and the latent class membership across other manipulated factors…………..………123

Figure 4.8d. Model selection percentage by the strength of relation between the continuous covariate and the person parameter across other manipulated factors…………..……………...124

Chapter 1 Introduction

In a wide variety of empirical studies in behavioral science and education, collateral information is collected about individuals in addition to the variables of primary interest to the researchers. This type of collateral information, also referred to as concomitant variables, auxiliary variables, or simply, covariates in the literature, usually contains individual background characteristics such as gender, ethnicity and years of education. Although these types of "outside variables" are sometimes ignored in statistical modeling, it is believed that they may potentially have important relations with the modeled variables of primary interest. The present study is an investigation of the potential benefits and limitations of including such covariate information in mixture item response theory (IRT) modeling, in complete and incomplete data scenarios.

1.1 Statement of the Problem

Mixture IRT models, which combine IRT and latent class analysis (LCA), have been increasingly used in psychometric research for analyzing item response data that may violate basic underlying assumptions of either modeling approach. Whereas IRT models assume the latent variable, a person's latent trait, to be continuous in nature, models in the LCA framework categorize respondents into qualitatively different latent groups based on their observed item responses. As a combination of the two modeling approaches, mixture IRT models estimate both the examinees' continuous latent trait and latent class membership of examinees simultaneously.

In recent years, the mixture IRT modeling approach has been applied to tackle a variety of important psychometric issues in test development, including the identification of latent differential item functioning (DIF) (Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002; Kelderman & Macready, 1990; Lu & Jiao, 2009; Samuelsen, 2005), the detection of testing speededness (Bolt, Cohen, & Wollack, 2002; Boughton & Yamamoto, 2007; Meyer, 2010; Yamamoto, 1989; Yamamoto & Everson, 1997) and the classification of test-takers with alternative cognitive attributes or strategies (e.g., Mislevy, & Verhelst 1990; Rijmen & De Boeck, 2003; Rost & von Davier, 1993). Further, this approach has also been applied to many practical situations. For instance, in psychopathological testing (e.g., Finch & Pierson, 2011; Maij-de Meij, Kelderman, & van der Flier, 2008), it is of great concern for researchers and clinicians to assign subjects to their most likely type of behavior disorders. In such scenarios, mixture IRT models can be used for diagnostic purposes from which an intervention program may be implemented.

Among all the applications of mixture IRT models, the identification of latent DIF is an important one related to test development. DIF refers to a phenomenon in which individuals with the same ability but from different subgroups do not have the same probability of a correct response to an item on a test. The presence of DIF does not indicate that the test is unfair, but it could be used as a warning flag, signifying potential threats which may jeopardize test fairness. This phenomenon could be attributed to the unintentional introduction of a nuisance dimension to the test items in addition to the dimension which is intended to be measured (Ackerman, 1992). Current DIF analysis is typically conducted based on manifest grouping variables, such as gender and ethnic groups. Potential DIF may be overlooked if it is caused by manifest grouping variables that are not included in the analysis or the interactions among multiple manifest grouping variables (Chen & Jiao, 2014; Jiao & Chen, 2014). Using mixture IRT models for latent DIF detection helps to deal with such issues (e.g., Cohen & Bolt, 2005; Jiao & Chen, 2014).

In practice, small sample sizes, short test lengths and small separations among latent classes often pose challenges for mixture IRT model estimation, especially in identifying latent group membership and obtaining accurate model parameter estimates (e.g., Smit, Kelderman, & Flier, 1999). The literature (e.g., Dai, 2009; Smit et al., 1999; Smit, Kelderman, & Flier, 2000) indicates that incorporating potentially important covariate information (e.g., demographic data) may yield desirable psychometric properties in the estimation of mixture IRT models, such as reducing standard errors in model estimation and improving the accuracy of latent class membership identification. Thus, an inclusion of effective covariates may be of theoretical and practical importance for the applications of mixture IRT models.

1.2 The Purpose of the Study

The purpose of the present study is to investigate different approaches to adding covariates into the mixture Rasch model (MRM), and the corresponding impact in model parameter estimation with both complete and incomplete response data. A Monte Carlo simulation is conducted in which data generated according to a two-class MRM with both dichotomous and continuous covariates are fitted to several misspecified MRMs with and without covariates.

As demonstrated by previous simulation studies and empirical research (Dai, 2009; Smit et al., 1999, 2000), incorporating potentially effective covariates in mixture IRT models may help relieve the rigid requirement of large sample size and large separations among latent classes in model parameter estimation, and obtain more accurate model parameter estimates and latent class assignment. However, certain important areas still remain unexplored in this line of research. First, previous studies in the mixture IRT modeling framework have exclusively focused on dichotomous covariates related to the latent class membership. No research includes continuous covariates. Second, the possibilities of relating dichotomous covariates with other model parameters have not been explored and no information is available about different potential approaches to including both dichotomous and continuous covariates in the model. Third, none of previous studies provide information about model fit and model selection with respect to covariate inclusion. Fourth, all previous studies are based on complete item response data sets. The manner in which dichotomous as well as continuous covariates enter a model and the impact of different approaches to covariate inclusion on model still have not been thoroughly explored in the mixture IRT modeling framework. Therefore, this study proposes to examine the impact of different approaches to incorporating dichotomous and continuous covariates in mixture IRT models on model performance, based on complete and incomplete item response data sets.

Both dichotomous and continuous covariates are included in the present study as predictors for the latent class membership and the person ability parameters. The impact of covariate specification is compared and analyzed in terms of model parameter recovery, latent class identification, and the relative overall model fit among competing models. Finally, an illustration of applying covariate inclusion approaches is demonstrated using the Programme for International Student Assessment (PISA) 2009 U.S. reading assessment data.

1.3 Significance of the Study

Three major advantages gained from the covariate inclusion approaches investigated in the present study can be summarized as follows. First, the present study uses one-step estimation of the conditional model, in which the model parameters and the relations between covariates (i.e., dichotomous and continuous) and model parameters are estimated simultaneously. Having covariates enter the model as predictors of the latent class membership could be used for simultaneous detection of latent DIF and explanation of latent DIF through the relations between manifest and latent groups. Previous mixture IRT modeling approaches for latent DIF detection usually involve two steps: identifying latent DIF and then probing the relations between latent classes and manifest groups (Chen & Jiao, 2014; Cohen & Bolt, 2005). However, as shown in a simulation study on a non-mixture Rasch model with covariates (Adams, Wilson, & Wu, 1997), two-step estimation tends to result in larger error variance for the item parameter estimation, larger mean squared error for the person parameter and underestimation of the regression coefficients for the covariates, especially when the test is short and the relation between the covariates and model parameters is strong. Thus, it is suggested that simultaneous estimation may potentially result in better model parameter estimates and more accurately capture the relation between manifest grouping variables and latent groups. As such, the cause of DIF may be more easily interpreted.

The second advantage is that the use of covariates, especially continuous covariates as predictors of the person parameter, may potentially relieve the rigid requirements of large sample size and latent class separation in mixture IRT model estimation, according to the literature in both non-mixture and mixture IRT framework (Adams et al., 1997; Mislevy, 1987; Mislevy & Sheehan, 1989a, 1989b). As such, more accurate parameter estimates and latent class assignment may be obtained.

Third, the impact of covariate inclusion on model performance may be more pronounced when missing data are present. In educational and psychological tests, covariates and demographic data have been found to account for as much as one third of the population variance and can increase the precision of model parameter estimation in the same amount as adding 2 to 6 items (Mislevy, 1987). This gain could be substantial in educational assessments/surveys or adaptive testing scenarios where only a small number of items (i.e., 5 to 15) are administered to each respondent (Mislevy & Sheehan, 1989a). The present study simulates different types and amounts of missing data that approximate those observed in large-scale assessment scenarios, so that the importance of covariate inclusion in mixture IRT models could be potentially revealed with regard to practical settings.

1.4 Overview of the Chapters

In the following chapters, the rationale of incorporating covariates in mixture IRT models and the different approaches to covariate inclusion are detailed after an introduction to the latent variable modeling framework.

In Chapter 2, IRT models, latent class models, mixture IRT models and how these modeling approaches are interrelated and integrated in the GDM framework are discussed in details. Further, the rationale of including covariates in latent class models, non-mixture IRT models and mixture IRT models is also elaborated. Covariate inclusion was first proposed in LCA as concomitant-variable latent-class model by Dayton and Macready (1988, 1989). This research area also emerged in non-mixture IRT framework as an explanatory modeling approach (e.g., Mislevy, 1987; Mislevy & Sheehan, 1989a, 1989b; Verhelst & Eggen, 1989). Later, this line of research has been extended to the mixture IRT modeling and other mixture modeling framework (e.g., growth mixture modeling). This chapter focuses on why covariate inclusion is promising for different types of latent variable models, how covariates enter a model and what desirable psychometric properties covariate inclusion may bring to mixture IRT models. Similar approaches from different perspectives (e.g., hierarchical GDM) are also briefly discussed. Since the present study intends to use one-step estimation of the conditional model, estimation will be carried out within a Bayesian framework vis-à-vis Markov chain Monte Carlo (MCMC) algorithms. An introduction to Bayesian inference, major sampling methods and relevant convergence diagnostic criteria are presented in the last part of this chapter.

Chapter 3 describes the technical issues regarding the model specification and model estimation implementation of the current study. The data generating model and five MRMs with misspecified covariates are illustrated. The second part of this chapter focuses on the implementation of model estimation in WinBUGS. The third part describes the simulation design with the purpose of investigating the impact of different approaches to covariate inclusion on fitting MRMs in complete and incomplete item response data scenarios. The evaluation criteria for model performance are also presented. Following the simulation study, an empirical example using the publicly available PISA 2009 reading assessment data is provided to demonstrate the impact of different covariate inclusion approaches in an MRM in real applications.

The results obtained from the simulation study are presented in Chapter 4. The influences of manipulated factors on latent class identification, model parameter recovery, and overall model fit are summarized. In addition, different approaches to covariate inclusion are compared based on the results from the 2009 PISA reading assessment data.

The last chapter discusses the findings, and points out potential limitation and future directions for this line of research. The implications of covariate inclusion in mixture IRT models to practical large-scale assessment settings are also provided in this chapter.

Chapter 2 Literature Review

In the 1960s, Rasch introduced a new statistical approach, item response theory, in his seminal paper (1960) and this model-based measurement theory has gained in popularity and prominence partially through its promotion in a textbook by Lord and Novick (1968). More recently, item response theory has gradually become the mainstay of modern measurement theories, and serves as the psychometric underpinning for much of large-scale testing today. With the advantage of population invariance and the ability of scaling persons and items on the same metric as compared with classical test theory, IRT models have increased attractiveness to many measurement practitioners. A substantial body of research has investigated the theoretical importance and practical applications of IRT, with the recent extensions of using more complicated IRT models to analyze item response data from more complex item and person populations.

The different approaches to covariate inclusion that are investigated in the present study are situated in the mixture IRT modeling framework. The mixture IRT modeling approach, as a combination of IRT and LCA, has been increasingly used in recent years for analyzing item response data that may violate underlying assumptions of either modeling approach. Mixture IRT models incorporate qualitative latent variables, which specify classes to which examinees belong, as well as quantitative variables which characterize latent trait within each class. Examples of mixture IRT models include the mixture Rasch model (Rost, 1990), the model presented by Mislevy and Verhelst (1990), which is a discrete mixture of a linear logistic test model (LLTM) identifying different latent trait variables across latent classes, and the loglinear modeling framework proposed by Kelderman and Macready (1990), which aims at detecting DIF through differences in item parameters or error rates across levels of manifest or latent grouping variables.

In this chapter, a review of IRT models and latent class models is presented first, followed by a summary based on the relation between these two modeling approaches and a unifying perspective of general diagnostic model (GDM). The mixture IRT modeling approach is also reviewed as a special case of the mixture distribution GDM. In the second section of Chapter 2, the rationale of including covariates in different lines of latent variable modeling research is discussed with a focus on the desirable psychometric properties resulted from covariate inclusion. The last section of this chapter focuses on the estimation method used in the present study. The Bayesian inference, major sampling methods and convergence diagnosis are described in this section.

2.1 Latent Variable Modeling Framework

In statistics, latent variables are defined as hypothetical constructs that are not directly observed but may be inferred from variables that are observed or directly measured. Latent variables could be categorical, ordinal or continuous in nature, and a latent structure model is a statistical model that relates a set of manifest variables to latent variables. When a latent variable is categorical, nominal latent classes are obtained; when a latent variable is continuous, a latent trait on a psychological continuum is assumed. Moreover, an example of ordinal latent variable is the use of ordered latent classes (Croon, 1990, 1991), such as classifying the scholastic aptitude of students into one of several levels of education.

The latent variable modeling framework embraces most commonly-used psychometric models with latent structure of different scale types, among which IRT models, latent class models and mixture IRT models are the focus of the present study. In the following sections, a brief review of the three types of models is provided with a discussion of the relations among these models. The rationale of including covariates in each modeling approach is detailed in later sections.

2.1.1 Item response theory (IRT) and latent class analysis (LCA).

Item response theory. Modern item response theory includes a variety of probability models (e.g., Birnbaum, 1968; Rasch, 1960), which characterize a nonlinear regression of item responses on a latent variable. Item responses are usually modeled in the form of the probability of a correct response from an examinee to a particular test item and the latent trait could be a single ability on a psychological continuum that an assessment instrument intends to measure. The individual performance differences on the instrument are attributed to different levels of underlying latent abilities of examinees.

There are three important assumptions underlying unidimensional IRT models: unidimensionality, monotonicity and local independence. The assumption of unidimensionality requires that the person parameter of an IRT model is restricted to only one latent dimension. Basically, a sufficient condition for this assumption is that only one common factor can be extracted from the matrix of tetrachoric item correlations (Lord & Novick, 1968). When this assumption is met, the conditional distributions for subpopulations are identical (Hambleton & Swaminathan, 1985).

Along with the assumption of unidimensionality, most IRT models also assume that the probability of correctly responding to a dichotomously scored item increases as individual latent ability increases (Reckase, 2009). This assumption is termed the monotonicity assumption; yet in real data, this assumption might hold with some small deviations (Sijtsma & Junker, 2006). As monotonicity is required for all items in an assessment instrument, this assumption is indeed a strong assumption, but sometimes it can be replaced by an alternative, weak monotonicity assumption (Stout, 1987; 1990), which assumes that the mean of item response functions is monotonically increasing as individual latent ability increases. The weak monotonicity assumption guarantees that there is enough information for the latent trait estimation (Sijtsma & Junker, 2006).

The last assumption, local independence, comprises two parts – item independence and person independence, which formulate the mathematical expression of the joint probability of examinees' responses to items (Mokken, 1996). Under this assumption, responses to all items across examines at a specified latent ability level are independent of each other. The first part, item independence or conditional independence, is the independence of responses within persons. This condition requires that for any single examinee at a given latent trait level, his or her response to any item on the test does not affect that person's responses to any other items on the test (Reckase, 2009). That is, all systematic variations in the responses are completely due to the variations of examinees over their latent trait levels (Mokken, 1996). On the other hand, the person independence, also called sampling independence, implies that the response of any examinee to a single item is not related to the responses to that item provided by any other examinee (Mokken, 1996). Taken together, the local independence assumption requires that the response of any examinee to any item on an exam only depends on the person's level on the latent trait and the item parameters which define the item response function (Reckase, 2009).

However, in practical testing situations, the unidimensionality and local independence assumptions are quite strong and may rarely be met. The major threat to the unidimensionality assumption is the possibility that the cognitive underpinning of an assessment instrument may include more than one dimension. In this situation, multidimensional IRT models can be used (Reckase, 1972, 1997, 2009).

On the other hand, the major challenge to the local person independence assumption is the nested data structure (e.g., Jiao, Kamata, Wang, & Jin, 2012). One example is the hierarchical data structure which is often observed in achievement test, with students nested in teachers and teachers nested in schools. In this situation, multilevel IRT models are used to deal with the person local dependence issue. The other special case of nested data structure pertains to heterogeneity of the population in which qualitative difference exists between latent subgroups of examinees. As such, examinees are conceived as nesting in latent classes. In this scenario, a standard IRT model cannot accurately predict the response patterns of all examinees (Kelderman & Macready, 1990; Mislevy & Verhelst, 1990; Rost, 1990), because there are qualitative inter-individual differences that are not captured by the model. This issue has given rise to the mixture IRT modeling approach which empirically identifies homogeneous latent clustering of examinees from a heterogeneous population based on response data. The following sections will discuss latent class analysis in relation to mixture IRT models in more details.

Latent class analysis. Before a detailed discussion of mixture IRT models, latent class analysis (LCA; Lazarsfeld & Henry, 1968), which provides the theoretical and empirical background for mixture distribution IRT models, is reviewed. All mixture IRT models described in later sections have their origins in latent class models. As a statistical method closely related to factor analysis and discrete mixture models, the simplest form of LCA can be considered either as a factor analytic model or a mixture of product-multinomial distributions (Dayton & Macready, 2007; Goodman, 1974a, 1974b; Haberman, 1979).

A basic assumption of latent class models is local independence, which requires that the probability of a response vector given a latent class is the product of class specific marginal response probabilities:

$$p(x_1,\ldots,x_I \mid g) = \prod_{i=1}^{I} p_{gi}(x_i)\,, \tag{2.1}$$

where x_i denotes the item response for item i as defined in the previous section, and g represents the unobserved categorical latent grouping variable. Thus, the joint probability of a response vector is given by:

$$P(x_1,\ldots,x_I) = \sum_{g=1}^{G} \pi_g \prod_{i=1}^{I} p_{gi}(x_i)\,, \tag{2.2}$$

where $\pi(g)$ indicates the mixing proportion with the restriction $\sum_{g=1}^{G} \pi_g = 1$.

Another common assumption of latent class models is that all the individuals within a latent class have the same item response probability. Under this assumption, the generalized form of the unrestricted latent class model can be formulated as follows (Dayton & Macready, 2007):

$$P(\mathbf{X}) = \sum_{g=1}^{G} \pi_g \prod_{i=1}^{I} \prod_{j=1}^{J} \lambda_{giy}^{\delta_{ijy}}\,, \tag{2.3}$$

$$\delta_{ijy} = \begin{cases} 1 & \text{iff } x_{ij} = y \\ 0 & \text{otherwise} \end{cases}$$

where λ_{giy} indicates the probability of a response y to item i conditional on the gth class membership with the restriction

$$\sum_{y=1}^{m_i} \lambda_{giy} = 1, \tag{2.4}$$

in which y takes mutually exclusive and exhaustive values from 1 to m_i for the ith item. m_i may take any value, depending on the number of response categories for item i. For example, for a dichotomous items, m_i =2. As such, Equation 2.3 and 2.4 could be generalized to mixed-format test (i.e., with dichotomous and polytomous items with different numbers of response categories). Additionally, x_{ij} denotes the score examinee j attains on item i. When x_{ij} equals y, δ_{ijy} takes the value of 1; otherwise, δ_{ijy} takes the value of 0.

In order to determine the classification of examinees, each individual is assigned to the class which yields the highest conditional probability given the response vector (Rost & Langeheine, 1997). For example, the posterior probability of a class membership by applying Bayes' rule is

$$p(g' \mid x_1,...,x_I) = \frac{\pi_{g'} p(x_1,...,x_I \mid g')}{\sum_{g=1}^{G} \pi_g p(x_1,...,x_I \mid g)}. \tag{2.5}$$

Compared with IRT models, LCA is less restrictive in terms of person-homogeneity because different classes of individuals can be described by totally different set of item parameters; however, LCA is more restrictive than IRT models in that all individual differences are explained by a limited number of classes (Rost & Langeheine, 1997). In IRT models, each person is assumed to have a distinct position on the latent psychological continuum; however, in LCA all persons within a latent class are treated as identical in terms of the probability of a specific response to an item (i.e., λ_{giy}). Statistically, the conditional probability of a response to an item given a latent class is restricted to be the same for all members of that specified latent class, but the item response probability is allowed to vary across different latent classes (Rost & Langeheine, 1997).

2.1.2 General diagnostic model (GDM).

In spite of the differences between IRT and LCA, both these two types of models fall under a flexible framework of general diagnostic models (GDM; von Davier & Yamamoto, 2004; von Davier, 2005a, 2008a) which incorporate a variety of latent structure models that describe the probability of observed responses in terms of conditional probabilities given one or more latent variables (von Davier, 2009).

In the GDM, the conditional probability of a response x_i from a person with attribute pattern a to item i is defined as

$$P(X_i = x_i \mid a) = \frac{\exp\left(\eta_{ix} + \sum_k \gamma_{ikx} h(q_{ik}, a_k)\right)}{1 + \sum_{y=1}^{m_i} \exp\left(\eta_{iy} + \sum_k \gamma_{iky} h(q_{ik}, a_k)\right)}. \tag{2.6}$$

In this definition, a is a K-dimensional latent variable with $a = (a_1,\ldots,a_K)$; X is the observed response variable with realizations $x_i \in \{0,\ldots,m_i\}$ and $i \in \{1,\ldots,I\}$; and q_{ik} denotes the vector of Q-matrix entries for item i in dimension $k \in \{1,\ldots,K\}$, which indicates the set of required person attributes. In addition, η_{ix} is a real valued difficulty parameter and γ_{ikx} is a K-dimensional slope parameter for each non-zero response category (von Davier, 2005a, 2008a, 2010). Thus, when there are m_i +1 categories in the response data, $m_i \times k$ slope parameters are specified for item i. In this generalized model, a could be a vector of latent abilities on the psychological continuum or a vector of dichotomous person attributes indicating whether a person has mastered a specific set of skills, and the response variable could be either dichotomous or polytomous (von Davier, 2005b).

In this sense, the GDM framework covers a variety of latent structure models for dichotomous and polytomous data, depending on the latent structure assumed (von Davier, 2005b). If *a* is a one-dimensional continuous variable, a dichotomous or polytomous IRT model is obtained; if *a* is multidimensional continuous, a multidimensional IRT model is obtained. Further, if *a* is binary with one single dimension, the generalized model conceptually reduces to a latent class model. In addition, if *a* is multidimensional with multiple binary components, a diagnostic classification model (DCM) is obtained. Thus, both IRT models and latent class models reviewed in this section are special cases of the GDM, depending on the dimension and the scale type of the latent variable.

As an extension of the GDM, the mixture distribution GDM (MGDM; von Davier, 2008b) was introduced to accommodate potential qualitative differences between subpopulations or clusters of observations when the data structure is complex. In the MGDM, the conditional probability of a response x_i given attribute pattern *a* and class *g* is specified as

$$P(X_i = x_i \mid a, g) = \frac{\exp\left(\eta_{ixg} + \sum_k \gamma_{ikxg} h(q_{ik}, a_k)\right)}{1 + \sum_{y=1}^{m_i} \exp\left(\eta_{iyg} + \sum_k \gamma_{ikyg} h(q_{ik}, a_k)\right)}. \qquad (2.7)$$

The η_{ixg}, and γ_{ikxg} are class-specific item difficulty and slope parameter respectively (von Davier, 2010). $g \in \{1,\ldots,G\}$ is the group index which could be manifest, latent or partially observed. For the mixture distribution model that no *g* is observed, a mixture GDM such as a mixture IRT model is derived; whereas for models with all *g* known a priori, a multiple-group GDM is obtained (e.g., multiple-group IRT). There is also possibility to build up a mixture GDM with group membership partially observed (von Davier & Yamamoto, 2004). The concept of partially observed mixtures is an attempt to combine multi-group and mixture IRT models (von Davier & Yamamoto, 2004). It is assumed that there is missing grouping information for some, but not all, observations in the population (von Davier & Yamamoto, 2004). This type of model is proposed to help recover such missing grouping information for those examinees (von Davier & Yamamoto, 2004). In the next section, the mixture IRT modeling approach, which is a special case of the mixture distribution GDM and the focus of the present study, is discussed in details.

2.1.3 Mixture IRT models.

One major weakness of LCA is the restrictive assumption that all examinees in a given latent class have the same response probability. This property of latent class models gives rise to several lines of research, one of which is the development of latent trait models within the latent class modeling framework (Rost & Langeheine, 1997). In the early 1990s, there were three cutting-edge articles which contributed to the early development of mixture IRT models. These articles were from Kelderman and Macready (1990), Mislevy and Verhelst (1990) and Rost (1990), and they proposed the idea of combining latent trait models with latent class models from different perspectives. Additionally, the HYBRID model proposed by Yamamoto (1987, 1989) is also considered an early development of mixture IRT modeling approach which is a mixture of latent groups of respondents who can be characterized by either an IRT model or a deterministic model with fixed response probability.

Mixture Rasch models (MRM). The basic idea of the MRM is to incorporate the Rasch model in a discrete mixture of latent subgroups (i.e., latent classes), with the Rasch model applied to each class but with different item parameters across latent classes (Rost, 1990). The MRM enables a partition of the examinees, which maximizes the qualitative difference between subpopulations (Rost, 1990). By using the MRM, the qualitative difference between examinees is taken into consideration, and simultaneously the individual abilities are quantified on a continuum.

Mathematically, the probability of a correct response for person *j* on item *i* conditional upon the class-specific item difficulty, person's latent ability and person's latent class membership in the MRM is specified as

$$P(X_{ij}=1\,|\,\theta_{jg},b_{ig},g)=\frac{1}{1+\exp[-(\theta_{jg}-b_{ig})]}, \qquad (2.8)$$

where b_{ig} is the difficulty parameter of item *i* conditional on latent class g ∈ {1,…,*G*}, and θ_{jg} is the person ability parameter which is estimated conditional on the categorical latent class membership. Correspondingly, the unconditional probability of a correct response for person *j* on item *i* is specified as

$$P(X_{ij}=1\,|\,\theta_{jg},b_{ig})=\sum_{g}\pi_g\frac{1}{1+\exp[-(\theta_{jg}-b_{ig})]}, \qquad (2.9)$$

where π_g denotes the mixing proportion and sums to 1 (i.e., $\sum_{g}\pi_g=1$).

For each person, the latent group membership is assigned by comparing the posterior probability of that person belonging to each latent class if maximum likelihood estimation is used. That person will be assigned to the latent class with the largest posterior probability, or the posterior probability weighted by utility or loss values. On the other hand, in Bayesian estimation, the latent class membership is considered as a parameter, the estimated value of which is drawn from a posterior multinomial distribution. As such, the latent class membership estimate is the mode of the posterior distribution.

Additionally, given that the Rasch model is assumed within each latent class in the MRM, the sum of the item difficulty parameters within each class could be constrained to be zero (i.e., $\sum_{i}^{I} b_{ig} = 0$) for model identification purpose.

A mixture extension of LLTM. Another important perspective of mixture distribution IRT models was introduced by Mislevy and Verhelst (1990) with respect to students' strategy-use in problem-solving. This is also an extension of LLTM in which individuals are assigned to different strategy classes with the item difficulty defined as a linear function of item features or cognitive operations involved in the problem-solving process. This approach concerns situations in which different groups of examinees tend to choose different strategies. The distinct nature of this approach is that different latent traits are assumed under different strategies (Mislevy & Verhelst, 1990). This is in contrast with the assumption of Rost's (1990) model where the same latent trait is assumed for all latent groups. Thus, according to Mislevy and Verhelst (1990), a comparison between persons within the same strategy class is meaningful, while a cross-class comparison of proficiency levels of all examinees is not valid because essentially two different latent traits are measured across latent classes. The strategy class under this approach is also not observable but can be inferred from response patterns and prior beliefs.

A loglinear modeling framework for DIF. Another perspective of mixture IRT models was proposed by Kelderman and Macready (1990) with respect to detecting DIF across manifest and latent examinee groups. As discussed in the previous chapter, DIF pertains to the phenomenon that items do not function in the same way for examinees in different subpopulations. The major difference between manifest DIF detection analysis and latent DIF based on mixture IRT models lies in the fact that the invariance of model parameters for a specified IRT model is usually compared across manifest subpopulations in traditional DIF analysis, whereas unobservable groupings of examinees are detected via qualitative differences implied by response patterns in mixture IRT models (Cohen & Bolt, 2005) and subsequently item parameter invariance is compared among the latent groups. This is a major advantage of mixture IRT models as compared with manifest DIF analysis as a detection method, as pointed out by Kelderman and Macready (1990). The general loglinear modeling framework they proposed enables the detection of differences in item parameters or error rates across levels of grouping variables, which may be manifest, latent or both, for models with either categorical or continuous attribute of interest (Kelderman & Macready, 1990). They further showed that this framework is flexible in that it is possible to incorporate additional interaction effects between item parameters and observed or unobserved grouping variables (Kelderman & Macready, 1990).

The HYBRID model. Different from the mixture distribution IRT models presented above, Yamamoto (1987, 1989) proposed the HYBRID model which allows different models in different components of the mixture distribution. For example, this model may incorporate two latent groups: one consists of random guessers for whom the independence of responses holds (i.e., this is simply a latent class grouping with a single set of conditional probabilities that define member likelihoods of item responses), while the other group's probabilities for item responses are characterized by an IRT model (Yamamoto, 1989). The implication of the HYBRID model is that it is more appropriate to model the guessing behavior on the person level rather than the item level (von Davier & Rost, 2006). The HYBRID model can also be extended to strategy switching, which means that an IRT model is appropriate for a subset of responses of an examinee, and a latent class model is suitable for the rest of the responses (e.g., guessing) of that person (Yamamoto & Everson, 1997).

In summary, the above mentioned research work has greatly contributed to the development of mixture IRT models and they serve as the theoretical building blocks for the present study. First, the MRM proposed by Rost (1990) lays the ground for the current study which investigates a two-class MRM with correctly-specified and misspecified covariates. Second, the present study employs a one-step estimation of the conditional model which allows the identification of latent DIF and the explanation of latent DIF using manifest grouping variables simultaneously. This perspective was built upon the work by Kelderman and Macready (1990) which promotes the use of mixture IRT models as a DIF modeling approach. Third, the present study incorporates both dichotomous and continuous covariates as predictors of model parameters (i.e., the latent class membership and the person parameter) and this approach is related to the perspective of Mislevy and Verhelst (1990). Although not directly discussed, the mixture extension of LLTM proposed by Mislevy and Verhelst (1990) is statistically an approach to include covariates as predictors of the item parameter, with the empirical meaning that the item parameter may be defined as a linear combination of prespecified cognitive operations. The current study, on the other hand, targets at covariate inclusion with respect to the person parameter and the latent class membership of the MRM, without considering item features.

2.2 The Rationale of Including Covariates

As a member of the mixture model family, the MRM shares many similarities with other types of mixture models (e.g., growth/factor mixture models; Muthén, 2001; Lubke & Muthén, 2005). The accurate identification of latent group membership and estimation of model parameters related to each latent class are critically important to all types of mixture models. However, small sample sizes, separations between latent classes and the interaction between the two may pose challenges for model parameter estimation and latent class identification in both growth mixture and mixture IRT models (e.g., Kohli, Harring, & Hancock, 2013; Li, Harring, & Macready, 2014; Lubke & Muthén, 2007; Smit et al., 1999). Thus, the inclusion of covariates has been proposed in both the mixture IRT and other mixture modeling framework in order to help relieve the rigid requirement of sample size and latent class structure in model estimation and to achieve different purposes, such as obtaining more accurate model parameter estimates (e.g., Dai, 2009, 2013; Smit et al., 1999, 2000), latent class assignment (e.g., Li & Hser, 2011), and enumeration of latent classes (e.g., Lubke & Muthén, 2007; Muthén, 2004).

The general approaches to covariate inclusion have been proposed in different lines of mixture modeling research. In the factor/growth mixture modeling framework, covariates may be included as predictors of the latent factor, the latent group membership, some distal outcomes or even observed variables (Li & Hser, 2011; Lubke & Muthén, 2005; Petras & Masyn, 2010). Similarly, in the IRT literature, both continuous and categorical covariates have been proposed to be incorporated in the models as predictors of the person and/or the item parameter, as well as predictors of the latent class membership in the case of mixture IRT models (Smit et al., 1999, 2000; Wilson & De Boeck, 2004). It has been suggested by previous research that the model estimation (e.g., enumeration of latent classes) may potentially benefit from the inclusion of observable covariates as predictors of the latent class membership and latent factors (e.g., Muthén, 2004). However, based on limited research in this area, it is still unclear whether and how different approaches to covariate inclusion would have beneficial or detrimental effects on model estimation.

Since the present study is on the MRM, which is a combination of IRT models and LCA, the following sections will provide a detailed discussion about the rationale of including covariates and the development of covariate inclusion approaches in latent class models, non-mixture IRT models as well as mixture IRT models.

2.2.1 Covariate inclusion in LCA.

The possibility of covariate inclusion was first introduced in the latent class modeling framework as concomitant-variable latent-class models (Dayton & Macready, 1988, 1989). As an extension of the simultaneous LCA (i.e., the application of LCA to multiway contingency tables simultaneously), concomitant-variable latent-class models allow the probability of latent class membership to be functionally related to covariates with known distribution (Dayton & Macready, 1988, 1989). The explanatory variables may be categorical, indicating manifest group membership (e.g., gender), or continuous such that the values may be different for each observation (Dayton & Macready, 1988, 1989). Further, the function linking covariates and the probability of latent class membership could be in any appropriate form such as logistic or exponential (Dayton & Macready, 1988, 1989).

Let *w* denote a vector of covariates. Then, a *G*-latent class concomitant-variable model can be specified as

$$P(x_1,...,x_I \mid w) = \sum_{g=1}^{G} \pi_{g|w} P(x_1,...,x_I \mid g), \tag{2.10}$$

with the functional relation between the mixing proportion and covariates defined as

$$\pi_{g|w} = f(w;\upsilon_p), \tag{2.11}$$

in which $\upsilon_p = \{\upsilon_0,\ldots,\upsilon_P\}$ indicates coefficients for *P* concomitant variables (Dayton & Macready, 1988, 1989). In Equation 2.10, the conditional probability $\pi_{g|w}$ is subject to the restriction that $\sum_{g=1}^{G} \pi_{g|w} = 1$ for all unique *w*. As shown in Equation 2.11, there is a conditional relation between the covariates *w* and the mixing proportion π_g based on parameters υ_p within the latent class model. The relation as expressed in Equation 2.11 can take a variety of mathematical forms (Dayton & Macready, 1988, 1989).

Later, this approach was extended by Hagenaars (1990) by relating more than one categorical covariate to the latent class parameter and restricting the probabilities by loglinear models. Van der Heijden, Mooijaart and de Leeuw (1992) also took up this line of research by including categorical and grouped continuous covariates using multinomial logit models. Formann (1992) further proposed a generalization of the above mentioned approaches by restricting both latent class probabilities and conditional probabilities by multinomial logit models. In terms of estimation, van der Heijden, Dressens, and Bockenholt (1996) proposed a simultaneous estimation of the model parameters and the relation between covariates and the latent class parameters using the EM algorithm. Further, Bolck, Croon, and Hagenaars (2004) demonstrated that separate estimation is also plausible, and the underestimation of the relation between covariates and class membership resulted from separate estimation can be adjusted by a specific correction method. More recently, Vermunt (2010) proposed a new maximum likelihood based correction method and it makes the separate estimation as efficient as the simultaneous estimation.

The covariate inclusion approach in LCA has been proved to be promising in that it potentially reduces the classification error for latent class models (Hagenaars, 1993). It also has the added advantage that it is more parsimonious than the multiple group latent class models which sometimes have zero degrees of freedom (Dayton & Macready, 1988, 1989). This covariate inclusion approach may result in a positive number of degrees of freedom, depending on the number of concomitant variables.

More importantly, the idea of relating covariates to the latent class parameter has also been brought to mixture IRT and other mixture modeling framework. Thus, it has provided theoretical foundations and empirical implications for the covariate inclusion approaches in the MRM under investigation in the present study.

2.2.2 Covariate inclusion in non-mixture IRT models.

Reasons for covariate inclusion. The incorporation of covariates in non-mixture IRT models was first introduced by Mislevy (1987) as an approach to increase the precision of item parameter estimates. In educational assessment and adaptive testing scenarios, each examinee usually responds to only a few items either because of the assessment design or the nature of adaptive testing. Thus, the sparse information provided by the observed item responses may pose challenge to an accurate recovery of item parameters. This research has found that covariates may increase the precision of model parameter estimation in the same amount as adding 2 to 6 items and covariates, such as demographic data, could account for as much as one third of the population variance in educational and psychological tests (Mislevy, 1987). This proportional gain that results from the use of covariate information is substantial for short tests where only a small number of items (i.e., 5 to 15) are administered to each respondent or missing data are present and the effects tend to become less impressive when test length increases, such as in individual achievement test scenarios (Mislevy, 1987).

As an extension of this seminal research work, Mislevy and Sheehan (1989a, 1989b) furthered this line of research by providing an important argument and mathematical proof about when such inclusion of covariate information is appropriate for measurement models. In the case that covariate information is used in neither the examinee sampling nor the test administration (i.e., no stratified sampling/targeted test administration), or the case that covariate information is only used for examinee sampling (e.g., stratified sampling of examinees), it is not necessary to include covariates in the estimation of the measurement model (Mislevy & Sheehan, 1989b). The parameter estimates of the measurement model remain consistent in these two scenarios if covariate information is not used. However, the inclusion of covariate information would improve model parameter estimation, and the benefit tends to be greater when the covariate information is more strongly related to the latent variable and less information is obtained from observed item responses (Mislevy & Sheehan, 1989b). On the other hand, if covariate information is used in both the examinee sampling and the test administration (e.g., stratified sampling and targeted test administration), the corresponding covariate information must be taken into account in the measurement model; otherwise, the item parameter estimates would be inconsistent (Mislevy & Sheehan, 1989b).

Moreover, whether to include covariates in IRT models is a theoretical debate involving the validity of the inference drawn about the population. Considering the test fairness issue, the estimation of individual ability should be independent of any variables beyond the response data per se. It may be desirable to use covariates to improve the precision of model parameter estimation, but it is less desirable to draw inference based on the conditional model, especially when high-stake decisions are made for individuals in competitive tests (Mislevy & Sheehan, 1989a). This is a separate issue. However, purely in the perspective of model parameter estimation, covariate inclusion is promising for IRT models with the potential benefit of improving the estimation of both the item and the person parameters by reducing standard errors.

Thus, based on this perspective, the present study aims at investigating different approaches to covariate inclusion and the corresponding impacts on model parameter estimation in the MRM. In the following section, the approaches to including covariates in non-mixture IRT models are detailed.

Approaches to covariate inclusion. Based upon the initial proposal of covariate inclusion in IRT models (Mislevy, 1987), an explanatory modeling approach in item response theory has been gradually developed (Adams et al., 1997; Verhelst & Eggen, 1989; Wilson & De Boeck, 2004; Zwinderman, 1997). Covariates are included in a variety of standard IRT models for different purposes such as explaining estimated effects or improving model parameter estimation (e.g., Adams et al., 1997; Wilson & De Boeck, 2004).

Within this explanatory perspective, three types of explanatory IRT models, which are the person explanatory, the item explanatory and the doubly explanatory Rasch models, have been proposed as extensions of the Rasch model as summarized in Table 2.1 (Wilson & De Boeck, 2004). The graphical representations of person explanatory, item explanatory and doubly explanatory, adapted from Wilson and De Boeck (2004) and the M*plus* manual (Muthén & Muthén, 1998-2012), are also presented.

Table 2.1. *Explanatory Rasch models*.

Item covariates	Person covariates	
	Absence of covariates	Inclusion of covariates
Absence of covariates	Doubly descriptive	Person explanatory
Inclusion of covariates	Item explanatory	Doubly explanatory

In the person explanatory model (i.e., the latent regression Rasch model), the person parameter θ_j is regressed on external person properties as predictors (Adams et al., 1997)

$$P(X_{ij}=1\,|\,\theta_j,b_i)=\frac{1}{1+\exp[-(\theta_j-b_i)]}\text{ with }\theta_j=\upsilon_0+\sum_{p=1}^{P}\upsilon_p w_{jp}+\varepsilon_j\ , \tag{2.12}$$

where w_{jp} is the value of person j on the pth person property ($p \in \{1,\ldots,P\}$), υ_p is the regression coefficient related to person property p and υ_0 is the intercept. ε_j is the random person effect after the fixed effects of person properties are accounted for with the distributional assumption that $\varepsilon_j \sim N(0, \sigma_\varepsilon^2)$ (Wilson & De Boeck, 2004). There could also be interaction terms among person properties in the model. When person properties are included as covariates, the error term is usually specified in the model in consideration of the individual variation. In this model, w_{jp} is the observed person attribute. The regression function may also be built upon latent variables underlying the observed person-level properties (Fox & Glas, 2003). A graphical representation of the latent regression Rasch model is shown in Figure 2.1.

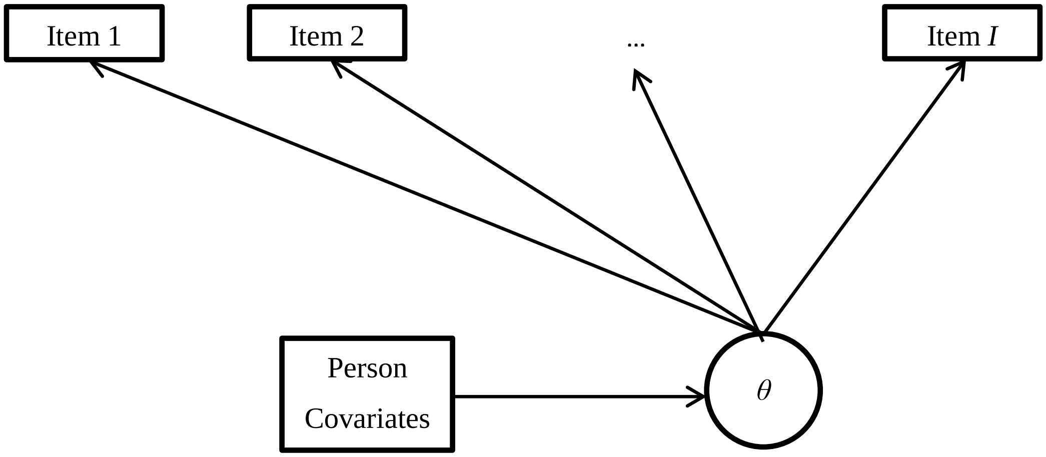

Figure 2.1. The person explanatory model.

In the item explanatory model (i.e., the LLTM), the item parameter (e.g., item difficulty) is specified as a linear combination of item properties (Fischer, 1973):

$$P(X_{ij} = 1 \mid \theta_j, b_i) = \frac{1}{1 + \exp[-(\theta_j - b_i)]} \text{ with } b_i = \upsilon_0 + \sum_{p=1}^{P} \upsilon_p w_{ip} + \varepsilon_i, \tag{2.13}$$

where each of the P item properties has a weight of υ_p. It is also possible to include interactions among item properties in the model. The early development of the LLTM conceives item difficulty as a linear combination of cognitive operations involved in the problem solving process. The cognitive features of each item and individuals' probability of response are assumed to be connected: each of the P features has a weight of υ_p, and it denotes the basic parameter of the model-the contribution of each operation to the item difficulty. As shown in Equation 2.11, an error term is specified in the regression function with $\varepsilon_i \sim N(0, {\sigma_\varepsilon}^2)$. This is a random-effect version of the LLTM (Janssen, Schepers, & Peres, 2004). Moreover, in certain situations, the error term is not specified in the model, indicating that the item difficulty is perfectly predicted by the item properties (Wilson & De Boeck, 2004). This assumption is so strong that it makes the applications of this type of LLTM quite restrictive. A graphical representation of the LLTM is given in Figure 2.2. The dotted arrows indicate the possibility of modeling the item discrimination parameter as a function of covariates in the case of a two parameter logistic (2-PL) IRT model. In the Rasch-based LLTM, the dotted arrows do not exist.

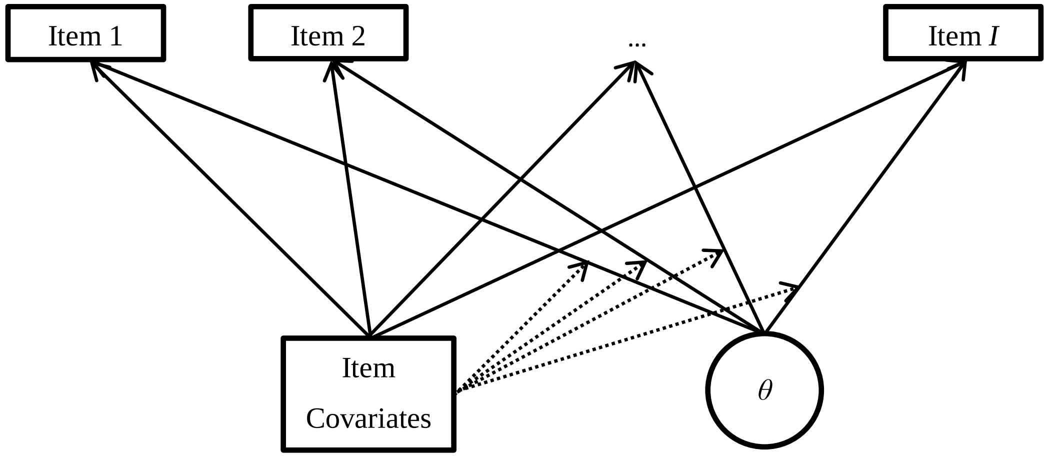

Figure 2.2. The item explanatory model.

Further, a doubly explanatory model is proposed as a combination of the two approaches presented above (i.e., the latent regression LLTM; Zwinderman, 1997). Figure 2.3 gives a graphical representation of this model.

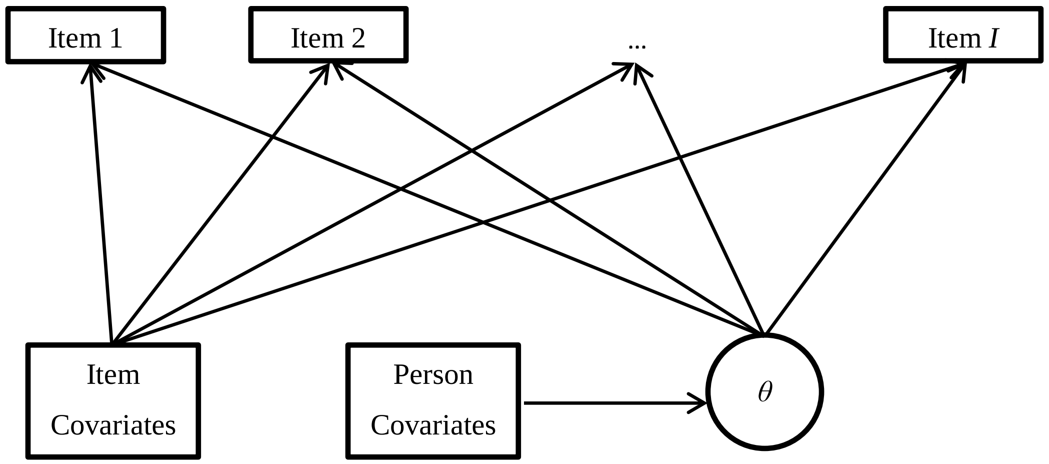

Figure 2.3. The doubly explanatory model.

The models presented above are the commonly used covariate inclusion approaches for non-mixture IRT models. As demonstrated by Adams et al. (1997), the use of covariate information as predictors of the person parameter substantially decreases the mean squared error of the person parameter estimates but has negligible effects on the accuracy of item parameter estimates in the Rasch model. The effects are more pronounced when the covariates are more strongly related to the person ability and the test length is short. Based on the findings, the present study incorporates only the person explanatory approach in the MRM, in addition to the approach to model latent class membership as illustrated in the concomitant-variable latent class models discussed above.

2.2.3 Covariate inclusion in mixture IRT models.

In mixture IRT models, covariates have been added to achieve similar goals as those in non-mixture IRT models. For example, Smit et al. (1999, 2000) explored the use of covariates in the mixture Rasch and 2-PL models by manipulating the association between latent classes and a dichotomous covariate in terms of bivariate probabilities. Their findings showed that the standard errors and the accuracy of latent class membership assignment could benefit from incorporating a covariate with a moderate to strong relation with the latent class variable. Samuelsen (2005) further explored the impact of covariates in the context of DIF. In her study, a dichotomous covariate was included in the MRM as an indicator of a manifest group membership and the level of overlapping between the manifest group and the latent group was under manipulation. Results showed that the power of detecting DIF decreases as the level of overlapping between the manifest and the latent group decreases, thus indicating that using the MRM to identify latent DIF may be a better approach than the current manifest DIF analysis.

More recently, latent class membership in the MRM was modeled using logistic regression with a dichotomous covariate as the sole predictor of latent class membership (Dai, 2009, 2013). Results indicated that the inclusion of dichotomous covariate in the MRM has positive effects on the correct recovery of latent structure (Dai, 2009, 2013). The strength of the relation between the covariate and the latent class membership also tends to impact the root mean squared error (RMSE) of the regression coefficients (Dai, 2009, 2013). Further, it was found that the DIF effect size has significant effects on both the identification of the latent structure and the accuracy of model parameter estimates (Dai, 2009, 2013). The most recent research work in this line was done by Tay and his colleagues (2011) in which they conducted a real data analysis exploring the mixture 2-PL IRT model in the context of DIF with both continuous and dichotomous covariates as predictors of the latent class membership.

2.2.4 Other relevant modeling approaches

Besides the covariate inclusion approaches, there are some other latent structure modeling approaches which can accommodate the complex data structure and heterogeneity of the population, and allow the inclusion of covariates at different levels. Such modeling approaches include multilevel LCA (e.g., Vermunt, 2003), multilevel IRT models (e.g., Kamata, 2001), all of which are special cases of the hierarchical GDM (HGDM; von Davier, 2007, 2010).

In the next section, the possible parameter estimation methods for mixture IRT models and the technical details of the Bayesian estimation used in the present study are described in details.

2.3 Model Estimation

There are two major model parameter estimation frameworks in statistics, the frequentist inference and the Bayesian inference. To a frequentist, model parameters are considered as fixed and unknown truth, and they are estimable by replications of data from experiments. In the frequentist point of view, data are repeatable random samples and underlying model parameters remain fixed in the repeatable process. On the other hand, the Bayesian perspective assumes that data are fixed from the realized sample, but model parameters are random and thus have distributions.

Maximum likelihood (ML), as a frequentist approach, has been used for a long time for the estimation of the models described above, including IRT models, latent class models as well as mixture IRT models. In ML estimation, parameter estimates are obtained by searching for values that maximize the likelihood function. ML is a consistent approach and it has been widely used for a variety of statistical models given the desirable mathematical properties offered by ML methods. Specifically, the ML estimators are asymptotically unbiased with minimum variance, and they approximate normal distributions and have sample variances which produce confidence intervals and enable hypothesis testing. Further, maximum likelihood estimation methods are computationally efficient. However, there are certain drawbacks for the ML estimation methods, such as the possibility of multiple local maxima, unbounded likelihood function and the difficulty to choose staring values; and these downsides are more pronounced for the estimation of complex models, such as the mixture IRT models.

On the other hand, there has been a surge in recent year to estimate complex psychometric model using Bayesian estimation methods. Compared with the ML estimation, the Bayesian inference has the following advantages. First, it allows the inclusion of prior information (i.e., prior knowledge or beliefs) into the current model parameter estimation. As more information is used, the Bayesian confidence intervals are supposed to be narrower than those produced by ML estimation. Second, the Bayesian inference via MCMC is unbiased with respect to sample size. Different from the ML estimation for which the desirable mathematical properties are asymptotic in nature, the Bayesian estimation can accommodate any sample sizes. Additionally, the Bayesian inference enables a better estimation of complex statistical model for which the ML estimation is sometimes problematic. For example, when the likelihood function is complicated across the parameter space, it may have several local maxima. It is very likely for the ML algorithms to arrive at the local maxima instead of the global maxima, resulting in the convergence issue. On the other hand, rather than searching for maxima, the Bayesian inference aims at exploring target distributions with the information from data and prior distributions. As such, the local maxima issue does not occur in Bayesian estimation of complex models. However, Bayesian estimation also has some disadvantages. For example, the selection of the prior distribution is sometimes arbitrary. If a highly informative prior is used, the posterior distribution may be heavily influenced by the prior if the sample size is small; on the other hand, if a flat prior is used, unreasonable values may be obtained for the parameter estimates. Further, the most serious drawback of Bayesian estimation is the large amount of time required. Therefore, it is often used for small data sets or in situations that ML estimation is less desirable.

Regarding the estimation of the MRM, several popular software packages based on ML algorithms may be used, such as M*plus* (Muthén & Muthén, 1998-2012), *mdltm* (von Davier, 2005b) and Latent GOLD (Vermunt & Magidson, 2000-2013), which are based on marginal maximum likelihood, and *WINMIRA* (von Davier, 2001), which is based on conditional maximum likelihood.

In the present study, there are certain trade-offs between ML and Bayesian estimation. On one hand, ML estimation is much more computationally efficient and less time-consuming than the Bayesian estimation, and this property makes ML estimation a desirable option. On the other hand, the present study aims at a one-step estimation of the MRMs with dichotomous and continuous covariates. Although, LCA literature has suggested that results from separate estimation may be as accurate as those from simultaneous estimation with the use of certain correction methods, IRT literature has found that two-step estimation tends to result in larger error variance for the item parameters, larger mean squared error for the person parameters, and underestimation of the regression coefficients for the covariates, especially when the test is short and the relation between the covariates and model parameters is strong (Adams et al., 1990). Given these drawbacks of two-step estimation, simultaneous estimation of the conditional model is implemented in the present study. As the model with covariates is even more complicated than the MRM, different types of convergence issues that are specific to the ML estimation (i.e., singularity of the information matrix, local maxima) are likely to occur. Thus, the use of Bayesian estimation may avoid these problems. Additionally, although the estimation of the MRM may be easily implemented in the above mentioned software packages, they have limited capacity of carrying out one-step estimation of the conditional model investigated in the present study. Given these issues and the trade-offs, the Bayesian approach is finally selected. The estimation of the current study is implemented in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003), which is elaborated in Chapter 3. In the following sections, the major sampling methods in Bayesian estimation and the convergence diagnosis are discussed in details.

2.3.1 Introduction to Bayesian inference.

Recall that the Bayes rule is expressed as

$$\underbrace{P(\mathbf{\Omega} \mid \mathbf{X})}_{\text{posterior}} \propto \underbrace{P(\mathbf{X} \mid \mathbf{\Omega})}_{likelihood} \times \underbrace{P(\mathbf{\Omega})}_{\text{prior}}, \tag{2.14}$$

where **Ω** denotes the parameters to be estimated for a model. *P*(*data*| **Ω**), the sampling density for the data, defines the likelihood function of a statistical model. *P*(**Ω**) denotes the prior distribution regarding the model parameters. Thus, the posterior distribution of **Ω**, given the observed data, is proportional to the likelihood function multiplied by the prior distribution of the model parameters.

Markov chain Monte Carlo (MCMC). In simple statistical models, it is easy to draw values from the posterior distribution, especially when conjugate priors are assumed. When multiple parameters are present, it is also possible to integrate out the parameter of interest from the joint posterior distribution. For posterior distributions in non-closed form with low-dimensional parameter space, either non-sampling methods or direct sampling methods (e.g., rejection sampling, importance sampling) can be used. However, for complex models with high-dimensional parameter space, in non-closed or closed form, such as the model in the present study, MCMC is usually used given that it provides a flexible way to draw values from the posterior distribution. It allows draws from low-dimensional distributions even when thousands of parameters are present.

MCMC is a general type of computing technique based on drawing values ω from approximate distributions and then improving these draws to better approximate the target posterior distribution $P(\mathbf{\Omega}|\mathbf{X})$ (Gelman, Carlin, Stern, & Rubin, 2003). It is an important innovation in statistical computing in recent year such that it has made Bayesian inference more widely applied in a variety of disciplines. The key to MCMC is to create a Markov chain which is a stochastic process with the Markov property on a finite state space and with a stationary distribution approximately equivalent to the target posterior distribution $P(\mathbf{\Omega}|\mathbf{X})$. The state space equals the parameter space, the states are draws of the parameter and the Markov property is that the next state in the process only depends on the current state with all previous states irrelevant.

MCMC methods are all based on this same general idea and the difference among methods is how the transitions between states in a Markov chain are created. There are three major categories of MCMC sampling methods: the Gibbs sampler, the Metropolis-Hastings algorithm and the Metropolis algorithm. The Gibbs sampler is the simplest MCMC algorithm and it is the primary choice for conditionally conjugate models, whereas the Metropolis algorithm can be used for models that are not conditionally conjugate (Gelman et al., 2003). The Metropolis-Hastings algorithm is a generalization of the Metropolis algorithm.

The Gibbs sampler is also called alternating conditional sampling which constructs Markov chains by cycling through all full conditional distributions (i.e., the distribution conditional on all other parameters and the data) and is supposed to reach the joint stationary distribution that approximates the joint posterior distribution $P(\mathbf{\Omega}|\mathbf{X})$. Suppose that $\mathbf{\Omega} =$

(ω_1,ω_2,ω_3) and the joint posterior distribution is $P(\mathbf{\Omega}|\mathbf{X})$. The steps of sampling unknown parameters using the Gibbs sampler are as follows:

(1) Identify the full conditional distributions.

$p(\omega_1|\omega_2,\omega_3,\mathbf{X})$, $p(\omega_2|\omega_1,\omega_3,\mathbf{X})$, $p(\omega_3|\omega_1,\omega_2,\mathbf{X})$

(2) Provide starting values ($\omega_1^0, \omega_2^0, \omega_3^0$).

(3) Draw ω_1^1 from $p(\omega_1 \mid \omega_2 = \omega_2^0, \omega_3 = \omega_3^0, \mathbf{X})$, ω_2^1 from $p(\omega_2 \mid \omega_1 = \omega_1^1, \omega_3 = \omega_3^0, \mathbf{X})$ and ω_3^1 from $p(\omega_3 \mid \omega_2 = \omega_2^1, \omega_1 = \omega_1^1, \mathbf{X})$.

(4) Repeat step (3) until the chains are convergent.

(5) The draws after chain convergence are a sample from the stationary distribution.

The five steps presented above constitute the Gibbs sampler algorithm and they are based on the prerequisite that the full conditional distributions from which values are drawn are all distributions in closed forms.

However, when one or more of the full conditional distributions is not a closed form, a more general MCMC sampling algorithm is needed, such as the Metropolis-Hastings (M-H) algorithm. The M-H algorithm has been around for years as the Gibbs sampler, and the difference between the two is that the M-H algorithm uses a proposal distribution for drawing values rather than sequentially drawing values from the full conditional distributions as the Gibbs sampler does. The steps of sampling using the M-H algorithm are as follows (Gelman et al., 2003):

(1) Given a draw ω^{t-1} in iteration *t*-1, a proposal draw ω^* is sampled from a proposal distribution $J(\omega^* \mid \omega^{t-1})$.

(2) This draw is accepted with probability

$$r = \frac{p(\omega^{*} \mid \mathbf{X}) / J(\omega^{*} \mid \omega^{t-1})}{p(\omega^{t-1} \mid \mathbf{X}) / J(\omega^{t-1} \mid \omega^{*})}. \tag{2.15}$$

(3) This draw is accepted with the probability min(*r*, 1) (i.e., $\omega^{t} = \omega^{*}$) and is rejected with the probability max(0, 1-*r*) (i.e., $\omega^{t} = \omega^{t-1}$).

In the M-H algorithm, the proposal distribution can have any form and the posterior distribution is not required to have a closed form.

Additionally, as a special case of the M-H algorithm, the Metropolis algorithm assumes the proposal distribution to be symmetric (Gelman et al., 2003), that is,

$$J_{t}(\omega_{a} \mid \omega_{b}) = J_{t}(\omega_{b} \mid \omega_{a}) \tag{2.16}$$

$$\forall \omega_{a}, \omega_{b}, t.$$

This simplifies the probability *r* to be the ratio of $p(\omega^{*}|\mathbf{X})$ to $p(\omega^{t-1}|\mathbf{X})$. However, the M-H algorithm is usually more efficient than the Metropolis algorithm because the use of asymmetrical proposal distribution increases the speed of convergence to a stationary posterior distribution (Gelman et al., 2003).

2.3.2 Convergence diagnosis.

The key to a successful Bayesian estimation is that the Markov chains have converged to the target posterior distribution. If the chains do not converge, the inference about parameters based on the sampled iterations after burn-in would be invalid. Thus, it is important to diagnose convergence of Markov chains before making inferences. Chain convergence may be evaluated by several diagnostic criteria, such as time-series plots, autocorrelation plots, density plots and the Gelman-Rubin statistic *R* (Brooks & Roberts, 1998; Cowles & Carlin, 1996).

Previous literature has suggested a serious problem, label switching, to the convergence of Markov chains in Bayesian estimation of mixture IRT models (Cho & Cohen, 2010; Cho, Cohen, & Kim, 2006; Dai, 2009; Li, Cohen, Kim & Cho, 2009). The first type of label switching is between-chain label switching in which the order of latent classes switches across replications or for different initial values (Cho et al., 2006). This type of label switching is frequently reported in previous simulation studies and is also observed in the present study. It is not considered as non-convergence. Previous literature has suggested a variety of solutions to this problem, including artificial identification constraints (e.g., Diebolt & Robert, 1994), label invariant loss functions (e.g., Celeux, Hurn, & Robert, 2000), relabeling using a k-means type clustering (Celeux, 1998) and random permutation samplers (Fruhwirth-Schnatter, 2001). However, as these solutions are computationally intensive and may cause ambiguous explanations of parameter estimates, they are not commonly used. There are two other simple solutions to this problem. One is to fix the class memberships of a small number of individuals in the sample (Chung, Loken, & Schafer, 2004), and the other is simply to compare the parameter estimates with the generating parameters (i.e., item or mixing proportion parameters for mixture IRT models) to determine the label of each latent class if a simulation study is carried out (Cho & Cohen, 2010). The present study adopts the latter approach because the former one may affect the calculation of the accuracy of latent class identification, which is an evaluation criterion of the model performance in the current study.

A second type of label switching is within-chain label switching in which the estimated value of a model parameter switches within a Markov chain (Cho & Cohen, 2007; see Figure 2.4). Sometimes this type of label switching is accompanied with non-systematical fluctuations in different chains, which may result in poor mixing cases. Label switching of this type occurs when the log-posterior is invariant to different permutations of model parameters, and it could be detected by checking the density plot that the parameter has multiple modes. This type of problem could be taken as an indication that the model may not provide a good fit to the data.

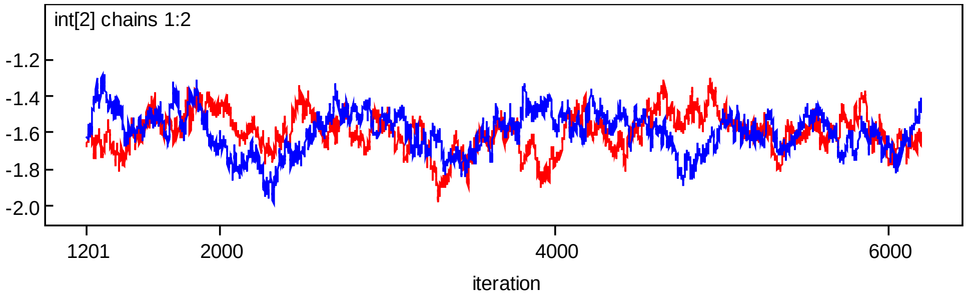

Figure 2.4. An example of within-chain label-switching.

Another type of non-convergence which is often observed in the estimation of complex models is that no mixing is observed between Markov chains. Figure 2.5 provides an example. It indicates that the likehood and the prior distributions do not offer enough information for model estimation.

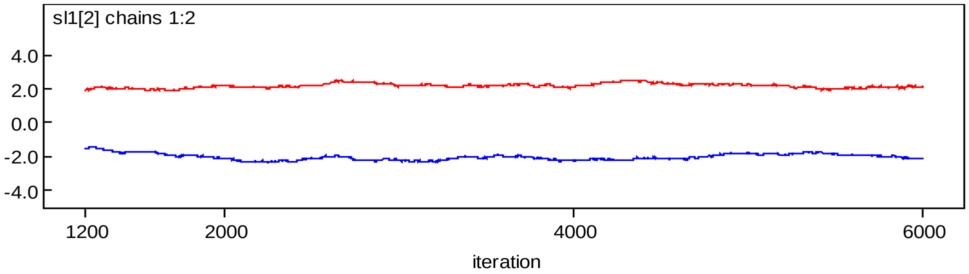

Figure 2.5. An example of no mixting between chains.

This issue is taken into consideration when the present study is designed. The amount of missingness simulated in response data is carefully selected so as to avoid this type of model non-convergence. The details of the factors of the simulation design and their corresponding levels are discussed in the next chapter. Figure 2.6 gives an example of converged Markov chains.

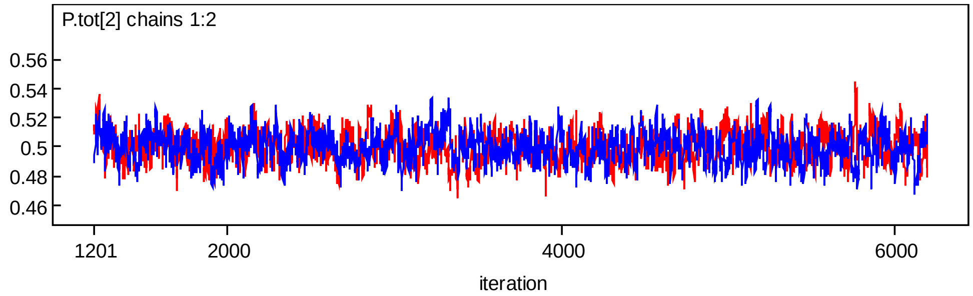

Figure 2.6. An example of converged Markov chains.

2.4 Summary

Given the fundamental relations among IRT, LCA and mixture IRT models, and the different approaches to covariate inclusion in each modeling approach, the present study takes the explanatory perspective and focuses on the impact of including dichotomous and continuous covariates on the estimation of the mixture Rasch model parameters. The estimation is carried out within a Bayesian framework.

Although previous simulation studies and empirical research have demonstrated that the inclusion of potentially important covariates may yield desirable psychometric properties in mixture IRT models. Certain areas remain unexplored in this line of research and thus give rise to the research questions that are of interest to the present study. Specifically, the research questions are presented as follows:

(1) What is the impact of including both a dichotomous covariate as predictor of the latent class membership and a continuous covariate as a predictor of the latent ability on the estimation of an MRM?

(2) What is the effect of misspecified covariates (e.g., mismatching covariates with model parameters) on fitting an MRM?

(3) What is the effect of covariate inclusion on overall model fit based on information-based model fit indices?

(4) What is the relative effect of covariate inclusion on the estimation of the MRM in complete versus incomplete data scenarios?

In the present study, both dichotomous and continuous covariates are included in the MRM as predictors for the latent class membership and the person parameter. Both complete and incomplete response data sets which approximate different types of missingness commonly observed in large-scale assessments are simulated. The impact of covariate specification is compared and analyzed in terms of model parameter recovery, latent class identification, and the relative overall model fit among alternative models. Finally, an illustration of applying the covariate inclusion approaches is demonstrated using the PISA 2009 reading assessment data.

In Chapter 3, a detailed description of the simulation design and the real data application is provided. Before the introduction of the simulation study and the real data example, the data generating model and five MRMs with misspecified covariates are detailed. Technical issues regarding the implementation of model estimation in WinBUGS are also discussed in this chapter.

Chapter 3 Methodology

The first chapter introduced the motivation of including dichotomous and continuous covariates as predictors of model parameters for the MRM and the potential contributions of covariate information. The second chapter reviewed the theoretical foundations of mixture IRT models as well as different covariate inclusion approaches in relevant modeling frameworks. This chapter focuses on four methodological issues of the MRM with covariates: the approaches to covariate inclusion under investigation in the present study, the implementation of model estimation in WinBUGS, the design of the simulation study and the analysis plan for the real data example.

3.1 Different Approaches to Covariate Inclusion in MRM

The MRM (Kelderman & Macready, 1990; Mislevy & Verhelst, 1990; Rost, 1990) assumes different latent classes and the Rasch model with different item parameters (i.e., item difficulties) holds within each class. In the MRM, the unconditional probability of a correct response from person *j* to item *i* is specified as

$$P(X_{ij}=1\,|\,\theta_{jg},b_{ig})=\sum_{g}\pi_{g}\frac{1}{1+\exp[-(\theta_{jg}-b_{ig})]}, \tag{3.1}$$

where π_g denotes the mixing proportion corresponding to the percentage of persons in each latent group g, θ_{jg} is a person's latent ability, and b_{ig} is the item difficulty for latent group g. Figure 3.1 shows the graphical representation of a MRM without covariates. The arrows from the latent group variable g to the items show that the item difficulty parameters are different for latent classes. If no difference is assumed among the latent groups with regard to the distribution of the latent ability, the arrow is not needed from the latent class variable g to the latent ability θ, as shown in the M*plus* manual (Muthén & Muthén, 1998-2012). On the other hand, if a mean difference in the person ability is assumed among the latent groups, there is an arrow from the latent class variable g to the latent ability θ (Tay et al., 2011). In the present study, the latter expression is adopted.

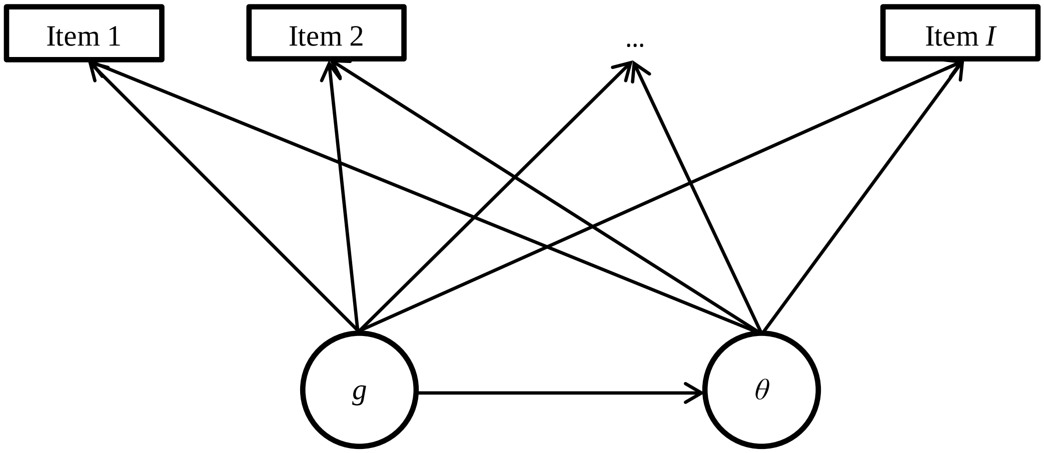

Figure 3.1. The mixture Rasch model.

In the present study, the covariates enter the MRM either as predictors of π_{jg}, the probability of a person belonging to a latent class (i.e., $\sum_g \pi_{jg} = 1$), or as predictors of the latent trait θ_{jg}. In the true model for data generation, a dichotomous covariate enters the model as a predictor of π_{jg} through a logistic function:

$$\pi_{jg} = \frac{\exp(\beta_{0g} + \beta_{1g} D_j)}{\sum_{g=1}^{G} \exp(\beta_{0g} + \beta_{1g} D_j)} \tag{3.2}$$

where D_j indicates the dichotomous covariate, such as gender, and β_{0g} and β_{1g} are corresponding regression coefficients in the logistic function. For model identification purpose, both β_{01} and β_{11} are fixed as 0.

Also, a continuous covariate enters the MRM as a predictor of the latent trait through a linear regression function:

$$\theta_{jg} = \alpha_{0g} + \alpha_{1g} C_j + e_{jg} \tag{3.3}$$

where C_j indicates the continuous covariate (e.g., intelligence or motivation), α_{0g} and α_{1g} indicate the intercept and the slope of the latent regression model corresponding to latent group g, and e_{jg} is the error term with the distributional assumption that $e_{jg} \sim N(0, \sigma_{eg}^2)$.

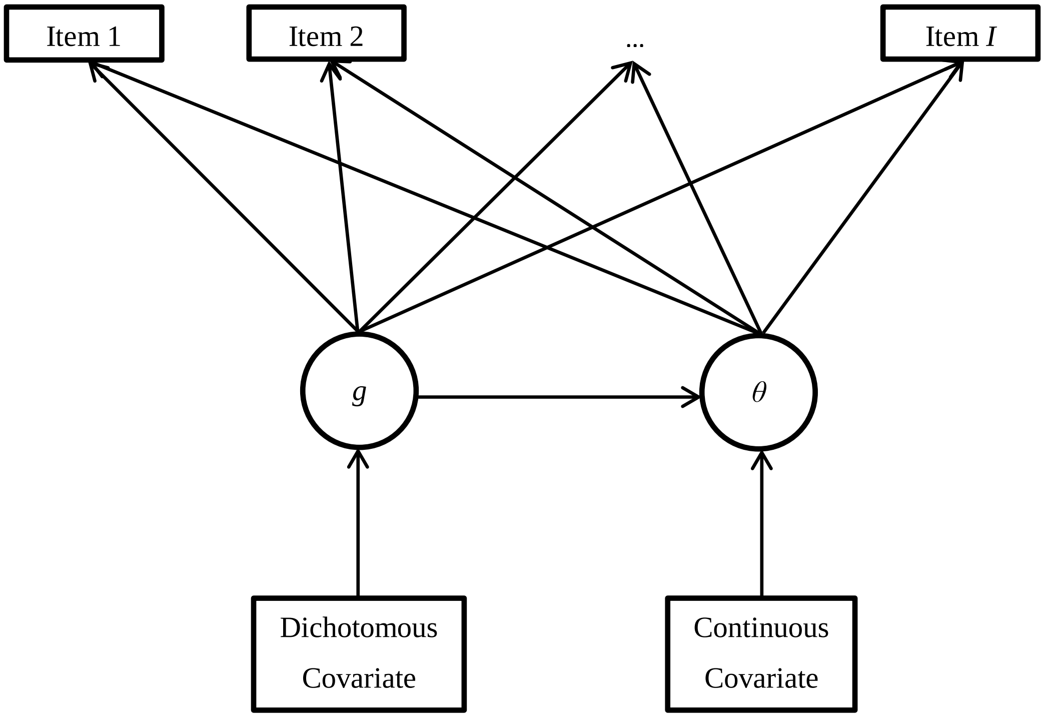

Figure 3.2. The data generating model.

In the present simulation study, two latent classes are assumed (G =2). Figure 3.2 provides the graphical representation of the data generating model. Equation 3.4 gives the mathematical expression of the expanded version of the data generating model.

$$P(X_{ij}=1\,|\,\theta_{jg},b_{ig})=\sum_{g}\left(\tfrac{1}{J}\sum_{J}\frac{\exp(\beta_{0g}+\beta_{1g}D_j)}{\sum_{g=1}^{G}\exp(\beta_{0g}+\beta_{1g}D_j)}\right)\left(\frac{1}{1+\exp\{-[(\alpha_{0g}+\alpha_{1g}C_j+e_{jg})-b_{ig}]\}}\right) \quad (3.4)$$

Six alternative models are used in the present study to fit the simulated data: the true model, an overspecified model that relates both covariates to both model parameters (i.e., π_{jg} and θ_{jg}), three underspecified models that are three constrained cases of the true model, and a model with mismatching covariates. In the model with mismatching covariates, C_j enters the model as a predictor of π_{jg} in a logistic function and D_j as a predictor of θ_{jg} in a linear function. Although covariates, in practical settings, are usually correlated, the continuous covariate and the dichotomous covariate explored in the present study are assumed to be independent. In the present study, the linear function and the logistic function are respectively used to link the continuous and dichotomous covariates and the model parameters because these two linking functions are typical in relevant studies (e.g., Adams et al., 1997; Dai, 2009, 2013). However, the functional form is not limited to these two types. Theoretically, any appropriate regression functions, such as polynomial regression or nonlinear regression, could be used to link covariates with model parameters, and interactions among covariates may be allowed. Different variable selection methods used in regression analysis could also potentially be used for the selection of the covariates in the measurement model. Future research may incorporate and compare different functions which link covariates and model parameters. Table 3.1 presents the mathematical functions of the six models under investigation in the current study.

Table 3.1. *True data-generating model and alternative models.*

Model Type	Model Specification
True model (TM)	the MRM with $\pi_{jg} = \frac{\exp(\beta_{0g} + \beta_{1g} D_j)}{\sum_{g=1}^{G} \exp(\beta_{0g} + \beta_{1g} D_j)}$ and $\theta_{jg} = \alpha_{0g} + \alpha_{1g} C_j + e_{jg}$ where $\beta_{01} = \beta_{11} = 0$
Overspecified model (OM)	the MRM with $\pi_{jg} = \frac{\exp(\beta_{0g} + \beta_{1g} D_j + \beta_{2g} C_j)}{\sum_{g=1}^{G} \exp(\beta_{0g} + \beta_{1g} D_j + \beta_{2g} C_j)}$ and $\theta_{jg} = \alpha_{0g} + \alpha_{1g} C_j + \alpha_{2g} D_j + e_{jg}$ where $\beta_{01} = \beta_{11} = \beta_{21} = 0$
Model with mismatch covariates (MISM)	the MRM with $\pi_{jg} = \frac{\exp(\beta_{0g} + \beta_{1g} C_j)}{\sum_{g=1}^{G} \exp(\beta_{0g} + \beta_{1g} C_j)}$ and $\theta_{jg} = \alpha_{0g} + \alpha_{1g} D_j + e_{jg}$ where $\beta_{01} = \beta_{11} = 0$
Underspecified models (UNM)	
1. UNM-N	the MRM without covariates
2. UNM-D	the MRM with $\pi_{jg} = \frac{\exp(\beta_{0g} + \beta_{1g} D_j)}{\sum_{g=1}^{G} \exp(\beta_{0g} + \beta_{1g} D_j)}$ where $\beta_{01} = \beta_{11} = 0$
3. UNM-C	the MRM with $\theta_{jg} = \alpha_{0g} + \alpha_{1g} C_j + e_{jg}$

3.2 WinBUGS Implementation of Model Parameter Estimation

In this study, R (version 2.15.2) is used to generate data sets based on the hypothesized distributions (i.e., the distributions for the person parameters, the item difficulty parameters and the covariates of interest) and the details about the design for data generation are provided in later sections. R2WinBUGS package in R is employed to interface with WinBUGS to carry out the Bayesian estimation of the item and person parameters, mixing proportions, respondents' latent group membership and the relations between covariates and model parameters.

Based on the Bayesian perspective of mixture IRT models provided by Mislevy, Levy, Kroopnick and Wise (2006), the data generating model in the present study may be expressed in the Bayesian framework as follows:

$$P(\boldsymbol{\theta},\mathbf{b},\boldsymbol{\varphi},\boldsymbol{\beta},\boldsymbol{\alpha} \mid \mathbf{X},\mathbf{C},\mathbf{D}) \propto P(\mathbf{X} \mid \boldsymbol{\theta},\mathbf{b},\boldsymbol{\varphi})P(\boldsymbol{\theta} \mid \mathbf{C},\boldsymbol{\alpha})P(\boldsymbol{\varphi} \mid \boldsymbol{\beta},\mathbf{D})P(\mathbf{b} \mid \boldsymbol{\varphi})P(\boldsymbol{\varphi})P(\boldsymbol{\alpha})P(\boldsymbol{\beta}). \qquad (3.5)$$

In this expression, $P(\mathbf{X}|\boldsymbol{\theta},\mathbf{b},\boldsymbol{\varphi})$ is the likelihood function of the measurement model, which denotes the probability of item responses conditional on individual latent class, person parameters, and item parameters. It is thus defined as:

$$P(\mathbf{X} \mid \boldsymbol{\theta},\mathbf{b},\boldsymbol{\varphi}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{g=1}^{G}\left\{(\frac{1}{1+\exp[-(\theta_{jg}-b_{ig})]})^{x_{ij}}(1-\frac{1}{1+\exp[-(\theta_{jg}-b_{ig})]})^{1-x_{ij}}\right\}^{\varphi_{jg}}. \qquad (3.6)$$

$\boldsymbol{\theta}$ and $\mathbf{b}$ are the vectors of person parameters and item parameters, respectively. $\boldsymbol{\varphi}$ is a design matrix indicating latent class membership with

$$\boldsymbol{\varphi} = \begin{bmatrix} \varphi_{11} & \cdots & \varphi_{J1} \\ \vdots & \ddots & \vdots \\ \varphi_{1G} & \cdots & \varphi_{JG} \end{bmatrix}, \qquad (3.7)$$

where φ_{jg} takes the value of 1 if person j belongs to latent class g, and 0 otherwise. For example, if there are two latent classes (i.e., G=2) and two persons (i.e., J=2), with the first person belonging to LC1 and the second person belonging to LC2, the $\boldsymbol{\varphi}$ matrix is specified as $\boldsymbol{\varphi} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Equation 3.5 is based on conditionally independent item responses given item parameters, person parameters and latent class membership. Further, $P(\boldsymbol{\theta}|\mathbf{C},\boldsymbol{\alpha})$ in Equation 3.4

provides the distribution of the person parameter conditional on the continuous covariate data matrix **C** and the linear regression parameters **α**, and $P(\boldsymbol{\varphi}|\boldsymbol{\beta},\mathbf{D})$ is the distribution of latent class membership conditional on the dichotomous covariate data matrix **D** and the logistic regression parameters **β**. Additionally, $P(\mathbf{b}|\boldsymbol{\varphi})$ is the distribution for the item parameter conditional on latent class membership, and $\mathrm{P}(\boldsymbol{\varphi})$ is the prior distribution for mixing proportion. $P(\boldsymbol{\alpha})$ and $P(\boldsymbol{\beta})$ are the prior distributions for the regression coefficients. Possible hyper-parameters may be defined in the model if necessary. Finally, the joint posterior distribution $P(\boldsymbol{\theta},\mathbf{b},\boldsymbol{\varphi},\boldsymbol{\alpha},\boldsymbol{\beta}|\mathbf{X},\mathbf{C},\mathbf{D})$ is obtained in the left part of Equation 3.4.

Chapter 2 has provided a discussion of the major MCMC sampling methods in Bayesian estimation. In WinBUGS, the primary method that is used is the Gibbs sampler. The sampling method employed by WinBUGS corresponds to a hierarchy such that a method is only used if no previous method is appropriate (Spiegelhalter et al., 2003). Table 3.2 presents the sampling hierarchy used by WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000; Spiegelhalter et al., 2003).

Table 3.2. *The hierarchy of sampling methods used by WinBUGS.*

Target distribution		Sampling method
Discrete	Finite upper bound	Inversion of cumulative distribution function
	Shifted Poisson	Direct sampling using standard algorithms
Continuous	Conjugate	Direct sampling using standard algorithms
	Log-concave	Derivative-free adaptive rejection sampling
	Restricted range	Slice sampling
	Unrestricted range	the M-H algorithms

As the MRM has both discrete and continuous latent variables, a combination of three sampling methods, the inversion method, the direct sampling and the derivative-free adaptive rejection sampling, are used (Cho & Cohen, 2010). The derivative-free adaptive rejection sampling is a type of the Gibbs sampler (Gilks, 1992).

In Bayesian estimation, starting values are required for each parameter being sampled to define the first state of each Markov chain (Gelman et al., 2003). In the present study, the starting values for all model parameters are randomly generated by the WinBUGS software. Since the initial sampled states in the Bayesian estimation are influenced by the starting values, a number of initial states (i.e., burn-in iterations) need to be discarded, and the estimates of item and person parameters are the means over the sampled iterations starting from the next iteration after the burn-in period (Kim & Bolt, 2007). For the latent class membership, the estimates are the modes of the sampled iterations after burn-in. To derive the posterior distributions for each model parameter, the following prior distributions are used for the estimation of the data generating model:

$$\begin{aligned}
&b_{i1} \sim Normal(0,1)\\
&b_{i2} - b_{i1} \sim Normal(0,.5)\\
&\tau_g \sim Gamma(.5,1)\\
&\beta_{0g} \sim Normal(0,1)\\
&\beta_{1g} \sim Normal(0,1)\\
&\alpha_{0g} \sim Normal(0,1)\\
&\alpha_{1g} \sim Normal(0,1),
\end{aligned}$$

where G=2. For the person parameter θ_{jg}, it is not directly estimated in the true model. It is decomposed by the linear regression function as shown in Equation 3.3, and the intercept parameter, slope parameter and variance of the error term are estimated instead. τ_g indicates the precision of the error term with $\mathrm{Var}(\theta_{jg}) = \mathrm{Var}(e_{jg}) = \sigma_{eg}^2$ and $\tau_g = 1/\sigma_{eg}^2$. For the MRM without covariates, additional prior distributions are used.

$$\begin{aligned}
&\theta_{jg} \sim Normal(\mu_g, \tau_g)\\
&\mu_g \sim Normal(0,1)\\
&(\pi_1, \pi_2) \sim Dirichlet(5,5)
\end{aligned}$$

These prior distributions are not highly informative. They are selected based on relevant mixture IRT research (e.g., Cho & Cohen) and the distributions for data generation used in the present study. Further, in the estimation, the sum of the item difficulty parameters within each latent class is constrained to be zero (i.e., $\sum_{i}^{I} b_{ig} = 0$) for model identification purpose.

Two chains of 31000 iterations are run, and the burn-in cycle for each chain is 6000. To reduce serial dependencies across iterations, a thinning of 5 is used. Thus, the final posterior sample size is 10000 (5000 iterations in each chain) on which model estimates are based. The number of iterations and the burn-in cycle are determined based on relevant research (e.g., Li et al., 2009) in this area and the results from the preliminary study. Each simulation run took up to 12 hours on a standard desktop PC with an Intel Core i7 3.40GHz processor. In the present study, no convergence problems discussed in Section 2.3.2 have been observed in any replications and the all of the results are based on converged estimation runs. Between-chain label switching has been observed for some replications, and this problem is handled by comparing the parameter estimates with the generating parameters to determine the correct latent class label.

3.3 Simulation Design

The present simulation study intends to explore the impact of different approaches to covariate inclusion on model parameter recovery, latent class assignment, and the overall model fit in the MRM context. Both complete and incomplete response data scenarios are considered.

3.3.1 Fixed factors.

To keep the simulation study manageable, certain factors are held constant in the simulation design, including the number of classes, the test length, the total number of subjects, the distribution of subjects' latent ability, and the distribution of covariates. Table 3.3 provides the factors that are fixed at a single level in the present study.

Table 3.3. *Factors that are fixed at a single level in the simulation.*

Factor	Fixed Value
The number of latent classes	2
The test length	30
The total number of subjects	2000
The distribution of subjects' latent ability	LC1: N(0,1); LC2: N(1,1)
The distribution of the covariate	dichotomous: 30%:70% continuous: LC1: N(0,1); LC2: N(0,1)

In the present study, the number of latent classes is set at two according to previous simulations in this line of research (e.g., Dai, 2009; Samuelsen, 2005; Smit et al., 1999, 2000). However, the exploration on covariates inclusion can be extended to more than two latent class scenarios using multinomial logistic functions as illustrated in concomitant-variable latent-class models (Dayton & Macready, 1988, 1989). The present simulation focuses on two latent classes as the first stage of investigation.

A total sample of 2,000 respondents responding to 30 dichotomously-scored items are simulated. It is a reasonable test length that is often seen in large-scale educational assessments. Also, the number of respondents is fixed at 2,000 to ensure that the model parameters could be accurately estimated so that the analysis of model performance would not be affected by the imprecision in model parameter estimates. The person parameters are drawn from a standard normal distribution Normal(0, 1) for one latent group and a normal distribution Normal(1, 1) for the other. The person parameters may be drawn from the same distributions or different distributions, as suggested by previous mixture IRT literature (e.g., Dai, 2009, 2013). In the present study, a mean difference of 1 is set for the person parameters of the two latent classes so that the estimation of the MRM can converge more easily. Some previous studies have manipulated test length and the number of respondents. However, because the present study focuses on different approaches to covariate inclusion and their corresponding impacts, these two factors are fixed to make the current simulation study manageable.

Additionally, the proportions of the dichotomous covariate are set to be .30 and .70 based on a previous study (Dai, 2009). In other words, 30% of the respondents are assigned a value of 0 on the covariate and 70% of the respondents are assigned a value of 1. The values of the continuous covariate are drawn from a normal distribution Normal(0, 1) respectively for the two latent classes.

3.3.2 Manipulated factors.

Other factors, including the distribution of latent classes, the average DIF effect size, the strength of relations between covariates and model parameters, the response data (i.e., complete response data and incomplete response data with different types of missingness), and the types of models for comparison purpose, are manipulated as shown in Table 3.4.

Table 3.4. *Manipulated factors in the simulation.*

Factor	Corresponding Values
model type (6 models)	true model over-specified model underspecified models (3) a model with mismatching covariates
mixing proportion (LC1% ; LC2%)	(50%; 50%) (30%; 70%)
strength of the relation between D_j and π_{jg}	strong (odds ratio = 10) weak (odd ratio=1)
strength of the relation between C_j and θ_{jg}	LC1: α_{11} = .2; LC2: α_{12} = .2 LC1: α_{11} = .8; LC2: α_{12} = .8
average DIF(i.e., the mean of $\|b_{i1}$-$b_{i2}\|$)	1.5 1.0
response data completeness (3 types)	complete data incomplete data with booklet design incomplete data with conditional omitted response

As for the mixing proportion, two levels (i.e., LC1%:LC2% = 50%:50% or 30%:70%) are considered. Extremely unequal mixes of latent classes are not included in the present study so as to ensure that the parameters could be accurately recovered for each latent group.

For the strength of relations between covariates and model parameters, odds ratios are used to indicate the magnitude of association between the dichotomous covariate and the latent class membership (Dai, 2009, 2013). In the DIF context, the odds ratio indexes the strength of the relation between manifest grouping variables and latent classes:

$$\text{odds ratio (OR)} = \frac{(P(g=1 \mid D_i = 0) / P(g=2 \mid D_i = 0))}{(P(g=1 \mid D_i = 1) / P(g=2 \mid D_i = 1))}. \tag{3.8}$$

Specifically, when the odds ratio equals 1, the covariate has no effect on the latent class membership; when the odds ratio equals 10, the relation between the covariate and the latent class membership is fairly strong. The details of using OR to generate the latent class membership and the values of the dichotomous covariate are presented in Appendix A. Regarding the relation between the continuous covariate and the latent trait in Equation 3.3, α_{1g} denotes the magnitude of the correlation. The strength of the correlation is also manipulated at two level as either weak (i.e., α_{11}=.2; α_{12}=.2), or strong (i.e., α_{11}=.8; α_{12}=.8).

The item parameters are drawn respectively from the distribution Normal(0,1) and then reordered to create two levels of average DIF effect sizes (i.e., the mean of $|b_{i1}$-$b_{i2}|$). When the average DIF effect size equals 1.5, 80% of the items have a difference in item difficulty greater than 1.0 between the two latent classes; when the average DIF effect size equals 1.0, 40% of the items have a difference in item difficulty greater than 1.0. These two relatively large DIF effect size levels are chosen based on the preliminary study in consideration of both the complete and incomplete response data scenarios. Both DIF size levels in the present simulation are quite large to ensure that the latent structure and parameters could be accurately recovered. The generated item parameters represent a wide range of parameter values that are observed in operational tests. The two sets of item difficulty values used to generate the item response data are both presented in Table 1 in Appendix A.

As shown in the table, the item parameters are specified for the corresponding simulation conditions as opposed to a random generation of different sets of item parameters based the distributions assumed. Given that the generation of item response data based on specified model parameter is in itself a random process, allowing the generation of item parameters to be random may bring more sampling error into the data generation. Thus, the specified item parameters with desired average DIF effect sizes are used in the current simulation to remove the potential sampling error in the data generation process.

Additionally, in the present study, both complete response data and incomplete response data with different types of missingness are considered because missing data scenarios are prevalent in practical assessments settings. The reasons for missing responses may generally be classified into two major categories: missingness by test design such as with matrix-sampled booklets, and nonresponses such as with omitted and not-reached items (Ludlow & O'Leary, 1999). Not-reached items usually occur when examinees fail to complete a test within a given time, whereas omitted items are associated with examinees' low ability levels or lack of motivation in low-stake assessments (e.g., De Ayala, Plake, & Impara, 2001). In this study, two general types of missing data are of primary interest: the missingness by design through balanced incomplete block spiraling (BIB) which is implemented in many large-scale assessments, such as the National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA), and the missingness by omitted items with low-ability individuals omitting difficult items which are essentially conditional missing. The not-reached item scenario is not considered in the present study, because it is suggested that not-reached items are not used in item calibration or scaling in practical settings (Lord, 1980). The missingness by test design is considered as missing completely at random, whereas the nonresponse by omitted items is considered missing not at random (Finch, 2008).

For the missingness due to booklet design, one condition is simulated based on a previous study (von Davier, Gonzalez, & Mislevy, 2009) and practical test settings (i.e. NAEP; NCES, 2009): items are randomly assigned to one of three blocks named A, B, and C, and each person responds to two of these blocks (i.e., a total of 20 items out of 30 items) such that the booklets are organized as (AB)(BC)(CA), to which are responded by 667, 667, and 666 examinees, respectively. The total proportion of missing data in this condition is .33.

Regarding the other type of missingness, omitted responses, previous literature indicates that this type of missingness usually affects 10% to 50% of the items in a test (e.g., Chen & Jiao, 2012; Finch, 2011). Thus, the omitted responses are simulated according to the upper bound: a total of 400 respondents omit 50% of the items (i.e., 15 items). The total proportion of missing data in this condition is .10. In both types of missingness, missing data only occur in the item responses but not in the covariate information.

In summary, all the levels of the manipulated factors are carefully chosen based on both the previous literature in this line of research and the preliminary simulation runs. Certain extreme levels, such as a large amount of missing data (i.e., 60%), extremely unequal latent classes (i.e., 15%:85%) and small DIF size (i.e., 0.5), are excluded from the present design because they have been found to result in serious convergence issues (i.e., non-mixing or within chain label switching) in the preliminary study.

The present study includes 2×2×2×2×3=48 simulation conditions. With 25 data sets for each condition, there are 1200 data sets generated. For each data set, 6 models are used to fit the data. Thus, there are a total of 6×48=288 simulation cells with 25×288=7200 replications. Table 2 in Appendix A presents a complete list of the simulation conditions for data generation in the present study. In the preliminary study, 100 replications are conducted in one condition for the data-generating model. Figure 3.3 and 3.4 present the standard errors of the item parameters for the two latent classes as a function of the number of replications. Three items, including an easy, a medium and a difficult item, are selected from each latent class. Although the average standard errors show different decreasing or increasing patterns when the replication number ranges between 2 to 15, they tend to be stable when the replication number reaches 20.

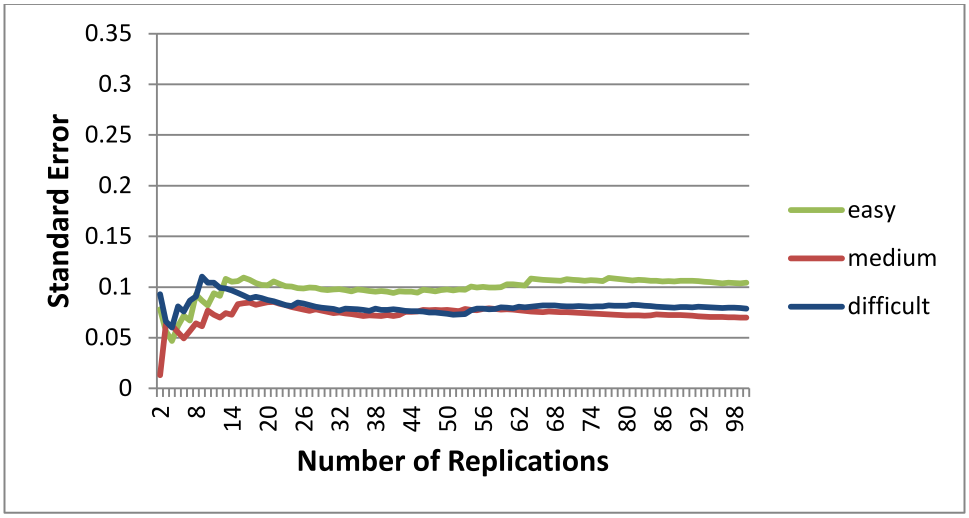

Figure 3.3. The standard errors for the item parameters by the number of replications (LC1).

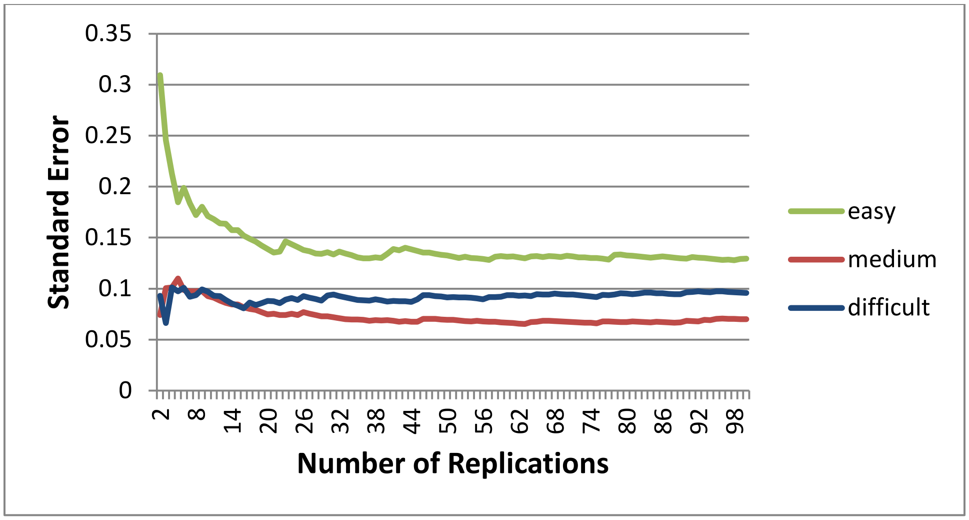

Figure 3.4. The standard errors for the item parameters by the number of replications (LC2).

Figure 3.5 displays the standard errors of the person parameters for a low-ability, a middle-ability, and a high-ability person. The patterns of the average standard errors for the person parameters are similar to those for the item parameters. Figure 3.6 shows the bias for the same three persons, and the average bias also tends to be stable when the replication number is around 20.

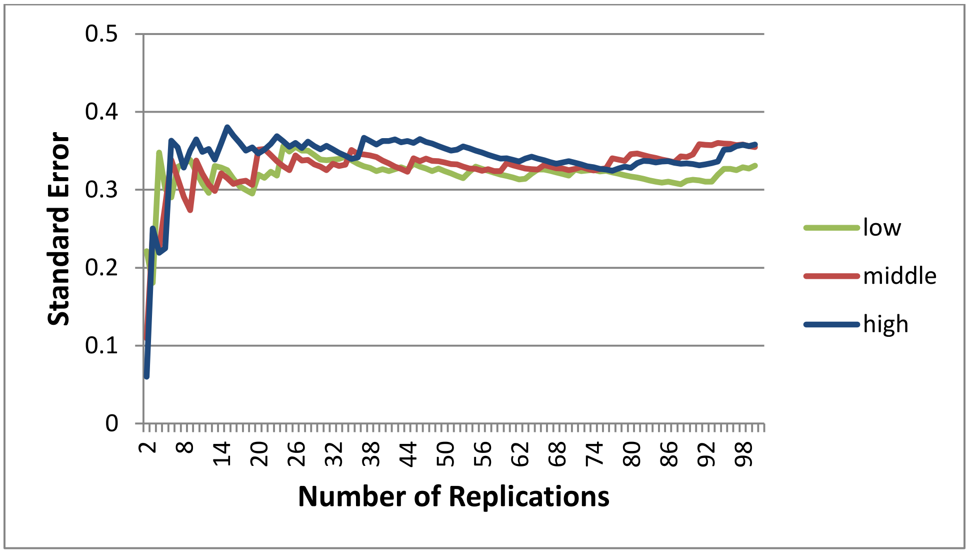

Figure 3.5. The standard errors for the person parameters by the number of replications.

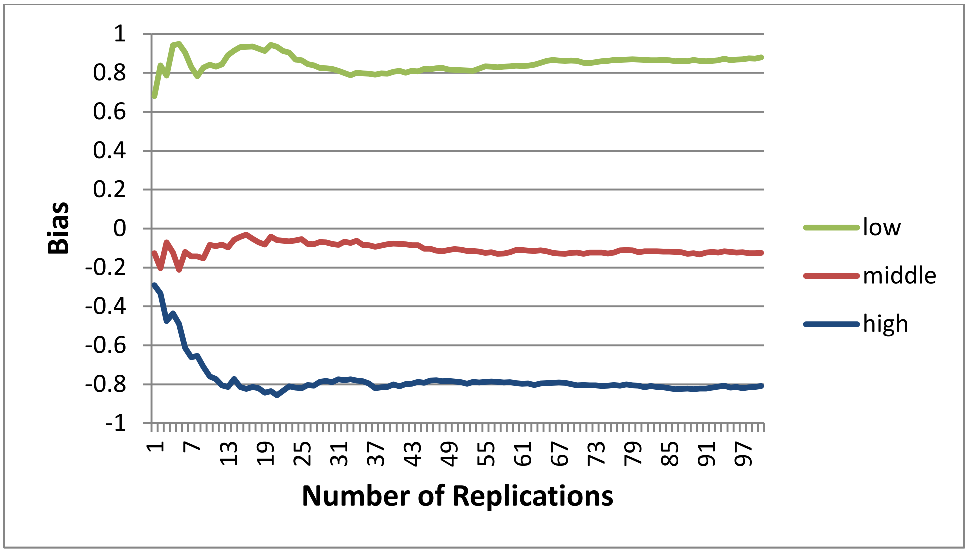

Figure 3.6. The bias for the person parameters by the number of replications.

Additionally, the average standard errors for the mixing proportion by the number of replications are presented in Figure 3.7. There is little fluctuation after the number of replications exceeds 15. Given all the plots displayed above and suggestions from previous literature on Bayesian estimation of IRT models (e.g., Jiao et al., 2012), 25 replications per cell are adopted by the present study considering the large amount of time required for the Bayesian estimation.

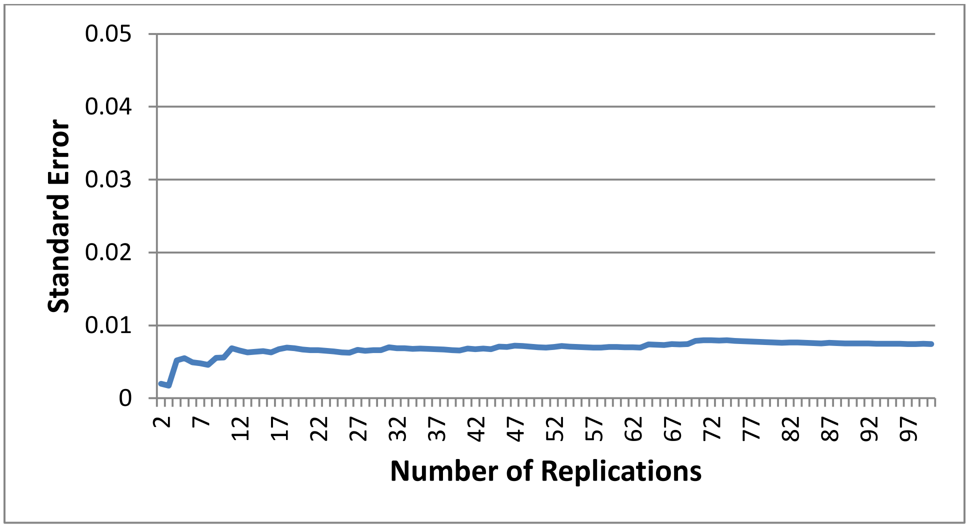

Figure 3.7. The standard errors for the mixing proportion by the number of replications.

3.3.3 Evaluation criteria.

The estimation effectiveness are analyzed in three outcomes: a) the accuracy of latent group classification, b) the accuracy of parameter recovery, and c) the overall model fit as indicated by the proportion of correct model selections.

Latent group classification. The accuracy of the latent group classification is assessed using the proportion of subjects that are assigned to their true latent class based on their estimated latent class membership.

Parameter recovery. The recovery of model parameters is evaluated with respect to four properties of the parameter estimates: a) the proportion of replications for which the 95% confidence interval around the item and person parameter estimates covered the true value for each simulation cell, b) the bias for the item, person and regression coefficient parameter estimates, if applicable, c) the standard error (SE) of the item, person parameter and regression coefficient estimates (i.e., precision), if applicable and d) the root mean squared error (RMSE) of the item, person parameter and regression coefficient estimates, if applicable. The mathematical equations for bias, SE and RMSE are provided in Equation 3.9 to 3.11 as follows:

$$\text{bias}(\hat{\omega}, \omega) = \frac{\sum_{rep=1}^{R} (\hat{\omega} - \omega)}{R} \tag{3.9}$$

$$\text{SE}(\hat{\omega}, \overline{\hat{\omega}}) = \sqrt{\frac{\sum_{rep=1}^{R} (\hat{\omega} - \overline{\hat{\omega}})^2}{R - 1}} \tag{3.10}$$

$$\text{RMSE}(\hat{\omega}, \omega) = \sqrt{\frac{\sum_{rep=1}^{R} (\hat{\omega} - \omega)^{2}}{R}} \tag{3.11}$$

where *R* is the number of replications in each simulation cell, $\hat{\omega}$ is the posterior estimate of a model parameter and $\overline{\hat{\omega}}$ is the mean of a model parameter estimates across replications. Different from the conventional way of computing SE in simulation studies, (*R*-1) is used in the present study as the denominator for the calculation of SE, as a correction for the degree of freedom when the replication number is small; otherwise, there would be a downward bias in the SE.

The SE represents random error in the estimation, whereas the bias represents systematic error. Asymptotically, the mean squared error (i.e., the squared RMSE) for each person or item may be decomposed into the variance (i.e., the squared standard error) and the squared bias. For unbiased estimator, the mean squared error equals the variance. However, for marginal statistics (i.e., RMSE, SE and bias averaged across persons or items), this relation does not hold. Although RMSE is considered as a combination of SE and bias, it has been suggested in IRT literature (e.g., Jiao et al., 2012) that RMSE have the potential of selecting the better fitting model (i.e., true model). In this regard, the RMSE may provide additional information above and beyond the bias and SE, and it is thus included in the present study. In the present, the average bias for the item parameter is supposed to be 0 because the item parameter is used for scale identification.

Overall model fit indices. The following fit statistics are also obtained for each model under different simulation conditions: Akaike's information criterion (AIC; Akaike, 1987), Bayesian Information Criterion (BIC; Schwartz, 1978), a correction of AIC based on sample size and the number of parameters (AICc; Burnham & Anderson, 2002), the consistent AIC (CAIC; Bozdogan, 1993), the sample-size adjusted BIC (SABIC; Sclove, 1987) and deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linden, 2002). These fit indices are the most commonly used indices for model selection purpose in mixture IRT, as well as in growth/factor mixture modeling literature. Among these information-based model fit indices, AIC only penalizes for the number of parameters and it tends to be negatively biased as the ratio of the number of parameters to sample size increases. BIC, AICc, CAIC and SABIC penalize for sample size and the number and parameters in different ways. AICc corrects for small samples and is preferred to be used when the ratio of sample size to the number of parameters is less than 40 (Burnham & Anderson, 2002). CAIC has a larger penalty for over-parameterization than AIC does and it is an asymptotically unbiased criterion (Bozdogan, 1993). DIC is a model fit index designed for Bayesian posterior estimates of model parameters. Some research (e.g., Li et al., 2009; Magidson & Vermunt, 2004) has recommended the use of BIC for mixture distribution model selection, because it outperforms other indices in terms of consistency; yet the choice of different model fit indices is still inconclusive because their relative performance of information indices is sometimes model and design specific. In the present study, 6 indices are included in order to provide a comprehensive overview of the model fit with regard to different approaches to covariate inclusion.

$$\mathrm{AIC} = \overline{D(\omega)} + 2p \tag{3.12}$$

$$\mathrm{BIC} = \overline{D(\omega)} + p(\ln N) \tag{3.13}$$

$$\mathrm{AICc} = \overline{D(\omega)} + \frac{2np}{n-p-1} \tag{3.14}$$

$$\mathrm{CAIC} = \overline{D(\omega)} + p(\ln N + 1) \tag{3.15}$$

$$\mathrm{SABIC} = \overline{D(\omega)} + p(\ln((N+2)/24)) \tag{3.16}$$

$$\mathrm{DIC} = \overline{D(\omega)} + p_D \tag{3.17}$$

where $\overline{D(\omega)}$ is the posterior mean of the deviance between the data and a model in Bayesian estimation, p is the number of parameters, N is the sample size and p_D denotes the difference between $\overline{D(\omega)}$ and $D(\hat{\omega})$. As model fit indices are not provided in the WinBUGS output, they are calculated in *R* (version 2.15.2) for the present study to select the best fitting model.

3.4 Real Data Examples

The real data analyses are based on the PISA 2009 U.S. data from students' reading assessment. This data set is selected because the Rasch model is used for the calibration of PISA dichotomous items (OECD, 2009). No specific inferences about items or respondents are drawn based on the real data examples in the present study and they are only used to illustrate different approaches to covariate inclusion in the MRM applications.

The first sample consisting of complete response data from 1,525 15-year-old students to 16 dichotomously-scored reading items is extracted from this data set. The items are obtained from booklet 2, 4, 5 and 7. As the first step of investigation, *ESCS* (i.e., a weighted likelihood estimate of student's economic, social and cultural status) is used as the continuous covariate. As for the dichotomous covariate, the preliminary study has explored a number of categorical variables, including gender (i.e., male or female), immigrant status (i.e., native, first-generation, second-generation), language at home (i.e., language at test or not), school type (public or private) and reading enjoyment time. Finally, *reading enjoyment time* is chosen as a moderately informative covariate. The variable *reading enjoyment time* is a five-category variable (i.e., 1=I don't read for enjoyment; 2=30 minutes or less a day; 3=between 30 and 60 minutes; 4=1 to 2 hours a day; 5=more than 2 hours a day) and is thus dichotomized for use in the present study (0=30 minutes or less; 1=more than 30 minutes). No missing data (i.e., missing by nonresponse or booklet design) are included in this sample.

The second sample is based on the dichotomous item response data with missingness by booklet design (i.e., 7=not administered) included, and the three covariates mentioned above. A subject is deleted either if any of the covariates has a missing value or any item response is missing as a non-reached item (i.e., 8=unreached). This sample includes a total of 4,892 students responding to 123 dichotomous items.

Ten models are fitted to the two samples and estimates of all the model parameters are obtained. The ten models include: 1) the Rasch model (RASCH); 2) the Rasch model with *ESCS* as the predictor of the person parameter (RASCH-C); 3) the Rasch model with *reading enjoyment time* as the predictor of the person parameter(RASCH-D); 4) the Rasch model with both *reading enjoyment time* and *ESCS* as predictors of the person parameter (RASCH-CD); 5) the two-class MRM without covariates (UNM-N); 6) the two-class MRM with *ESCS* as the predictor of the person parameter (UNM-C); 7) the two-class MRM with *reading enjoyment time* as the predictor of the latent class membership (UNM-D); 8) the two-class MRM with *ESCS* as the predictor of the person parameter and *reading enjoyment time* as the predictor of the latent class membership (TM); 9) the two-class MRM with both *reading enjoyment time* and *ESCS* as predictors of both the person parameter and the latent class membership (OM); and 10) the two-class MRM with *reading enjoyment time* as the predictor of the person parameter and *ESCS* as the predictor of the latent class membership (MISM). The preliminary study has found non-convergence issues when fitting the data using the MRM with covariates and more than two latent classes. Considering sample size, percentage of missing data, and model complexity, the mixture Rasch models with more than two latent classes are not further explored in the real data applications.

Chapter 4 Results

Covariate inclusion in IRT models is not a new topic, yet the exploration of different approaches to including covariates in mixture IRT models still needs further research. The present study focuses on an investigation of different approaches to adding dichotomous and continuous covariates into the mixture Rasch model, with both complete and incomplete response data. The simulation study described in Chapter 3 explored the impact of six manipulated factors on the model performance as indexed by three categories of evaluation criteria, including latent group classification, parameter recovery and overall model fit. Two empirical data analyses were conducted to illustrate the impact of different specifications of covariate effects for an MRM in real applications.

4.1 Results of the Simulation Study

For a clear presentation of the results, abbreviations of manipulated factors and model names were used in the tables and figures presented in the following sections. Table 4.1 listed all the abbreviations and their corresponding explanations.

Table 4.1. *Variable and model name abbreviations*.

Abbreviations	Explanation
Manipulated factors:	
Model	Model type
Prop	Mixing proportion
OR	Strength of the relation between D_j and π_{jg}
Corr	Strength of the relation between C_j and θ_{jg}
DIF	Average DIF(i.e., the mean of $\|b_{i1}$-$b_{i2}\|$)
Data	Response data completeness
Model names (see Table 3.1 for details):	
TM	Data-generating model
UMN-N	The MRM without covariates
UMN-C	The MRM with the continuous covariate only
UMN-D	The MRM with the dichotomous covariate only
OM	The over specified model
MISM	The MRM with mismatching covariates

For all the outcome measures, descriptive statistics were provided in the following sections. In order to identify statistically significant effects of the manipulated factors on the model performance evaluation criteria (except overall model fit), several repeated measures analyses of variance (ANOVA) were performed in SPSS (version 19.0). The manipulated factors, including mixing proportion, strength of the relation between D_j and π_{jg}, strength of the relation between C_j and θ_{jg}, DIF and response data completeness, were used as between-replication variables. Model was used as a within-replication variable because each generated data set was fitted by six models repeatedly. The sphericity assumption was checked for repeated measures ANOVA, and the Huynh-Feldt correction was used to adjust the degrees of freedom if necessary. The correction does not change the value of the *F*-statistic, but changes the degrees of freedom of the *F*-distribution. The smaller degrees of freedom in both the numerator and the denominator result in larger critical values, so that the inflation of Type I error due to the violation of the sphericity assumption can be adjusted.

Additionally, although repeated measures ANOVA assumes that the dependent variable is normally distributed for each level of the with-replication variable, it is also known that ANOVA is robust to moderate violation of the normality assumption (Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992; Lix, Keselman, & Keselman, 1996). Considering that the sample size with respect to the repeated variable (i.e., model) was relatively large in the current simulation, the normality assumption was not a big concern for the present study. However, arcsine transformation was still implemented in the preliminary analyses of the simulation results and the repeated ANOVA findings were compared with those obtained from the original data, and no much difference was found in terms of the statistically significant effects. In order to make the interpretation more meaningful, the results based on the original data were reported in later sections.

For all the main effects and interactions, effect sizes as indicated by Cohen's *f* were calculated according to Equation 4.1:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}}\,, \tag{4.1}$$

where η^2 describes the ratio of variance explained in the dependent variable by a manipulated factor (i.e., η^2=SS$_{\text{factor}}$/SS$_{\text{total}}$).

Only those statistically significant effects with at least an *f* value of .1 (i.e., small effect size) were reported. The effect size cutting values are negligible (f<.1), small (.1≤f<.25), moderate (.25≤f<.4), and large (f≥.4) in the ANOVA (Cohen, 1988). Nonsignificant or significant results with negligible effect sizes were not presented in the ANOVA tables in the present study. The effect sizes of all four-way and five-way interactions were negligible in the current study, thus none of them were presented. Effect sizes were reported in the following sections both in the table and inside the parenthesis when interpreting the effects of studied factors.

4.1.1 Latent group classification.

The accuracy of latent group classification was evaluated using the correct classification rate, which is the proportion of subjects in the sample that are assigned to their true latent class based on their estimated latent class membership.

Table 4.2 presented the descriptive statistics of correct classification rate by six manipulated factors. Overall, the average correct classification rate was very high and usually above .900, with the exception of the booklet design condition. This may be due to the large DIF size and separations between latent classes simulated in the present study. As expected, the true model with both dichotomous and continuous covariates correctly specified resulted in in the most accurate latent class assignment, although the difference between the true model and the overspecified model was almost negligible. Also, when either of the covariates was correctly specified in the model (i.e., UNM-C and UNM-D), the accuracy of latent group classification was better than the conditions in which no covariates were included (i.e., UNM-N). The underspecified model with only the dichotomous covariate included resulted in slightly higher correct classification rate than that with only the continuous covariate included. The MRM with mismatching covariates resulted in the worst correct classification rate, and it was even worse than not including any covariates.

Table 4.2. *The descriptive statistics of correct classification rate by manipulated factors*.

Factors	Levels	M	SD
Model	TM	0.936	0.031
	UMN-N	0.923	0.037
	UMN-C	0.929	0.033
	UMN-D	0.931	0.034
	OM	0.935	0.031
	MISM	0.907	0.068
Prop	.5/.5	0.932	0.033
	.3/.7	0.921	0.049
OR	1	0.923	0.043
	10	0.930	0.041
Corr	.2;.2	0.926	0.037
	.8;.8	0.928	0.047
DIF	1	0.912	0.049
	1.5	0.941	0.027
Data	Complete	0.960	0.014
	Booklet Design	0.893	0.053
	Omitted Response	0.927	0.011

Notes: TM: the data-generating model; UNM-N: the MRM without covariates; UNM-C: the MRM with the continuous covariate only; UNM-D: the MRM with the dichotomous covariate only; OM: the over specified model; MISM: the MRM with mismatching covariates.

In addition, equal mixing proportion resulted in better latent group classification, and larger DIF size also helped assign subjects to their correct latent class. Regarding odds ratio and correlation which indexed the strength of relations between the covariates and the model parameters, the stronger the relations were, the higher the correct classification rate. The average correct latent group classification rates for all the simulation cells were fully presented in Table 3 in Appendix A.

Table 4.3. *ANOVA results of manipulated factors on the correct classification rate.*

Source	F value	*p* value	η^2	Cohen's *f*
Within(Huynh-Feldt Correction):				
Model	1584.284	<.001	0.054	0.239
Model* Data Completeness	579.142	<.001	0.039	0.203
Model*OR	302.962	<.001	0.010	0.102
Model*Corr	656.639	<.001	0.022	0.151
Model*DIF	418.804	<.001	0.014	0.120
Data*Model*Corr	314.765	<.001	0.021	0.148
Data*Model*DIF	287.232	<.001	0.020	0.141
Between:				
Data Completeness	8019.088	<.001	0.421	0.853
Prop	639.619	<.001	0.016	0.131
DIF	4440.138	<.001	0.117	0.363
Data Completeness*Prop	582.768	<.001	0.031	0.178
Data Completeness*DIF	1730.391	<.001	0.091	0.316

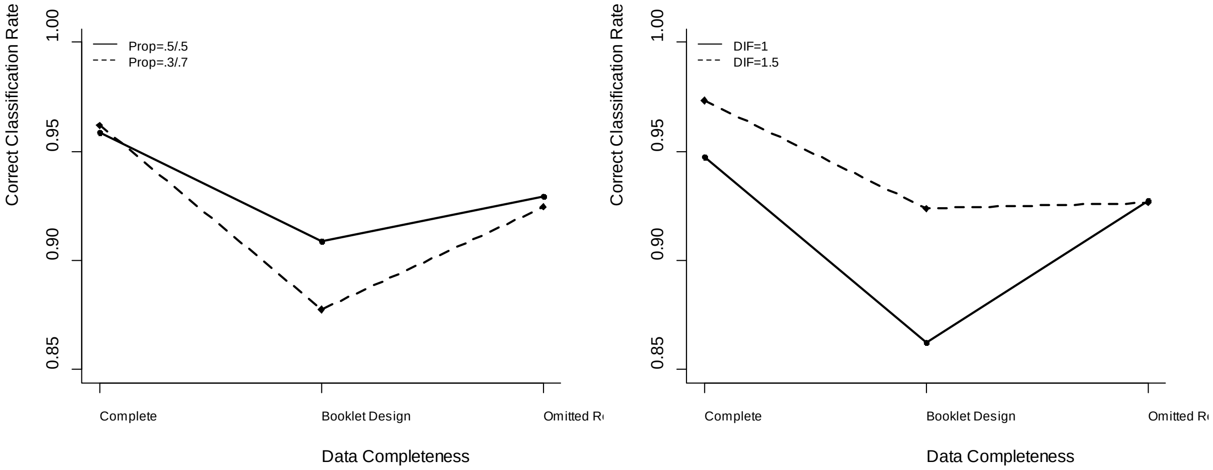

Figure 4.1a. Statistically significant two-way interactions among between-replication variables on the correct classification rate.

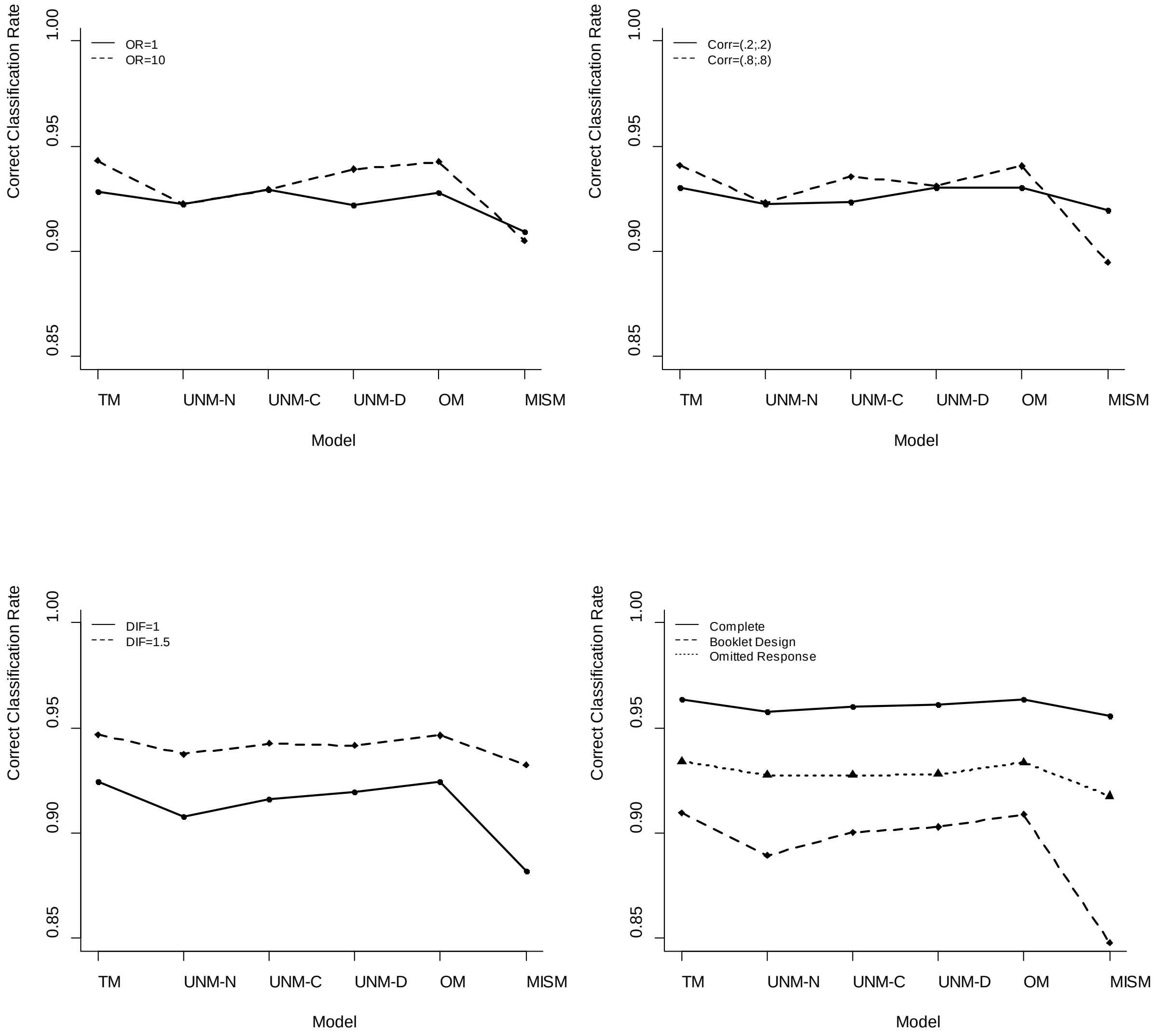

Figure 4.1b. Statistically significant two-way interactions between Model and other between-replication variables on the correct classification rate.

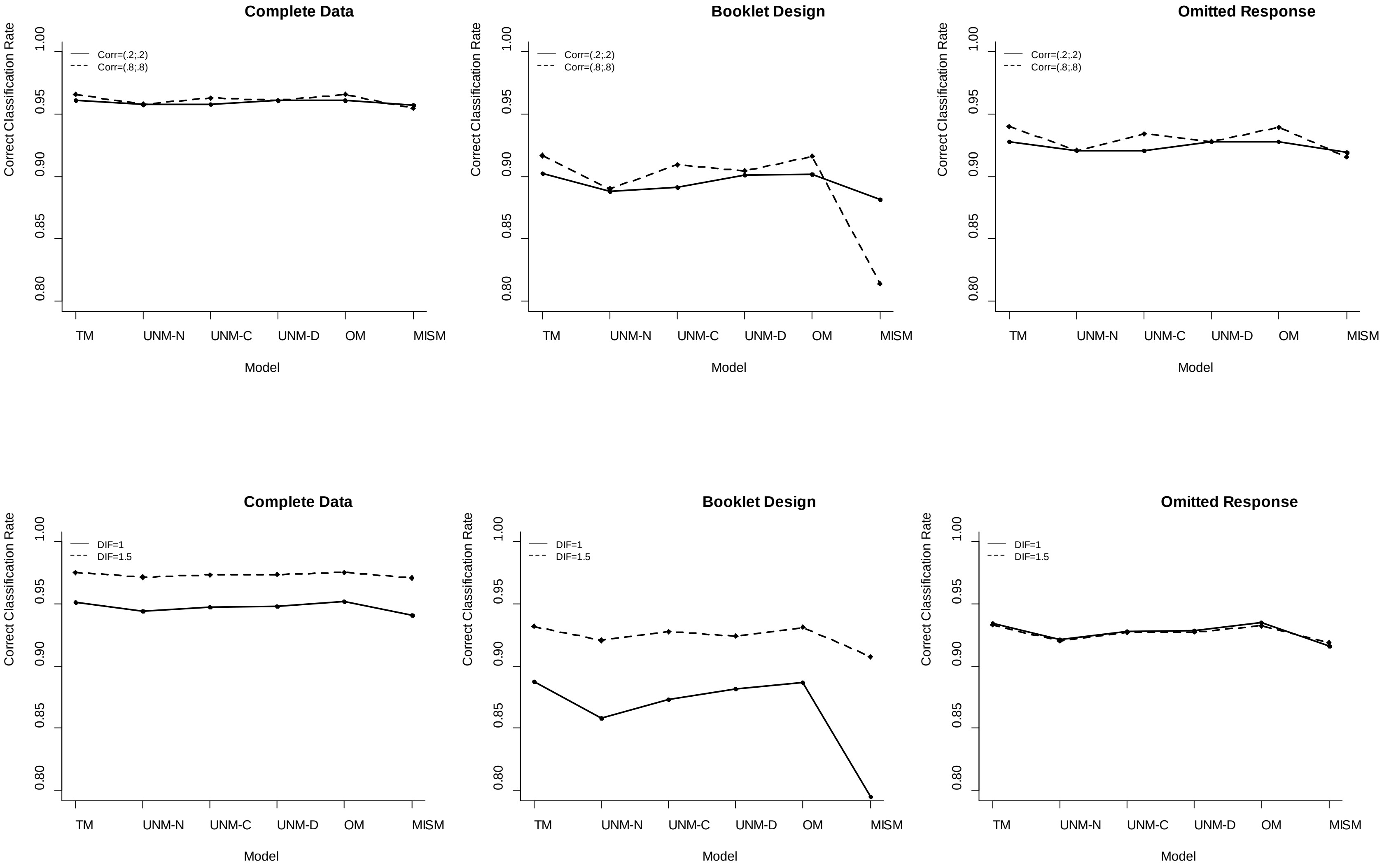

Figure 4.1c. Statistically significant three-way interactions among between-replication variables on the correct classification rate.

The ANOVA results presented in Table 4.3 indicated that estimation model, data completeness, mixing proportion, and DIF size had significant effects on the correct classification rate. The effect sizes were large for data completeness (*f*=0.853), moderate for DIF (*f*=0.363) and small for model (*f*=0.239) and mixing proportion (*f*=0.131). The complete data scenario resulted in the most accurate latent class assignment, and booklet design led to the worst latent group classification. Besides, larger DIF and equal mixing proportion tended to result in higher correct classification rate. Post-hoc pairwise comparison showed that all pairwise differences were statistically significant, thus they were not reported in details here.

In addition to the main effects, the interaction terms of model by data completeness (*f*=0.203), odds ratio (*f*=0.102), correlation (*f*=0.151) and DIF (*f*=0.120), and the interactions of data completeness by mixing proportion (*f*=0.178) and DIF (*f*=0.316) were also found to be significantly related to the accuracy of latent class assignment. Further, three-way interactions among data completeness, model and correlation (*f*=0.148), and among data completeness, model and DIF (*f*=0.141) were also found to be statistically significant. The two-way interactions were depicted in Figure 4.1a and 4.1b, and the three-way interactions were presented in Figure 4.1c. As shown in Figure 4.1a, the effect of data completeness tended to be stronger when the mixing proportion was unequal, with the booklet design resulted in remarkably worse latent class assignment. Similarly, the booklet design also resulted in dramatically low correct classification rate in the case of smaller DIF. However, the latent class assignment in the omitted response conditions seemed largely unaffected by DIF or mixing proportion.

In Figure 4.1b, it was shown that the correct classification rate was much higher for the true model, the MRM with only the dichotomous covariate and the overspecified model when the relation between the dichotomous covariate and the latent class membership was stronger;

and similarly, the correct classification rate was higher for the true model, the MRM with only the continuous covariate and the overspecified model when the relation between the continuous covariate and the person parameter was stronger. The latent class assignment of the MRM without covariates remained unaffected by odds ratio or correlation, whereas the classification for the MRM with mismatching covariates was worse when the relations between the covariates and the model parameters were stronger. Moreover, the two plots at the bottom of Figure 4.1b indicated that the effect of model on the correct classification rate was more pronounced when DIF was smaller or the booklet design was present, with the MRM without covariates and the MRM with mismatching covariates performing even worse in terms of latent class assignment in these two situations.

With regard to the three-way interactions, the plots at the upper part of the panel in Figure 4.3c showed that there was merely an interaction between model and correlation in the complete data scenario, and even the main effects of correlation and model on the latent class assignment were negligible in this situation. However, in the omitted response situation, the correct classification rate was higher for the true model, the MRM with only the continuous covariate and the overspecified model when the correlation was stronger, whereas the latent class assignment of the MRM without covariates, the MRM with only the dichotomous covariate and the MRM with mismatching covariates remained unaffected by the correlation. In the booklet design scenario, the interaction between model and correlation was even more pronounced, with the correct classification rate much higher for the true model, the MRM with only the continuous covariate and the overspecified model and dramatically lower for the MRM with mismatching covariates when the correlation was stronger, as compared with the weaker correlation situations. Finally, the bottom part of the panel indicated that the interaction between model and DIF was almost negligible when omitted responses were present. Moreover, the main effects of DIF and model were also minimal in this situation. In the complete data scenario, the effect of DIF was consistent regardless of model types. The interaction between model and DIF was not observed in this situation. However, in the booklet design, the interaction was much more pronounced, with the MRM without covariates and the MRM with mismatching covariates performing much worse in the latent class assignment than the other models when DIF size was smaller. The importance of covariate inclusion on latent group classification and its relations with other factors are discussed in more details in Chapter 5.

In the following sections, the ANOVA results of manipulated factors on the evaluation criteria of model parameter recovery, including item parameter recovery, person parameter recovery and regression parameter recovery, are respectively presented.

4.1.2 Model parameter recovery.

In the present simulation, the recovery of item, person and regression parameters was evaluated separately in order to investigate the effects of manipulated factors on different aspects of model parameter recovery. Item parameters were examined in terms of SE, RMSE, and the proportion of replications for which the 95% confidence interval around the parameter estimates covered the true value (i.e., 95% coverage). As the item parameters were used for scale identification purpose, on average there was no bias involved in the item parameter estimates. For person parameters, they were similarly evaluated with respect to bias, SE, RMSE and 95% coverage. As presented in Equation 3.8 to 3.10, bias quantifies the systematic error in the estimation, and SE represents the random error, whereas RMSE combines the information of bias and SE to reflect overall item or person parameter recovery. For the 95% coverage, it is expected that the coverage rate approximates the nominal level (i.e., 0.950) if the model is correctly specified. The discrepancy between the coverage rate and the nominal coverage probability may indicate model misspecification or deviations from normality for the estimator distribution.

Additionally, the regression parameters which linked the covariates with the model parameters were also evaluated in terms of bias, SE and RMSE. It was of interest that whether the regression parameters were underestimated or overestimated in the one-step estimation of the conditional model used in the present study.

Item parameter recovery. Table 4.4 presented the summary statistics of item parameter recovery evaluation criteria by six manipulated factors, and the average item parameter SE, RMSE and 95% coverage were completely displayed in Table 4a through 4c in Appendix A. Across all the other factors, the true model resulted in the smallest SE and RMSE, and the highest 95% coverage rate, followed by the overspecified model with negligible difference. Also, when either of the covariates was correctly specified in the model (i.e., UNM-C or UNM-D), the item parameter recovery was better than the MRM with no covariates included. The MRM with only the continuous covariate resulted in slightly better recovery than the MRM with only the dichotomous covariate. Similar to the performance of the MRM with mismatching covariates on the latent class assignment, this model also resulted in the worst item parameter recovery, and it was even worse than the MRM without covariates.

Moreover, equal mixing proportion or stronger relations between the covariates and the model parameters tended to result in better item parameter recovery. It was within expectation because a sample size of 1,000 per class was more adequate than unequal sample sizes for the item parameter estimation of the mixture Rasch model. Also, smaller DIF size helped recover item parameters. As expected, in the booklet design condition which was accompanied by the largest amount of missing data, the item parameter recovery was the worst, whereas the complete data scenario led to the best recovery of item parameters.

Table 4.4. *The descriptive statistics of item parameter by manipulated factors*.

Factors	Levels	Item Parameter SE		Item Parameter RMSE		95% Coverage	
		M	SD	M	SD	M	SD
Model	TM	0.089	0.024	0.136	0.103	0.868	0.103
	UMN-N	0.091	0.024	0.144	0.115	0.849	0.116
	UMN-C	0.089	0.023	0.140	0.107	0.860	0.109
	UMN-D	0.092	0.029	0.142	0.110	0.859	0.112
	OM	0.089	0.024	0.136	0.103	0.867	0.104
	MISM	0.094	0.029	0.158	0.132	0.823	0.145
Prop	.5/.5	0.085	0.020	0.100	0.032	0.892	0.055
	.3/.7	0.097	0.029	0.185	0.142	0.816	0.145
OR	1	0.091	0.026	0.144	0.112	0.850	0.118
	10	0.090	0.025	0.141	0.110	0.859	0.114
Corr	.2;.2	0.091	0.025	0.143	0.113	0.852	0.115
	.8;.8	0.091	0.026	0.142	0.110	0.856	0.117
DIF	1	0.083	0.017	0.121	0.074	0.864	0.088
	1.5	0.098	0.030	0.164	0.136	0.844	0.138
Data	Complete	0.072	0.004	0.073	0.005	0.953	0.005
	Booklet Design	0.119	0.025	0.250	0.137	0.749	0.138
	Omitted Response	0.082	0.007	0.104	0.020	0.861	0.022

Notes: TM: the data-generating model; UNM-N: the MRM without covariates; UNM-C: the MRM with the continuous covariate only; UNM-D: the MRM with the dichotomous covariate only; OM: the over specified model; MISM: the MRM with mismatching covariates.

Table 4.5a. *Statistically significant ANOVA results of manipulated factors on the SE of item parameters*.

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,2832}$=407.610	<.001	0.189	0.483
Prop	$F_{1, 2832}$=69.398	<.001	0.016	0.128
DIF	$F_{1, 2832}$=118.199	<.001	0.034	0.187
Data*DIF	$F_{2, 2832}$=25.578	<.001	0.015	0.122

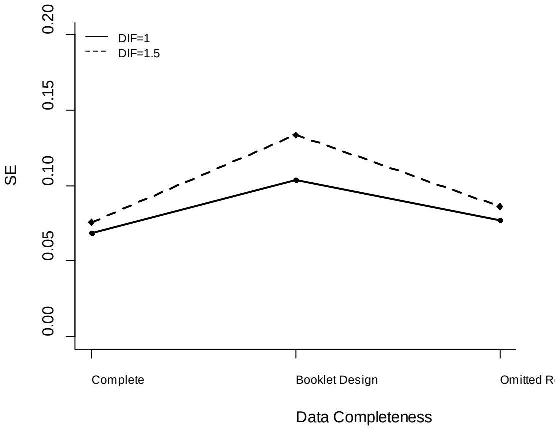

Figure 4.2a. Statistically significant two-way interactions among between-replication variables on the SE of item parameters.

Table 4.5b. *Statistically significant ANOVA results of manipulated factors on the RMSE of item parameters*.

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,2832}$=173.949	<.001	0.097	0.327
Prop	$F_{1, 2832}$=103.804	<.001	0.029	0.172
Data*DIF	$F_{2, 2832}$=75.260	<.001	0.042	0.209

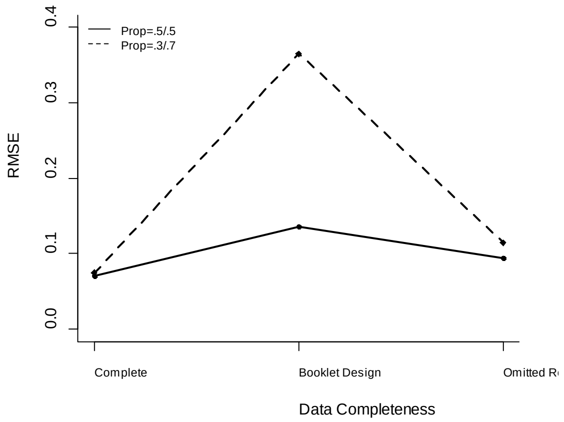

Figure 4.2b. Statistically significant two-way interactions among between-replication variables on the RMSE of item parameters.

Table 4.5c. *Statistically significant ANOVA results of manipulated factors on the 95% coverage of item parameters.*

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,2832}$=292.612	<.001	0.140	0.403
Prop	$F_{1, 2832}$=123.318	<.001	0.029	0.174
Data*Prop	$F_{1, 2832}$=100.812	<.001	0.048	0.225
Data*DIF*Prop	$F_{2, 2832}$=31.418	<.001	0.015	0.123

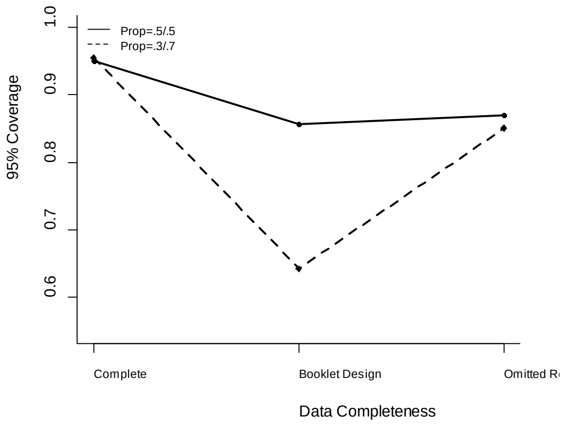

Figure 4.2c. Statistically significant two-way interactions among between-replication variables on the 95% coverage of item parameters.

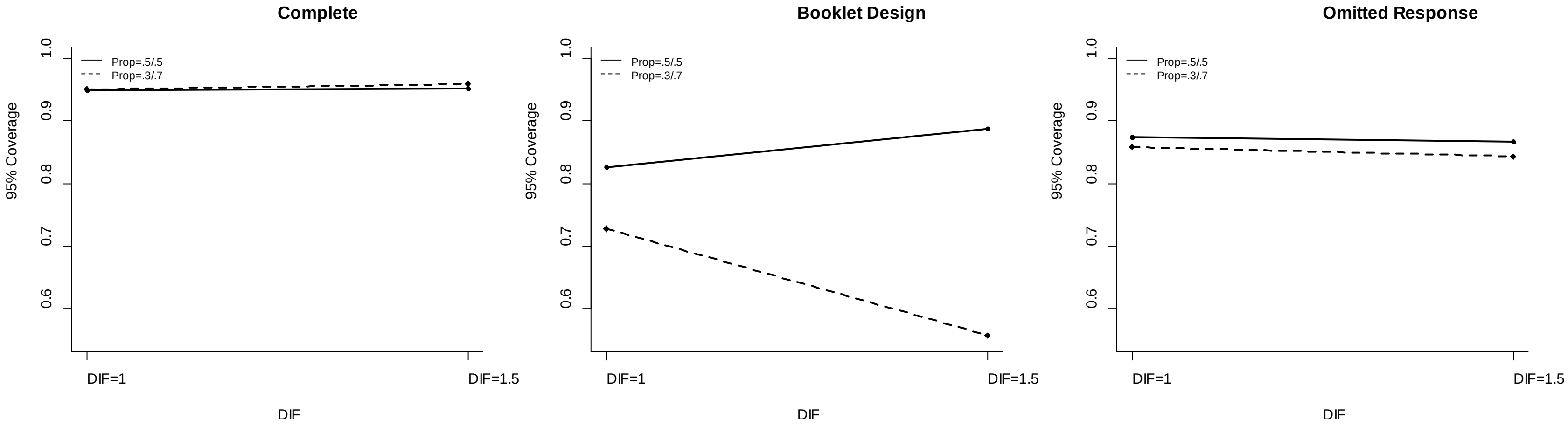

Figure 4.2d. Statistically significant three-way interactions among between-replication variables on the 95% coverage of item parameters.

The ANOVA results presented in Table 4.5a to 4.5c indicated that data completeness, mixing proportion, and DIF had statistically significant impacts on the SE of item parameters. Among them, data completeness had a large effect size (*f*=0.483), and mixing proportion (*f*=0.128) and DIF (*f*=0.187) respectively had a small effect size. In addition, data completeness interacted significantly with DIF with a small effect size (*f*=0.122). As shown in Figure 4.2a, the effect of data completeness on the SE of item parameters was more pronounced when DIF was larger, with the booklet design scenario resulting in much larger item parameter SE than the complete data and omitted response conditions.

Regarding the RMSE of item parameters, data completeness and mixing proportion had statistically significant effects, respectively with a moderate (*f*=0.327) and a small effect size (*f*=0.172). The interaction term of data completeness by mixing proportion was also statistically significant with a small effect size (*f*=0.172). Figure 4.2b indicated that the effect of data completeness was remarkably large in unequal mixing proportion conditions, with the booklet design resulting in very large RMSE of item parameters.

Similar to the results of RMSE, data completeness, mixing proportion and their interactions also had statistically significant effects on the 95% coverage of item parameters. The interaction among data completeness, mixing proportion and DIF was also statistically significant. The effect size was large for data completeness (*f*=0.403), and small for mixing proportion (*f*=0.174), the two-way interaction (*f*=0.225) and the three-way interaction (*f*=0.123). Figure 4.2c indicated that the effect of data completeness on the 95% coverage of item parameters was large in unequal mixing proportion conditions, with the booklet design leading to a very low coverage rate of around 0.650. However, the 95% coverage in the complete data and omitted response conditions remained largely unaffected by mixing proportion. Further, as shown in Figure 4.2d, there seemed not to be any effects of DIF, mixing proportion or their interaction in the complete data and omitted response scenarios. Nevertheless, in the booklet design condition, it was interesting to find that larger DIF resulted in slightly higher 95% coverage rate for item parameters when the mixing proportion was equal; whereas larger DIF resulted in dramatically lower 95% coverage rate when the mixing proportion was unequal.

Although the descriptive statistics showed some differences in the item parameter evaluation criteria with respect to model, in the repeated measures ANOVA, the effect size of model was not large enough to claim a practical significance.

Person parameter recovery. Table 4.6 presented the descriptive statistics of person parameter recovery evaluation criteria by six manipulated factors. The average person parameter bias, SE, RMSE and 95% coverage rate were completely displayed in Table 5a to 5d in Appendix A. Overall, there tended to be a positive bias in the person parameter estimates, indicating an overestimation of the person parameters. However, the marginal bias for the booklet design condition across other manipulated factors was negative, suggesting an underestimation in this condition. For the SE of person parameters, there were negligible differences among the true model, the MRM with only the continuous covariate and the overspecified model, and they resulted in smaller SE than the other three model. It indicated that the inclusion of the continuous covariate, rather than the dichotomous covariate, may potentially lead to a reduction in the SE of person parameter estimates. With regard to the RMSE which indicated the overall recovery of person parameters, it was found that the true model and the overspecified model resulted in the best person parameter recovery, followed by the MRM with only the continuous covariate with negligible difference. The MRM with mismatching covariates again performed worse than the MRM without covariates in terms of person parameter recovery.

For other manipulated factors, unequal mixing proportion, stronger relations between covariates and model parameters or smaller DIF resulted in better recovery of person parameters. As with other evaluation criteria, the person parameter recovery was the best for the complete data scenario, followed by the omitted response condition, and the booklet design was the worst. The 95% coverage seemed not to differ much according to different levels of manipulated factors. Additionally, it was also interesting to find that the omitted response condition led to the largest magnitude of bias and the lowest 95% coverage rate as compares with the other two data completeness conditions, suggesting that the conditional missing data mechanism may have an impact on the recovery of person parameters but not item parameters.

Table 4.6. *The descriptive statistics of person parameter by manipulated factors*.

Factors	Levels	Person Parameter Bias		Person Parameter SE		Person Parameter RMSE		95% Coverage	
		M	SD	M	SD	M	SD	M	SD
Model	TM	0.025	0.061	0.144	0.040	0.184	0.056	0.948	0.011
	UMN-N	0.022	0.075	0.162	0.042	0.202	0.065	0.946	0.012
	UMN-C	0.031	0.065	0.143	0.042	0.185	0.059	0.947	0.012
	UMN-D	0.022	0.071	0.163	0.047	0.198	0.066	0.946	0.011
	OM	0.024	0.063	0.145	0.041	0.184	0.058	0.948	0.011
	MISM	0.027	0.080	0.163	0.043	0.207	0.070	0.942	0.014
Prop	.5/.5	0.043	0.035	0.179	0.029	0.217	0.048	0.948	0.010
	.3/.7	0.007	0.087	0.128	0.040	0.170	0.067	0.944	0.013
OR	1	0.026	0.071	0.154	0.044	0.195	0.064	0.946	0.012
	10	0.025	0.067	0.153	0.042	0.191	0.062	0.946	0.011
Corr	.2;.2	0.022	0.072	0.161	0.042	0.199	0.065	0.945	0.012
	.8;.8	0.028	0.065	0.146	0.041	0.188	0.060	0.946	0.012
DIF	1	0.029	0.023	0.135	0.032	0.158	0.035	0.948	0.008
	1.5	0.022	0.095	0.172	0.046	0.229	0.064	0.944	0.014
Data	Complete	0.010	0.003	0.134	0.033	0.146	0.035	0.955	0.001
	Booklet Design	-0.016	0.091	0.181	0.046	0.234	0.067	0.950	0.007
	Omitted Response	0.082	0.029	0.145	0.034	0.200	0.046	0.933	0.009

Notes: TM: the data-generating model; UNM-N: the MRM without covariates; UNM-C: the MRM with the continuous covariate only; UNM-D: the MRM with the dichotomous covariate only; OM: the over specified model; MISM: the MRM with mismatching covariates.

Table 4.7a. *Statistically significant ANOVA results of manipulated factors on the bias of person parameters*.

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,95952}$=2833.373	<.001	0.048	0.223
Data*Prop	$F_{2, 95952}$=1422.504	<.001	0.024	0.156
Data*DIF	$F_{2, 95952}$=1176.261	<.001	0.020	0.142
Data*Prop*DIF	$F_{2, 95952}$=977.362	<.001	0.016	0.129

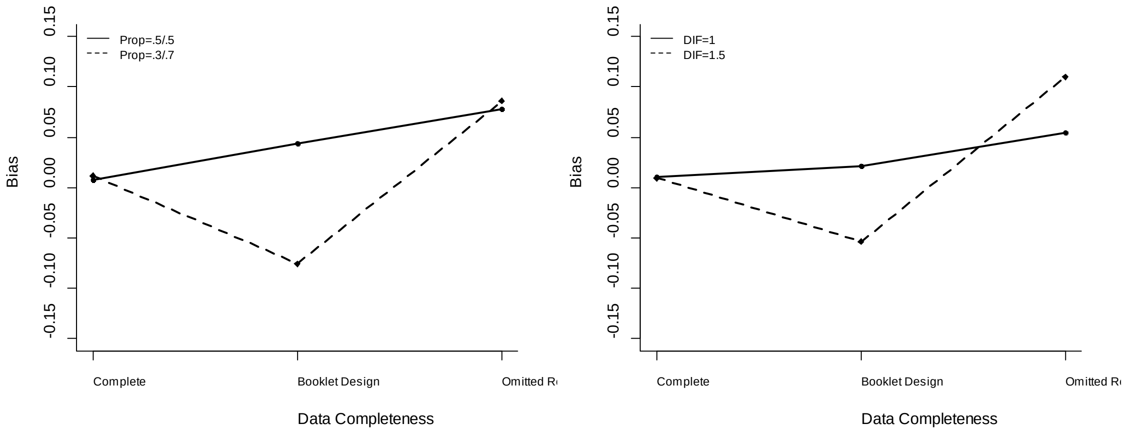

Figure 4.3a. Statistically significant two-way interactions among between-replication variables on the bias of person parameters.

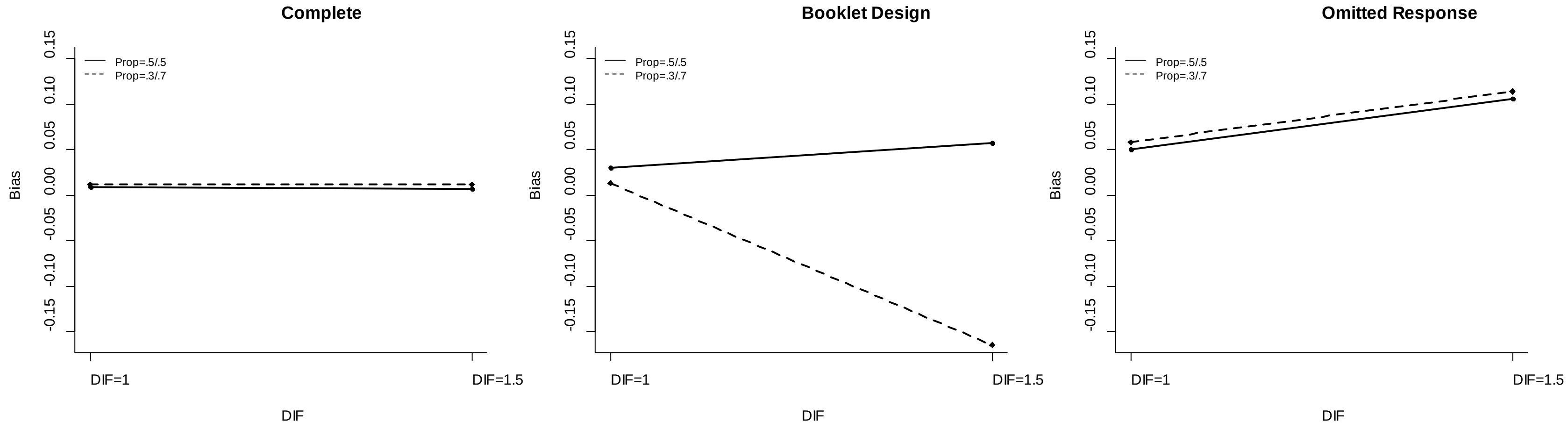

Figure 4.3b. Statistically significant three-way interactions among between-replication variables on the bias of person parameters.

Table 4.7b. *Statistically significant ANOVA results of manipulated factors on the SE of person parameters*.

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,95952}$=707.581	<.001	0.014	0.118
Prop	$F_{1,95952}$=2181.871	<.001	0.021	0.147
DIF	$F_{1,95952}$=1151.505	<.001	0.011	0.106

Table 4.7c. *Statistically significant ANOVA results of manipulated factors on the RMSE of person parameters*.

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,95952}$=1367.923	<.001	0.026	0.163
Prop	$F_{1,95952}$=1133.556	<.001	0.011	0.104
DIF	$F_{1,95952}$=264.484	<.001	0.025	0.159

Table 4.7d. *Statistically significant ANOVA results of manipulated factors on the 95% Coverage of person parameters*.

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,95952}$=873.849	<.001	0.015	0.122

The ANOVA results presented in Table 4.7a to 4.7d showed that data completeness had a statistically significant effect on the bias of person parameters with a moderate effect size (*f*=0.223). In addition, data completeness interacted significantly with mixing proportion (*f*=0.156) and DIF (*f*=0.142) with respect to the bias of person parameters. The three-way interaction term of data by mixing proportion and DIF were also found to be statistically significant with a small effect size (*f*=0.129). The two-way interactions were displayed in Figure 4.3a. With equal mixing proportion, the effect of data completeness was small, and the booklet design resulted in slight positive bias in the person parameter estimates. However, when the mixing proportion was unequal, there tended to be a larger negative bias in the person parameter estimates. For the complete data and omitted response conditions, the bias of person parameters was largely unaffected by mixing proportion. Additionally, when DIF size was smaller, there was consistently a small positive bias in the person parameter estimates regardless of data completeness; whereas when DIF was larger, the booklet design led to a negative bias in the parameter estimates, and the omitted response condition resulted in a larger positive bias, but the bias in the complete data scenario remained unchanged. Further, the three-way interaction plot in Figure 4.3b indicated that the interaction between mixing proportion and DIF was stronger in the booklet design condition, with unequal mixing proportion resulting in larger negative bias in the person parameter estimates when DIF was larger. However, in the other DIF, mixing proportion and data completeness level combinations, the bias of person parameters was positive and relatively small.

Regarding the SE and RMSE of person parameters, data completeness (*f*=0.118; *f*=0.163), mixing proportion (*f*=0.147; *f*=0.104) and DIF (*f*=0.106; *f*=0.159) were found to have significant effects on these two measures with small effect sizes. No significant interaction terms were observed. For the 95% coverage, only data completeness was found to be statistically significant with a small effect size (*f*=0.122).

Similar to the results of item parameter recovery, the effect size of model on person parameter evaluation criteria did not exceed 0.100, and thus was not practically significant. However, some meaningful differences were still observed in the descriptive statistics with regard to covariate inclusion as presented in Table 4.6.

Regression parameter recovery. The evaluation of regression parameter recovery included two parts: 1) the linear regression parameters which linked the continuous covariate with the person parameter, and 2) the logistic regression parameters which linked the dichotomous covariate with the latent class membership. For part 1, the true model, the MRM with the continuous covariate only and the overspecified model were under investigation; and for part 2, the true model, the MRM with the dichotomous covariate only and the overspecified model were evaluated. The MRM without covariates and the MRM with mismatching covariates were not included in the evaluation because no regression parameters were involved in the former one and the regression parameters corresponded to wrong covariates in the latter one.

Table 4.8 presented the descriptive statistics of the linear regression parameter measures. Overall, there was a positive bias in the intercept and slope parameters of the linear regression, suggesting a tendency of overestimation of these parameters. The only exception was the omitted response condition in which the marginal bias across the other manipulated factors was negative. With regard to model, the true model and the MRM with only the continuous covariate resulted in the smallest bias in the intercept, whereas the overspecified model had the smallest bias in the slope parameter estimates. In addition, unequal mixing proportion was associated with a smaller bias in both parameters. A smaller odds ratio or a larger DIF size would also result in smaller bias in the intercept and slope parameter estimates. Regarding the overall recovery of the linear regression parameters as indicated by the RMSE, the true model recovered the intercept parameter the best and the overspecified model recovered the slope parameter the best. Also, better recovery of linear regression parameters may be obtained when the mixing proportion was unequal, the odds ratio was small or the correlation was large. Large DIF size tended to result in better recovery of the intercept parameter but worse recovery of the slope parameter. Finally, the booklet design condition led to the worst linear regression parameter recovery as expected.

Table 4.8. *The descriptive statistics of linear regression parameter (linking C_j with θ_{jg}) by manipulated factors*.

Factors	Levels	$\alpha_{0.}$ (intercept)						$\alpha_{1.}$ (slope)					
		Bias		SE		RMSE		Bias		SE		RMSE	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Model	TM	0.030	0.072	0.028	0.008	0.073	0.051	0.014	0.027	0.021	0.005	0.035	0.013
	UMN-C	0.030	0.077	0.028	0.008	0.077	0.056	0.013	0.028	0.021	0.005	0.037	0.014
	OM	0.060	0.073	0.043	0.011	0.098	0.055	0.010	0.024	0.021	0.005	0.033	0.011
Prop	.5/.5	0.045	0.085	0.036	0.012	0.096	0.058	0.025	0.014	0.022	0.005	0.036	0.012
	.3/.7	0.035	0.063	0.030	0.010	0.069	0.049	0.000	0.030	0.021	0.005	0.034	0.014
OR	1	0.004	0.044	0.033	0.012	0.058	0.024	0.003	0.021	0.017	0.002	0.027	0.010
	10	0.076	0.083	0.033	0.011	0.108	0.065	0.022	0.028	0.025	0.004	0.043	0.010
Corr	.2;.2	0.046	0.081	0.033	0.011	0.086	0.061	0.012	0.027	0.022	0.005	0.036	0.013
	.8;.8	0.034	0.068	0.034	0.013	0.080	0.049	0.013	0.025	0.020	0.004	0.034	0.012
DIF	1	0.059	0.078	0.037	0.011	0.087	0.062	0.016	0.022	0.021	0.004	0.034	0.010
	1.5	0.021	0.067	0.030	0.011	0.079	0.079	0.009	0.030	0.021	0.006	0.036	0.015
Data	Complete	0.015	0.028	0.027	0.008	0.046	0.022	0.023	0.014	0.018	0.003	0.031	0.013
	Booklet Design	0.029	0.108	0.042	0.012	0.115	0.063	0.026	0.013	0.024	0.005	0.037	0.013
	Omitted Response	0.076	0.050	0.031	0.009	0.087	0.047	-0.012	0.028	0.021	0.004	0.037	0.011

Notes: TM: the data-generating model; UNM-C: the MRM with the continuous covariate only; OM: the over specified model.

Table 4.9a. *Statistically significant ANOVA results of manipulated factors on the bias of* $\alpha_{0\cdot}$.

Source	F value	*p* value	η^2	Cohen's *f*
Within(Huynh-Feldt Correction):				
Model	$F_{1.2,2750.9}$=668.607	<.001	0.019	0.138
Between:				
Data	$F_{2,2352}$=171.804	<.001	0.062	0.257
Prop	$F_{1,2352}$=642.007	<.001	0.116	0.362
DIF	$F_{1,2352}$=176.427	<.001	0.032	0.181
Data*Prop	$F_{2,2352}$=67.887	<.001	0.025	0.159
Data*DIF	$F_{2,2352}$=349.649	<.001	0.126	0.380
Prop*DIF	$F_{1,2352}$=74.724	<.001	0.013	0.117
Data*Prop*DIF	$F_{2,2352}$=160.567	<.001	0.058	0.248

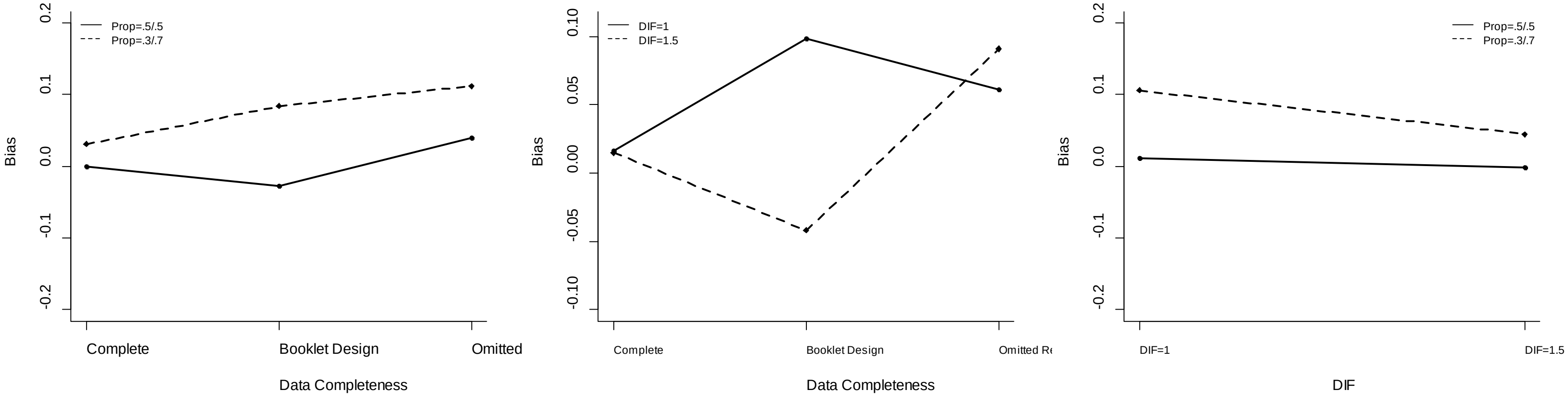

Figure 4.4a. Statistically significant two-way interactions among between-replication variables on the bias of $\alpha_{0\cdot}$.

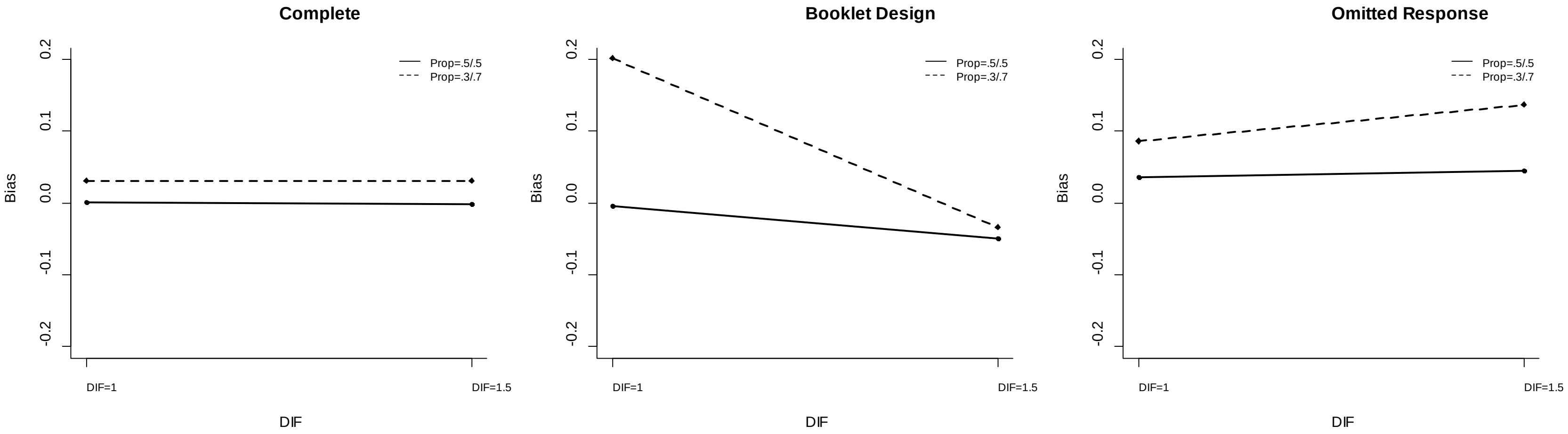

Figure 4.4b. Statistically significant three-way interactions among between-replication variables on the bias of $\alpha_{0\cdot}$.

Table 4.9b. *Statistically significant ANOVA results of manipulated factors on the SE of* $\alpha_{0\cdot}$.

Source	F value	*p* value	η^2	Cohen's *f*
Within(Sphericity Assumed):				
Model	$F_{2,4}$=1659.536	<.001	0.368	0.764
Between:				
Data	$F_{2,2}$=934.644	<.01	0.316	0.679
Corr	$F_{1,2}$=376.889	<.01	0.053	0.236
DIF	$F_{1,2}$=582.156	<.01	0.105	0.343

Table 4.9c. *Statistically significant ANOVA results of manipulated factors on the RMSE of* $\alpha_{0\cdot}$.

Source	F value	*p* value	η^2	Cohen's *f*
Within(Sphericity Assumed):				
Model	$F_{2,4}$=275.516	<.001	0.040	0.203
Between:				
Data	$F_{2,2}$=706.655	<.01	0.270	0.608
Corr	$F_{1,2}$=322.303	<.01	0.060	0.254
Prop	$F_{1,2}$=1099.124	<.01	0.209	0.514
Data*Prop	$F_{2,2}$=391.266	<.01	0.149	0.418
Data*DIF	$F_{2,2}$=246.889	<.01	0.095	0.325
Data*Prop*DIF	$F_{2,2}$=143.679	<.01	0.056	0.243

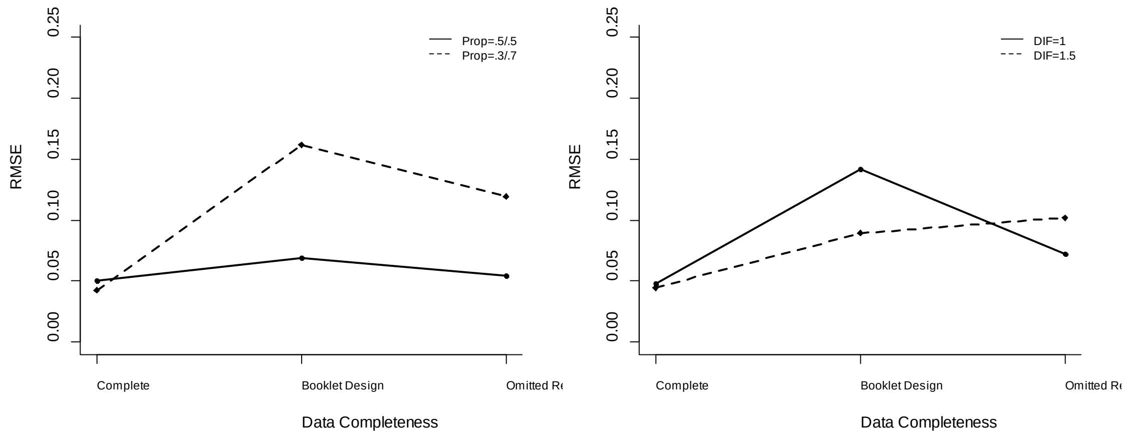

Figure 4.4c. Statistically significant two-way interactions among between-replication variables on the RMSE of $\alpha_{0\cdot}$.

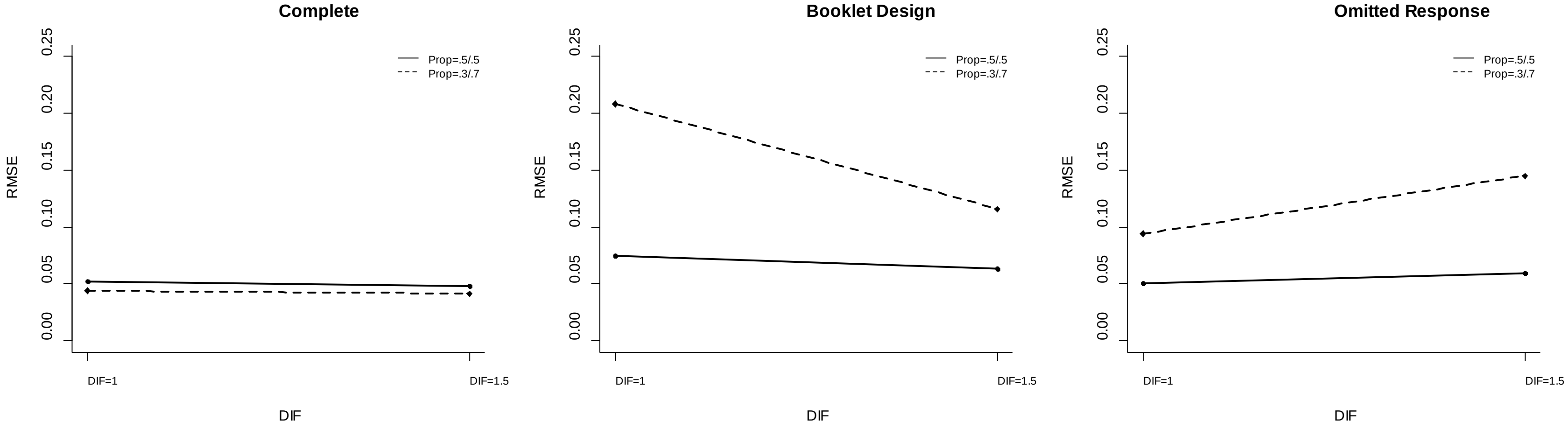

Figure 4.4d. Statistically significant three-way interactions among between-replication variables on the RMSE of $\alpha_{0\cdot}$.

Table 4.10a. *Statistically significant ANOVA results of manipulated factors on the bias of* $\alpha_{1\cdot}$.

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,2352}$=510.398	<.001	0.218	0.528
Corr	$F_{1,2352}$=497.665	<.001	0.106	0.345
Prop	$F_{1,2352}$=311.178	<.001	0.066	0.267
Data*Corr	$F_{2,2352}$=128.036	<.001	0.055	0.241

Table 4.10b. *Statistically significant ANOVA results of manipulated factors on the SE of* $\alpha_{1\cdot}$.

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,2}$=874.827	<.01	0.333	0.707
Prop	$F_{1,2}$=5265.557	<.001	0.667	1.414

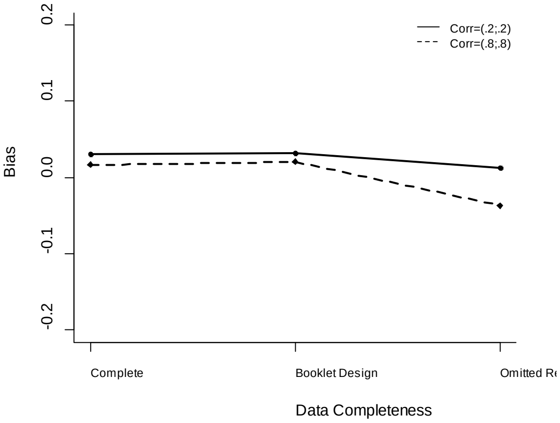

Figure 4.5a. Statistically significant two-way interactions among between-replication variables on the bias of $\alpha_{1\cdot}$.

Table 4.10c. *Statistically significant ANOVA results of manipulated factors on the RMSE of* $\alpha_{1.}$.

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,2}$=441.984	<.01	0.043	0.213
Prop	$F_{1,2}$=7177.970	<.001	0.391	0.802
Data*Corr	$F_{2,2}$=1728.281	<.01	0.217	0.527
Data*Prop	$F_{2,2}$=1251.440	<.01	0.130	0.387
Data*DIF	$F_{2,2}$=437.524	<.01	0.043	0.213
Corr*DIF	$F_{1,2}$=466.685	<.01	0.043	0.213

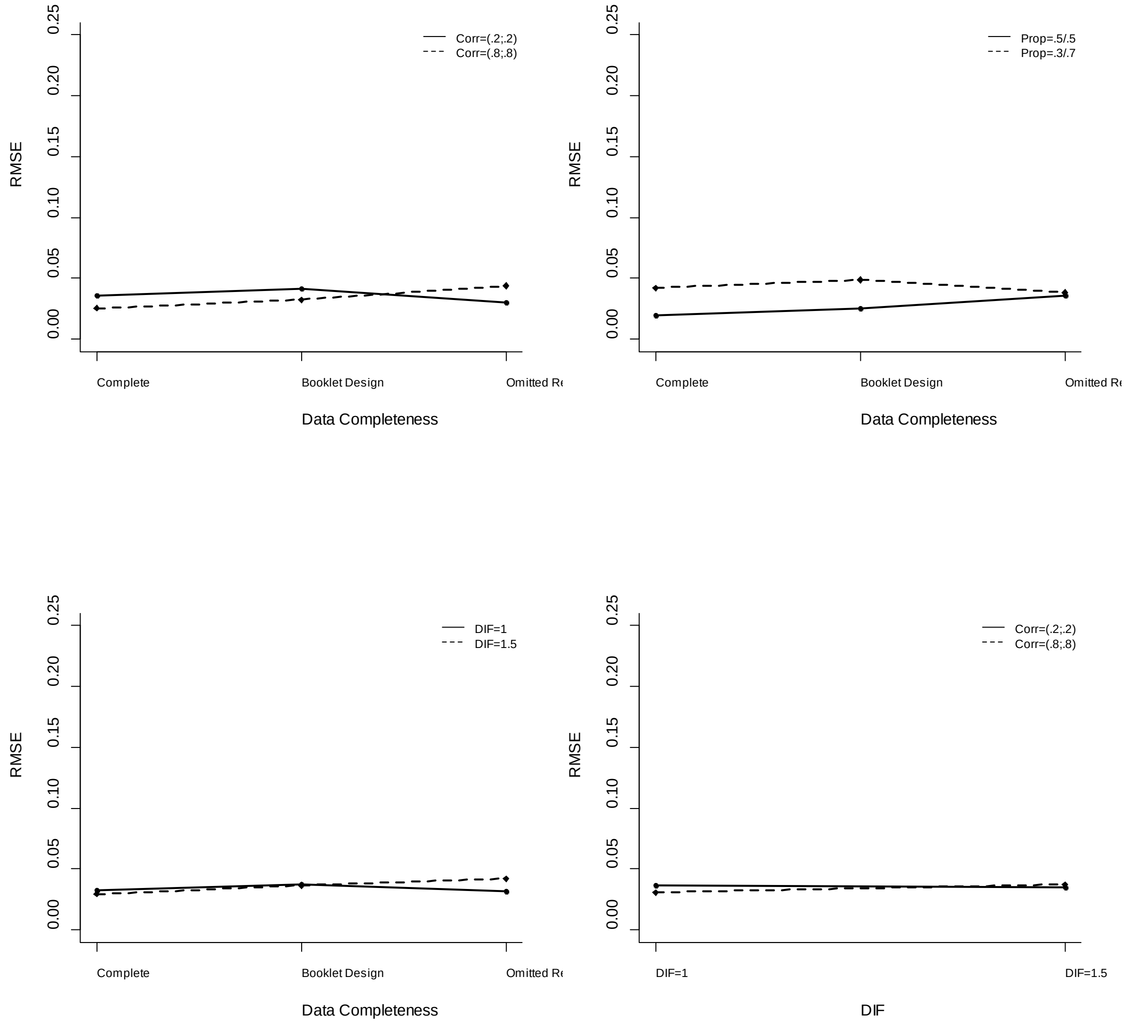

Figure 4.5b. Statistically significant two-way interactions among between-replication variables on the RMSE of $\alpha_{1.}$.

Table 4.9a to 4.9c presented the ANOVA results of the manipulated factors on the evaluation criteria of the intercept parameter. The main effects of model, data completeness, mixing proportion and DIF on the bias of the intercept parameter were statistically significant. Data completeness (*f*=0.257) and mixing proportion (*f*=0.362) had moderate effect sizes, and the effect sizes for model (*f*=0.138) and DIF (*f*=0.181) were small. In addition, the interactions of data completeness by mixing proportion (*f*=0.159) and DIF (*f*=0.380), and the interaction between mixing proportion and DIF (*f*=0.117) were significantly related to the bias of the intercept. The three-way interaction among data completeness, mixing proportion and DIF was also found to be significant with a small effect size (*f*=0.248). Figure 4.4a showed that the effects of mixing proportion and DIF were more pronounced in the booklet design condition. In equal mixing proportion or large DIF conditions, the intercept parameter tended to be underestimated, whereas unequal mixing proportion or smaller DIF led to an overestimation of the intercept. For the three-way interaction, it was observed in Figure 4.4b that the interaction between mixing proportion and DIF was more pronounced in the omitted response condition. There was a relatively large overestimation of the intercept parameter when unequal mixing proportion was paired with smaller DIF. With respect to the SE of the intercept parameter, the ANOVA results indicated that only main effects of model (*f*=0.764), data completeness (*f*=0.679), correlation (*f*=0.236) and DIF (*f*=0.343) were statistically significant. In addition, the manipulated factors that significantly affected the bias of the intercept similarly impacted the RMSE of the intercept parameter, with the exception that correlation, instead of DIF, was found to be statistically significant for the RMSE. The interaction effects were graphically displayed in Figure 4.4c and 4.4d.

As for the recovery of the slope parameter, Table 4.10a to 4.10c separately presented the effects of manipulated factors on the bias, SE and RMSE of the slope parameter. Data completeness, correlation and mixing proportion had statistically significant effects on the bias of the slope parameter, respectively with large (*f*=0.528) and moderate (*f*=0.345; *f*=0.267) effect sizes. The interaction term of data completeness by correlation was also statistically significant, with the effect of correlation on the bias more pronounced in the omitted response condition, as shown in Figure 4.5a. Regarding the SE of the slope parameter, only data completeness (*f*=0.707) and mixing proportion (*f*>1) were found to be statistically significant. Finally, for the RMSE of the slope parameter, data completeness (*f*=0.213) and mixing proportion (*f*=0.802) tended to have significant impacts. Moreover, the interaction terms of data completeness by correlation (*f*=0.527), mixing proportion (*f*=0.387) and DIF (*f*=0.213), and the interaction between correlation and DIF (*f*=0.213) also significantly impacted the RMSE of the slope parameter as shown in Figure 4.5b.

The next part of this section presented the evaluation of the logistic regression parameters. The descriptive statistics of bias, SE and RMSE of the intercept and slope parameters were included in Table 4.11. Different from the linear regression parameters which were mostly overestimated, the marginal bias values for most manipulated factors were negative for both the intercept and the slope parameter in the logistic regression, indicating a tendency of underestimation. However, when the mixing proportion was equal, the odds ratio was large, or the missing data were either not present or caused by omitted responses, the average bias across other manipulated parameters was positive for the intercept parameter. With regard to model, the MRM with only the dichotomous covariate resulted in the least biased estimates of the intercept and the overspecified model had the least biased estimates of the slope parameter. However, for overall recovery of the regression parameters as indicated by RMSE, the true model was the best in terms of both the intercept and the slope parameter. Similar to other model parameters, the recovery of logistic regression parameters was the worst in the booklet design condition.

Table 4.11. *The descriptive statistics of logistic regression parameter by manipulated factors*.

Factors	Levels	β_{02} (intercept)						β_{12} (slope)					
		Bias		SE		RMSE		Bias		SE		RMSE	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Model	TM	-0.025	0.196	0.062	0.030	0.147	0.148	-0.045	0.087	0.063	0.023	0.101	0.059
	UMN-D	-0.023	0.201	0.068	0.031	0.151	0.151	-0.055	0.072	0.068	0.023	0.104	0.050
	OM	-0.033	0.203	0.066	0.033	0.153	0.154	-0.029	0.092	0.070	0.027	0.105	0.060
Prop	.5/.5	0.060	0.076	0.068	0.029	0.104	0.061	-0.033	0.080	0.069	0.023	0.100	0.045
	.3/.7	-.114	0.241	0.063	0.033	0.196	0.194	-0.053	0.088	0.065	0.026	0.106	0.066
OR	1	-0.064	0.212	0.060	0.029	0.145	0.178	0.000	0.066	0.059	0.021	0.080	0.041
	10	0.010	0.178	0.071	0.032	0.155	0.116	-0.086	0.078	0.075	0.025	0.126	0.060
Corr	.2;.2	-0.037	0.207	0.067	0.034	0.150	0.164	-0.046	0.080	0.070	0.027	0.104	0.055
	.8;.8	-0.017	0.191	0.064	0.028	0.151	0.136	-0.040	0.089	0.065	0.022	0.102	0.058
DIF	1	-0.051	0.212	0.077	0.034	0.156	0.173	-0.081	0.073	0.074	0.023	0.116	0.064
	1.5	-0.003	0.183	0.054	0.023	0.145	0.124	-0.005	0.077	0.060	0.024	0.090	0.044
Data	Complete	0.017	0.025	0.042	0.014	0.049	0.019	-0.027	0.032	0.046	0.013	0.059	0.021
	Booklet Design	-0.194	0.258	0.099	0.028	0.281	0.187	-0.050	0.123	0.091	0.020	0.149	0.058
	Omitted Response	0.096	0.083	0.056	0.013	0.121	0.068	-0.072	0.072	0.064	0.013	0.101	0.041

Notes: TM: the data-generating model; UNM-D: the MRM with the dichotomous covariate only; OM: the over specified model.

Table 4.12a. *Statistically significant ANOVA results of manipulated factors on the bias of* β_{02} .

Source	F value	*p* value	η^2	Cohen's *f*
Within(Sphericity Assumed):				
Model	$F_{12,2304}$=150.201	<.001	0.010	0.100
Between:				
Data	$F_{2,1152}$=18.244	<.001	0.011	0.103
OR	$F_{1,1152}$=544.643	<.001	0.158	0.433
DIF	$F_{1,1152}$=414.510	<.001	0.120	0.369
Data*OR	$F_{2,1152}$=19.556	<.001	0.011	0.107
Data*Prop	$F_{2,1152}$=82.955	<.001	0.048	0.225
Data*DIF	$F_{2,1152}$=217.724	<.001	0.126	0.380
Data*Prop*OR	$F_{2,1152}$=54.674	<.001	0.032	0.181
Data*Prop*DIF	$F_{2,1152}$=19.739	<.001	0.011	0.108

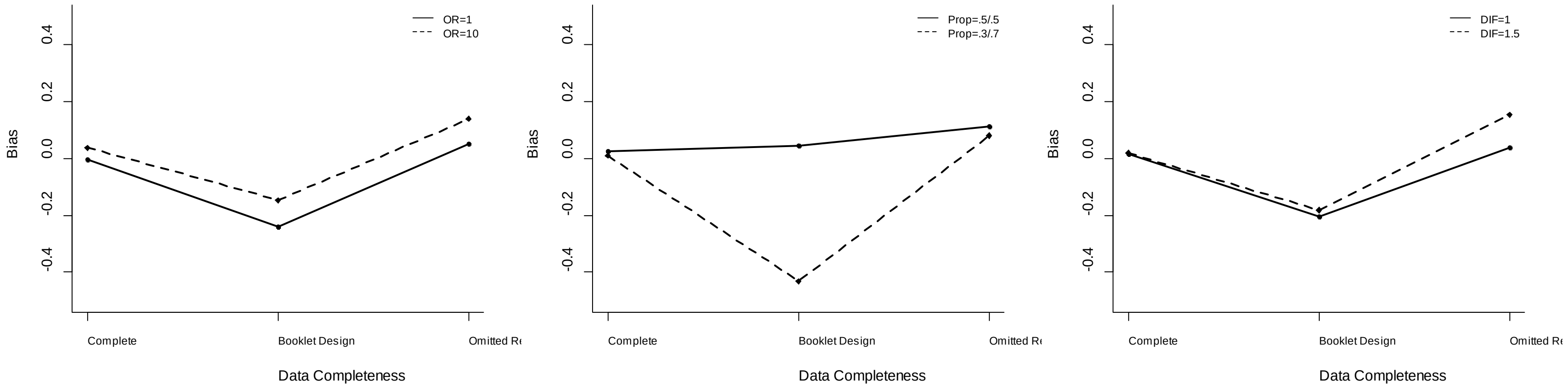

Figure 4.6a. Statistically significant two-way interactions among between-replication variables on the bias of β_{02} .

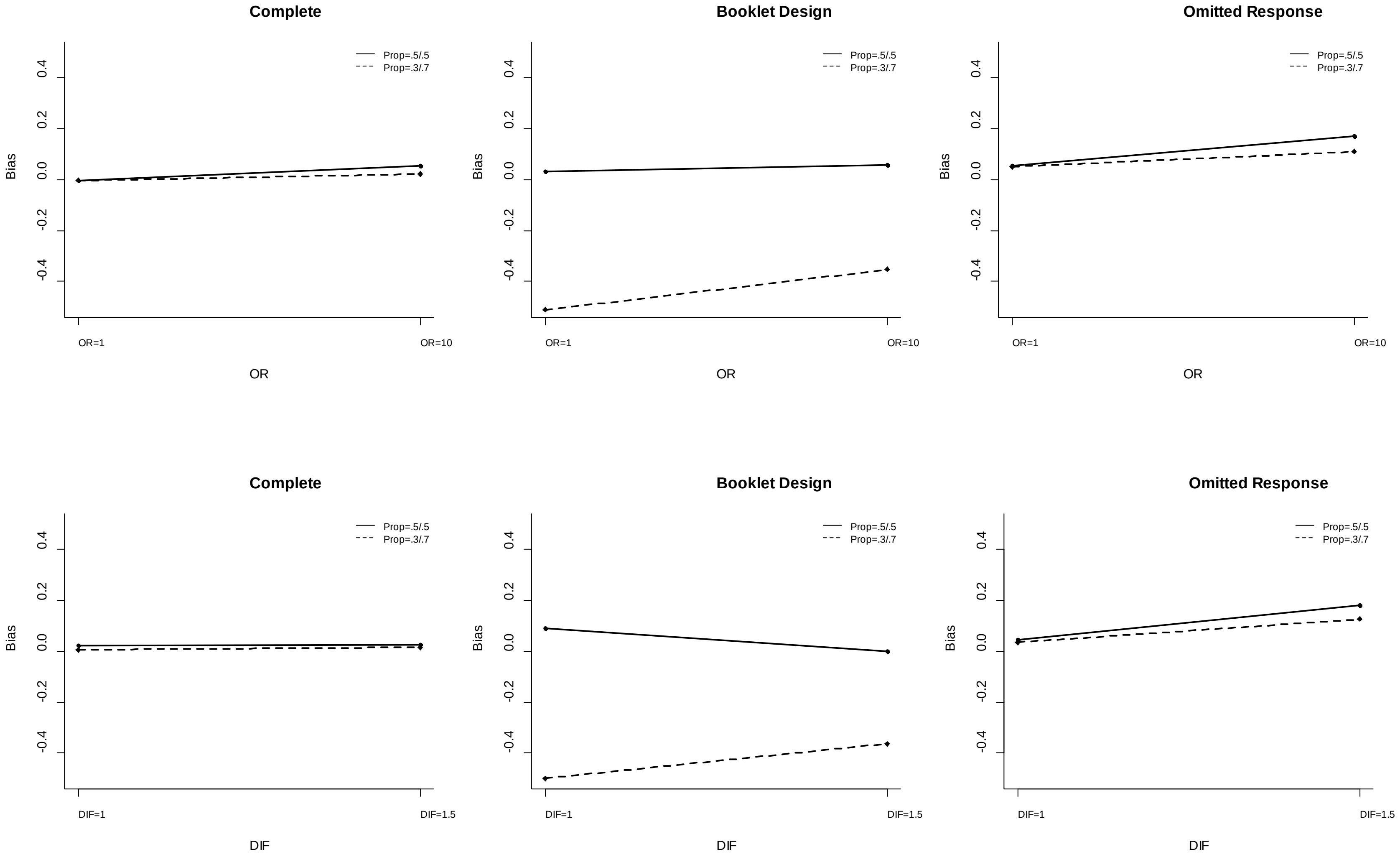

Figure 4.6b. Statistically significant three-way interactions among between-replication variables on the bias of β_{02} .

Table 4.12b. *Statistically significant ANOVA results of manipulated factors on the SE of* β_{02} .

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,2}$=1022.075	<.001	0.609	1.247
OR	$F_{1,2}$=122.398	<.01	0.036	0.194
DIF	$F_{1,2}$=492.393	<.01	0.145	0.412
Data*Prop	$F_{2,2}$=22.923	<.05	0.014	0.121
Data*DIF	$F_{2,2}$=77.299	<.001	0.043	0.213
OR*Prop	$F_{1,2}$=72.071	<.05	0.022	0.149

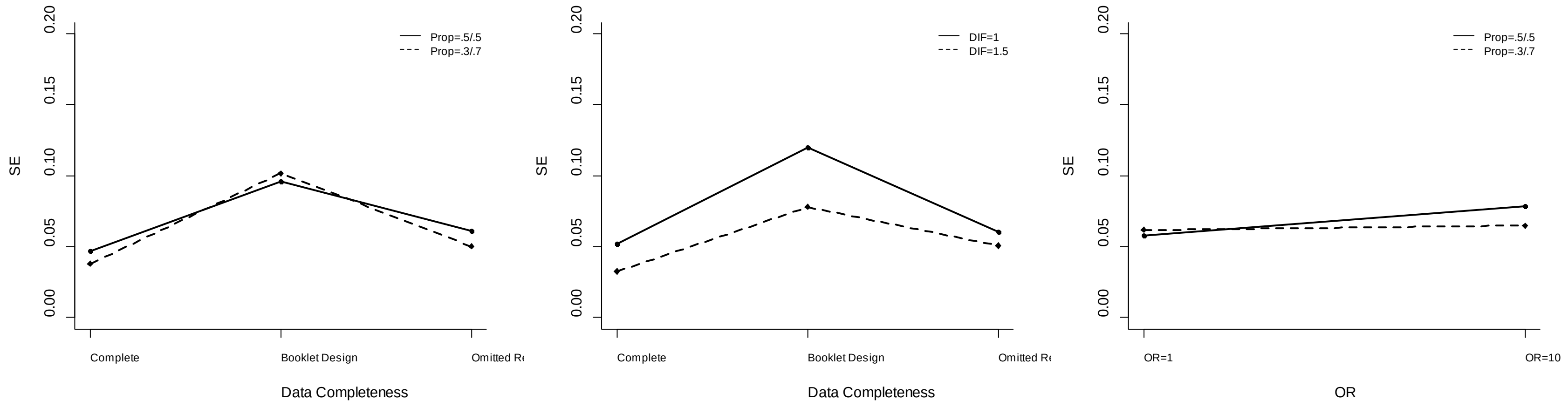

Figure 4.6c. Statistically significant two-way interactions among between-replication variables on the SE of β_{02} .

Table 4.12c. *Statistically significant ANOVA results of manipulated factors on the RMSE of* β_{02} .

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
Data	$F_{2,2}$=12889.293	<.001	0.417	0.846
Data*OR	$F_{2,2}$=1125.760	<.01	0.037	0.195
Data*Prop	$F_{2,2}$=9549.635	<.001	0.309	0.669
Data*DIF	$F_{2,2}$=1992.927	<.01	0.065	0.263
OR*Prop	$F_{1,2}$=1274.609	<.01	0.021	0.146
Data*OR*Prop	$F_{2,2}$=342.579	<.01	0.011	0.106

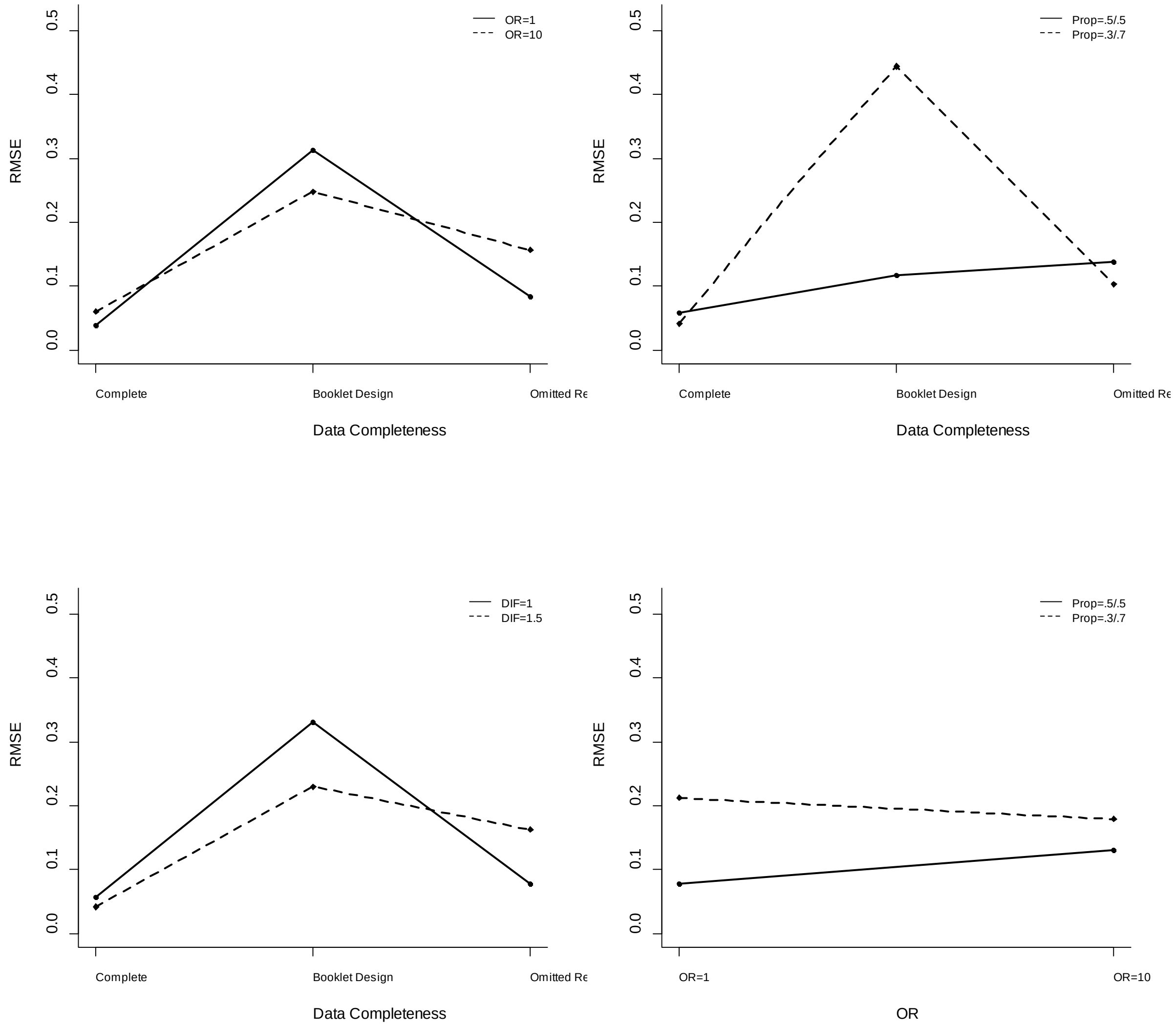

Figure 4.6d. Statistically significant two-way interactions among between-replication variables on the RMSE of β_{02} .

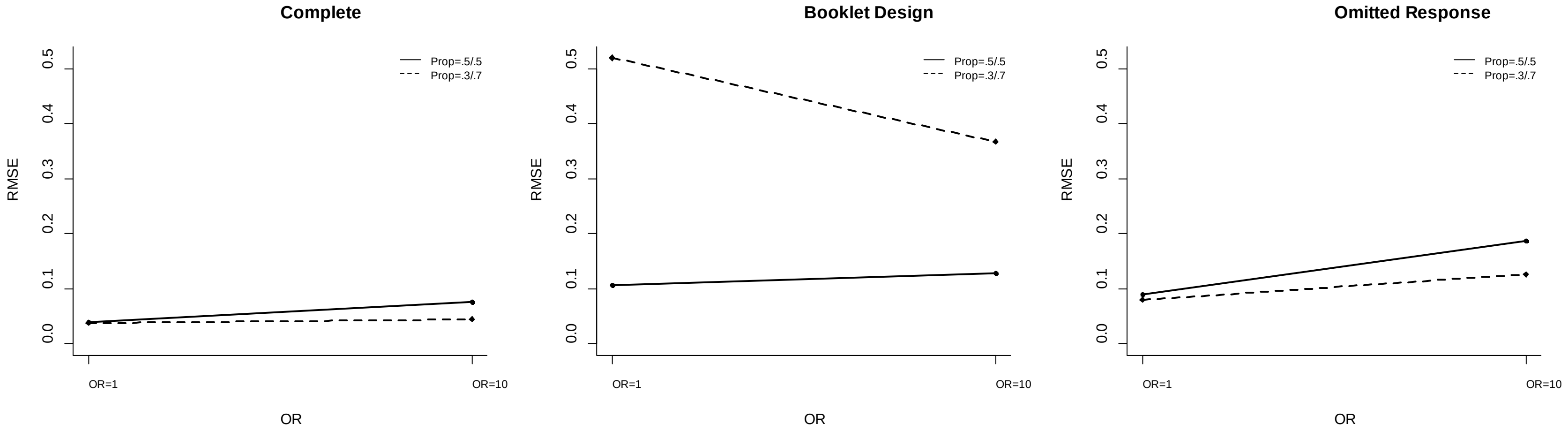

Figure 4.6e. Statistically significant three-way interactions among between-replication variables on the RMSE of β_{02} .

Table 4.13a. *Statistically significant ANOVA results of manipulated factors on the bias of* β_{12} .

Source	F value	*p* value	η^2	Cohen's *f*
Between:				
OR	$F_{1,1152}$=544.643	<.001	0.043	0.211
DIF	$F_{1,1152}$=414.510	<.001	0.032	0.183
Data*Prop	$F_{2,1152}$=82.955	<.001	0.013	0.115
Data*DIF	$F_{2,1152}$=217.724	<.001	0.034	0.188

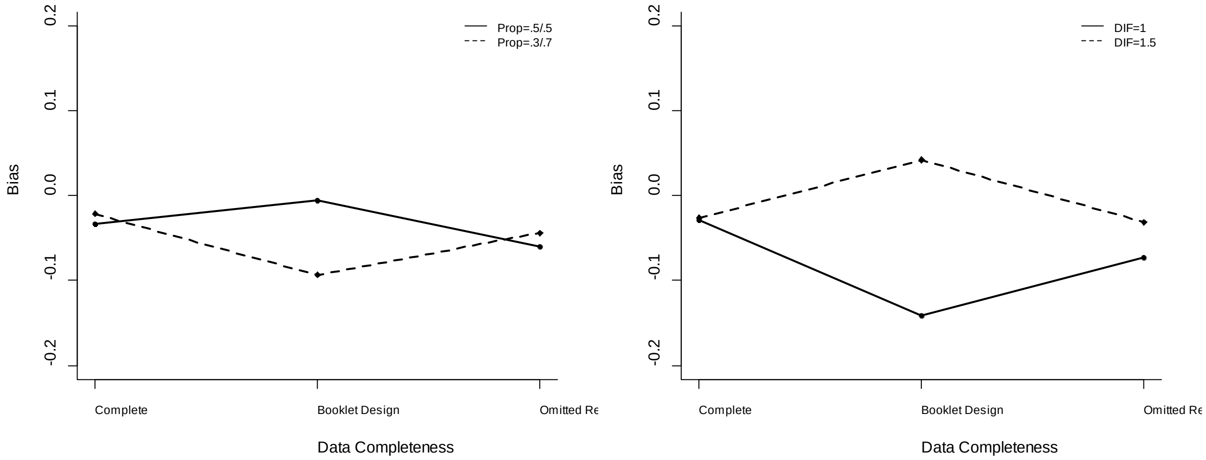

Figure 4.7a. Statistically significant two-way interactions among between-replication variables on the bias of β_{12} .

Table 4.13b. *Statistically significant ANOVA results of manipulated factors on the SE of* β_{12} .

Source	F value	*p* value	η^2	Cohen's *f*
Within(Huynh-Feldt Correction):				
Model	$F_{2,4}$=202.035	<.001	0.012	0.109
Model*Data	$F_{4,4}$=41.718	<.01	0.012	0.109
Model*DIF	$F_{2,4}$=101.770	<.001	0.012	0.109
Between:				
Data	$F_{2,2}$=279.782	<.01	0.588	1.195
OR	$F_{1,2}$=106.226	<.01	0.106	0.344
DIF	$F_{1,2}$=75.667	<.05	0.082	0.300

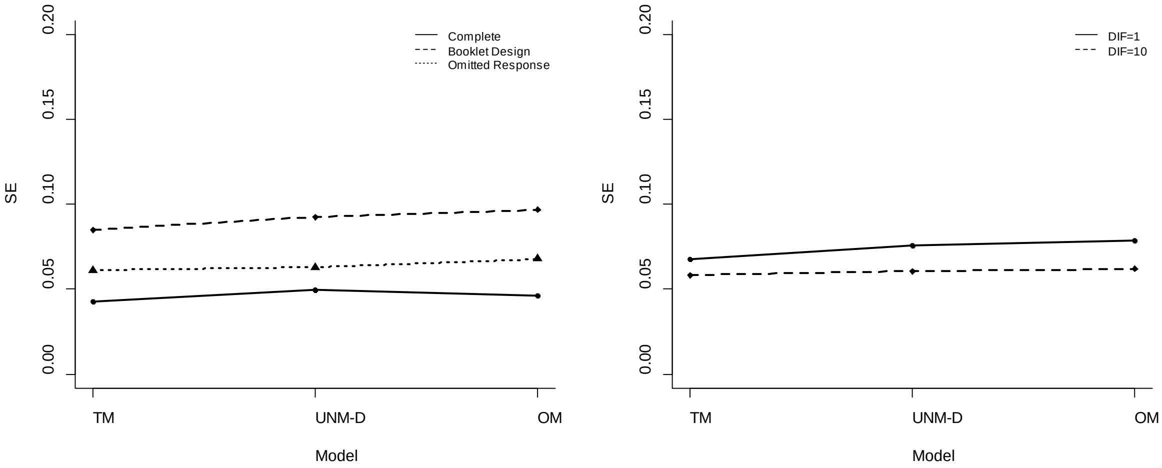

Figure 4.7b. Statistically significant two-way interactions between model and between-replication variables on the SE of β_{12} .

Table 4.13c. *Statistically significant ANOVA results of manipulated factors on the RMSE of* β_{12}.

Source	F value	*p* value	η^2	Cohen's *f*
Within(Sphericity Assumed):				
Model*Data	$F_{2,4}$=45.015	<.01	0.013	0.116
Between:				
Data	$F_{2,2}$=1681.218	<.01	0.433	0.874
OR	$F_{1,2}$=1297.059	<.01	0.167	0.448
DIF	$F_{1,2}$=429.121	<.01	0.055	0.241
Data*OR	$F_{2,2}$=40.777	<.05	0.011	0.105
Data* Corr	$F_{2,2}$=53.736	<.05	0.013	0.116
Data* Prop	$F_{2,2}$=91.546	<.05	0.024	0.157
Data*DIF	$F_{2,2}$=110.125	<.01	0.029	0.171

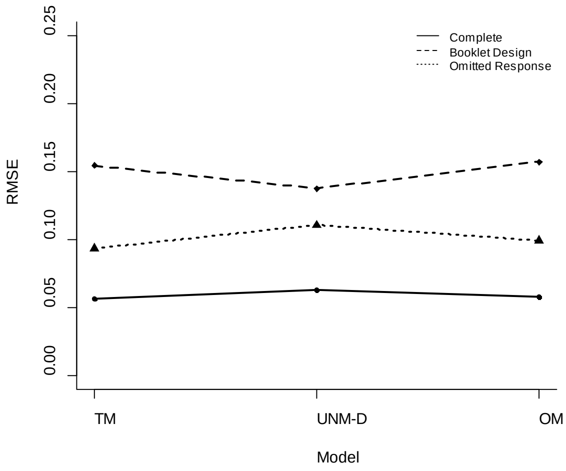

Figure 4.7c. Statistically significant two-way interactions between model and data completeness on the RMSE of β_{12}.

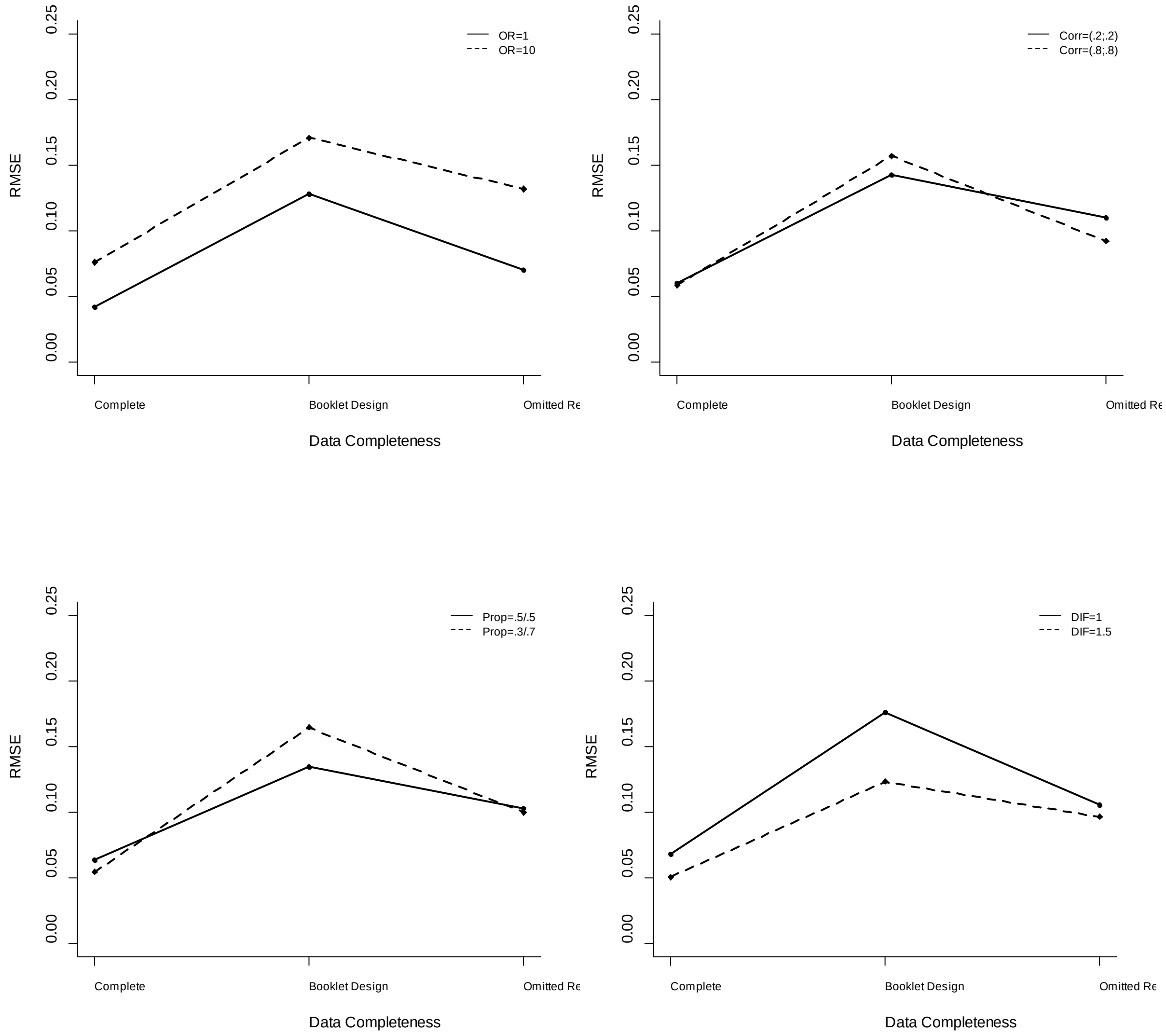

Figure 4.7d. Statistically significant two-way interactions between model and between-replication variables on the RMSE of β_{12}.

As with the other model parameters, repeated measures ANOVA were used to identify significant effects with regard to the evaluation criteria of the logistic regression parameters. Table 4.12a to 4.12c presented the results for the intercept parameter. The effects of model (*f*=0.100), data completeness (*f*=0.103), odds ratio (*f*=0.433) and DIF (*f*=0.369) on the bias of the intercept parameter were statistically significant. In addition, the interaction terms of data completeness by odds ratio (*f*=0.107), mixing proportion (*f*=0.225) and DIF (*f*=0.380), and the three-way interactions among data completeness, mixing proportion and odds ratio (*f*=0.181) and among data completeness, mixing proportion and DIF (*f*=0.108) were significantly related to the bias of the intercept. The interaction effects were displayed in Figure 4.6a and 4.6b. It was interesting to find that the intercept bias was slightly positive when the mixing proportion was equal in the booklet design condition. However, when the mixing proportion was unequal, there was a negative bias of around 0.400 in the intercept parameter. In addition, the two-way interactions of mixing proportion by odds ratio and DIF were also more pronounced in the booklet design condition.

Regarding the SE, data completeness had a statistically significant effect with a very large effect size (*f*>1). The effects of odds ratio and DIF were also significant, respectively with a small (*f*=0.194) and a large effect size. Moreover, data completeness significantly interacted with mixing proportion (*f*=0.121) and DIF (*f*=0.213), and odds ratio significantly interacted with mixing proportion (*f*=0.149) on the SE of the intercept parameter with small effect sizes (see Figure 4.6c).

Further, for the RMSE of the intercept parameter, the two-way interaction terms of data completeness by odds ratio (*f*=0.195), mixing proportion (*f*=0.669) and DIF (*f*=0.263), and the interaction between odds ratio and mixing proportion (*f*=0.146) were found to be statistically significant. The three-way interactions among data completeness, mixing proportion and odds ratio (*f*=0.106) was also statistically significant. In addition, data completeness itself was significant with a large effect size (*f*=0.846). The interaction effects were shown in Figure 4.6d and 4.6e.

As for the slope parameter in the logistic regression function, odds ratio (*f*=0.211) and DIF (*f*=0.183) tended to have significant main effects on its bias with small effect sizes. Moreover, data completeness had significant interactions with mixing proportion (*f*=0.115) and DIF (*f*=0.188) in impacting the bias. As shown in Figure 4.7a, the slope bias was slightly positive when the mixing proportion was equal or when DIF was larger in the booklet design condition. However, when the mixing proportion was unequal or the DIF was smaller, there was a negative bias of around 0.100 in the slope parameter.

Regarding the SE of the slope parameter, the effects of model (*f*=0.109), data completeness (*f*>1), odds ratio (*f*=0.344) and DIF (*f*=0.300) were statistically significant. The interactions of model by data completeness (*f*=0.109) and DIF (*f*=0.109) were also significant with small effect sizes, so that they were barely observable in Figure 4.7b.

Finally, for the RMSE of the slope parameter, the effects of data completeness (*f*=0.874), odds ratio (*f*=0.448) and DIF (*f*=0.241) were statistically significant, among which the effect sizes of data completeness and odds ratio were large. Also, data completeness interacted significantly with odds ratio (*f*=0.105), correlation (*f*=0.116), mixing proportion (*f*=0.157) and DIF (*f*=0.171) in impacting the RMSE of the slope parameter. The interaction between model and data completeness (*f*=0.116) was also statistically significant, yet its effect size was quite small so that it was not obvious in Figure 4.7c. Additionally, Figure 4.7d showed an interesting pattern that the effects of odds ratio, correlation, mixing proportion and DIF were all more pronounced in the booklet design condition.

In summary, regarding the recovery of item and person parameters, data completeness, mixing proportion, DIF and certain interactions among these manipulated factors tended to have great impacts. The effect size of model was not large enough to claim a practical significance of covariate inclusion on item and person parameter recovery, yet the marginal descriptive statistics indicated a tendency of better recovery if covariates were correctly specified in the MRM. These results were largely in line with the findings of previous research (Adams et al., 1997; Mislevy & Sheehan, 1989a, 1989b; Smit et al., 1999, 2000) that the incorporation of important covariates could reduce the mean squared error of person parameter estimates and the standard error of item parameter estimates. In the current simulation, the MRM with only the continuous covariate tended to perform better than the MRM with only the dichotomous covariate in both item and person parameter recovery. Not surprisingly, the MRM with mismatching covariates performed the worst in parameter recovery as it did in latent class assignment.

With regard to the recovery of regression parameters, it was interesting to find a tendency of overestimation of the linear regression parameters as well as a tendency of underestimation of the logistic regression parameters. Another important finding was that the quality of regression parameter recovery was very sensitive to the manipulated factors, especially DIF and mixing proportion, in the booklet design condition. Last but not least, the model with only one covariate correctly specified in the model tended to result in the least biased intercept parameter estimates in both the linear regression and the logistic regression functions, while the overspecified model had the least biased slope parameter estimates in both functions. Further, the true model performed the best in recovering the intercept parameters in both functions, and the slope parameter in the logistic function. Possible explanations to these finding would be detailed in Chapter 5.

4.1.3 Overall model fit indices.

The frequency of each model being selected as the best-fitting model with respect to the six overall model selection indices, including AIC, BIC, AICc, CAIC, SABIC and DIC, were summarized in Table 4.14b. The selection decision for each simulation condition, the marginal totals by data completeness and the totals were provided in Table 4.14b.

In addition, Figure 4.8a graphically displayed the overall selection decision in terms of percentage. The reason to provide a percentage figure was to give a clear presentation of the performance of different overall model fit indices under a variety of simulation conditions. Frequency results may be easily obtained by multiplying the percentage by the number of replications, which was 25 in the present study.

Table 4.14a. *Model selection frequency for simulation cells.*

					Model Selection Indices																																			
					AIC						BIC						AICc						CAIC						SABIC						DIC					
Data	DIF	OR	Corr	Prop	TM	M2	M3	M4	M5	M6	TM	M2	M3	M4	M5	M6	TM	M2	M3	M4	M5	M6	TM	M2	M3	M4	M5	M6	TM	M2	M3	M4	M5	M6	TM	M2	M3	M4	M5	M6
Complete	1.5	10	.2;.2	.5;.5	2	12	0	11	0	0	0	19	0	6	0	0	1	13	0	11	0	0	0	19	0	6	0	0	0	13	0	12	0	0	15	0	0	0	10	0
				.3;.7	4	10	3	8	0	0	0	16	0	9	0	0	3	9	3	10	0	0	0	17	0	8	0	0	0	14	0	11	0	0	15	0	3	0	7	0
			.8;.8	.5;.5	4	5	10	6	0	0	0	13	7	5	0	0	4	6	9	6	0	0	0	14	6	5	0	0	3	7	9	6	0	0	17	0	1	0	7	0
				.3;.7	18	2	5	0	0	0	5	6	8	6	0	0	18	2	5	0	0	0	4	8	6	7	0	0	13	2	7	3	0	0	13	0	2	0	10	0
		1	.2;.2	.5;.5	1	20	3	1	0	0	0	25	0	0	0	0	1	21	2	1	0	0	0	25	0	0	0	0	0	25	0	0	0	0	7	0	12	0	6	0
				.3;.7	1	21	3	0	0	0	0	25	0	0	0	0	1	21	3	0	0	0	0	25	0	0	0	0	0	25	0	0	0	0	12	0	3	0	10	0
			.8;.8	.5;.5	0	11	14	0	0	0	0	16	9	0	0	0	0	12	13	0	0	0	0	19	6	0	0	0	0	13	12	0	0	0	6	0	10	0	9	0
				.3;.7	16	1	8	0	0	0	0	13	12	0	0	0	16	1	8	0	0	0	0	16	9	0	0	0	8	8	9	0	0	0	13	0	3	0	9	0
	1	10	.2;.2	.5;.5	1	7	0	17	0	0	0	10	0	15	0	0	1	7	0	17	0	0	0	11	0	14	0	0	0	9	0	16	0	0	17	0	0	1	7	0
				.3;.7	10	3	3	8	0	1	0	11	0	14	0	0	10	3	3	8	0	1	0	12	0	13	0	0	2	7	2	14	0	0	14	0	0	2	9	0
			.8;.8	.5;.5	9	2	3	11	0	0	1	6	2	16	0	0	9	2	3	11	0	0	1	7	1	16	0	0	7	2	3	13	0	0	17	0	0	0	8	0
				.3;.7	16	1	8	0	0	0	6	4	11	4	0	0	15	1	9	0	0	0	6	5	10	4	0	0	9	3	11	2	0	0	16	0	0	0	9	0
		1	.2;.2	.5;.5	0	23	1	1	0	0	0	25	0	0	0	0	0	23	1	1	0	0	0	25	0	0	0	0	0	25	0	0	0	0	12	1	1	0	11	0
				.3;.7	1	14	6	4	0	0	0	25	0	0	0	0	0	14	7	4	0	0	0	25	0	0	0	0	0	22	2	1	0	0	3	0	13	0	9	0
			.8;.8	.5;.5	14	8	1	2	0	0	1	23	1	0	0	0	13	9	1	2	0	0	0	24	1	0	0	0	8	16	1	0	0	0	17	0	1	0	7	0
				.3;.7	7	3	15	0	0	0	0	9	16	0	0	0	6	4	15	0	0	0	0	10	15	0	0	0	0	6	19	0	0	0	4	0	11	0	10	0
Booklet Design	1.5	10	.2;.2	.5;.5	0	0	23	2	0	0	0	12	10	3	0	0	0	0	23	2	0	0	0	18	5	2	0	0	0	4	19	2	0	0	11	0	0	0	14	0
				.3;.7	0	9	4	4	0	8	0	21	1	3	0	0	0	9	4	4	0	8	0	22	1	2	0	0	0	17	2	3	0	3	6	0	0	1	18	0
			.8;.8	.5;.5	0	2	23	0	0	0	0	4	21	0	0	0	0	2	23	0	0	0	0	4	21	0	0	0	0	4	21	0	0	0	1	0	11	0	13	0
				.3;.7	3	9	5	2	6	0	1	12	6	6	0	0	3	9	5	2	6	0	1	12	6	6	0	0	3	9	7	4	2	0	4	0	1	0	20	0
		1	.2;.2	.5;.5	0	0	17	8	0	0	0	3	6	16	0	0	0	0	17	8	0	0	0	12	3	10	0	0	0	0	16	9	0	0	0	0	18	1	6	0
				.3;.7	0	0	1	24	0	0	0	6	0	19	0	0	0	0	1	24	0	0	0	7	0	18	0	0	0	1	1	23	0	0	0	0	3	1	21	0
			.8;.8	.5;.5	0	0	21	4	0	0	0	1	20	4	0	0	0	0	21	4	0	0	0	1	20	4	0	0	0	0	21	4	0	0	0	0	18	0	7	0
				.3;.7	1	0	10	14	0	0	0	1	7	17	0	0	1	0	10	14	0	0	0	4	7	14	0	0	0	0	10	15	0	0	7	0	4	0	14	0
	1	10	.2;.2	.5;.5	10	0	0	15	0	0	0	0	0	25	0	0	9	0	0	16	0	0	0	0	0	25	0	0	1	0	0	24	0	0	22	0	0	1	2	0
				.3;.7	2	3	1	1	0	18	0	14	0	2	0	9	2	4	1	1	0	17	0	16	0	1	0	8	1	7	1	3	0	13	12	0	2	0	11	0
			.8;.8	.5;.5	24	0	0	1	0	0	17	0	0	8	0	0	24	0	0	1	0	0	17	0	0	8	0	0	23	0	0	2	0	0	17	0	1	0	7	0
				.3;.7	3	12	7	1	2	0	1	18	3	3	0	0	3	12	7	1	2	0	1	18	3	3	0	0	3	14	6	1	1	0	12	0	2	0	11	0
		1	.2;.2	.5;.5	2	12	8	3	0	0	0	25	0	0	0	0	0	15	7	3	0	0	0	25	0	0	0	0	0	20	3	2	0	0	4	0	21	0	0	0
				.3;.7	2	18	4	1	0	0	0	25	0	0	0	0	2	18	4	1	0	0	0	25	0	0	0	0	0	21	4	0	0	0	7	0	13	3	2	0
			.8;.8	.5;.5	0	1	24	0	0	0	0	3	22	0	0	0	0	1	24	0	0	0	0	5	20	0	0	0	0	2	22	1	0	0	4	0	21	0	0	0
				.3;.7	3	13	8	1	0	0	0	21	4	0	0	0	3	13	8	1	0	0	0	22	3	0	0	0	0	15	8	2	0	0	11	0	14	0	0	0
Omitted Responses	1.5	10	.2;.2	.5;.5	1	0	0	24	0	0	0	0	0	25	0	0	1	0	0	24	0	0	0	0	0	25	0	0	0	0	0	25	0	0	14	0	0	0	11	0
				.3;.7	2	0	0	20	3	0	0	0	0	25	0	0	2	0	0	20	3	0	0	0	0	25	0	0	1	0	0	24	0	0	11	0	0	0	14	0
			.8;.8	.5;.5	13	0	12	0	0	0	8	0	17	0	0	0	13	0	12	0	0	0	8	0	17	0	0	0	12	0	13	0	0	0	16	0	0	0	9	0
				.3;.7	20	0	1	0	4	0	24	0	1	0	0	0	20	0	1	0	4	0	24	0	1	0	0	0	24	0	1	0	0	0	10	0	0	0	15	0
		1	.2;.2	.5;.5	0	2	20	3	0	0	0	23	0	2	0	0	0	2	20	3	0	0	0	23	0	2	0	0	0	15	4	6	0	0	0	0	13	0	12	0
				.3;.7	0	11	4	9	0	1	0	24	0	1	0	0	0	12	2	10	0	1	0	25	0	0	0	0	0	20	0	5	0	0	2	0	8	0	15	0
			.8;.8	.5;.5	0	0	25	0	0	0	0	0	25	0	0	0	0	0	25	0	0	0	0	0	25	0	0	0	0	0	25	0	0	0	3	0	22	0	0	0

			.3;.7	4	0	21	0	0	0	0	0	25	0	0	0	3	0	22	0	0	0	0	0	25	0	0	0	0	0	25	0	0	0	5	0	9	0	11	0
1	10	.2;.2	.5;.5	6	3	3	12	0	1	0	9	0	16	0	0	6	3	4	12	0	0	0	10	0	15	0	0	0	5	1	19	0	0	17	0	0	0	8	0
			.3;.7	9	2	2	10	1	1	0	5	0	20	0	0	9	3	2	10	1	0	0	5	0	20	0	0	4	4	1	16	0	0	9	0	1	2	13	0
		.8;.8	.5;.5	16	0	9	0	0	0	11	0	14	0	0	0	16	0	9	0	0	0	11	0	14	0	0	0	12	0	13	0	0	0	18	0	0	0	7	0
			.3;.7	17	0	8	0	0	0	11	0	14	0	0	0	17	0	8	0	0	0	11	0	14	0	0	0	15	0	10	0	0	0	10	0	0	0	15	0
	1	.2;.2	.5;.5	1	15	8	1	0	0	0	25	0	0	0	0	1	15	7	2	0	0	0	25	0	0	0	0	0	23	2	0	0	0	11	0	9	0	4	1
			.3;.7	3	9	12	1	0	0	0	25	0	0	0	0	3	11	10	1	0	0	0	25	0	0	0	0	0	22	3	0	0	0	8	0	10	0	7	0
		.8;.8	.5;.5	10	0	15	0	0	0	1	0	24	0	0	0	10	0	15	0	0	0	0	0	25	0	0	0	3	0	22	0	0	0	7	0	10	0	8	0
			.3;.7	13	0	12	0	0	0	1	0	24	0	0	0	12	0	13	0	0	0	0	0	25	0	0	0	3	0	22	0	0	0	10	0	7	0	8	0

Notes: TM: the data-generating model; M2: the MRM without covariates; M3: the MRM with the continuous covariate only; M4: the MRM with the dichotomous covariate only; M5: the over specified model; M6: the MRM with mismatching covariates.

Table 4.14b. *Model selection decision by simulation condition and its frequency*.

					Model Selection Indices																																			
					AIC						BIC						AICc						CAIC						SABIC						DIC					
Data	DIF	OR	Corr	Prop	TM	M2	M3	M4	M5	M6	TM	M2	M3	M4	M5	M6	TM	M2	M3	M4	M5	M6	TM	M2	M3	M4	M5	M6	TM	M2	M3	M4	M5	M6	TM	M2	M3	M4	M5	M6
Complete	1.5	10	.2;.2	.5;.5		X						X						X						X						X					X					
				.3;.7		X						X								X				X						X					X					
			.8;.8	.5;.5			X					X							X					X							X				X					
				.3;.7	X								X				X							X					X						X					
		1	.2;.2	.5;.5		X						X						X						X						X							X			
				.3;.7		X						X						X						X						X					X					
			.8;.8	.5;.5			X					X							X					X						X							X			
				.3;.7	X							X					X							X							X				X					
	1	10	.2;.2	.5;.5				X						X						X						X						X			X					
				.3;.7	X									X			X									X						X			X					
			.8;.8	.5;.5				X						X						X						X						X			X					
				.3;.7	X								X				X								X						X				X					
		1	.2;.2	.5;.5		X						X						X						X						X					X					
				.3;.7		X						X						X						X						X							X			
			.8;.8	.5;.5	X							X					X							X						X					X					
				.3;.7			X						X						X						X						X						X			
Marginal Totals					5	6	3	2	0	0	0	10	3	3	0	0	5	5	3	3	0	0	0	11	2	3	0	0	1	8	4	3	0	0	12	0	4	0	0	0
Booklet Design	1.5	10	.2;.2	.5;.5			X					X							X					X							X								X	
				.3;.7		X						X						X						X						X									X	
			.8;.8	.5;.5			X						X						X						X						X								X	
				.3;.7		X						X						X						X						X									X	
		1	.2;.2	.5;.5			X							X					X					X							X						X			
				.3;.7				X						X						X						X						X							X	
			.8;.8	.5;.5			X						X						X						X						X						X			
				.3;.7				X						X						X						X						X							X	
	1	10	.2;.2	.5;.5				X						X						X						X						X			X					
				.3;.7						X		X										X		X										X	X					
			.8;.8	.5;.5	X						X						X						X						X						X					
				.3;.7		X						X						X						X						X					X					
		1	.2;.2	.5;.5		X						X						X						X						X							X			
				.3;.7		X						X						X						X						X							X			
			.8;.8	.5;.5			X						X						X						X						X						X			
				.3;.7		X						X						X						X						X							X			
Marginal Totals					1	6	5	3	0	1	1	8	3	4	0	0	1	6	5	3	0	1	1	9	3	3	0	0	1	6	5	3	0	1	4	0	6	0	6	0
Omitted Responses	1.5	10	.2;.2	.5;.5				X						X						X						X						X			X					
				.3;.7				X						X						X						X						X							X	
			.8;.8	.5;.5	X								X				X								X						X				X					
				.3;.7	X						X						X						X						X										X	
		1	.2;.2	.5;.5			X					X							X					X						X							X			
				.3;.7		X						X						X						X						X									X	
			.8;.8	.5;.5			X						X						X						X						X						X			

			.3;.7			X						X						X						X						X								X	
1	10	.2;.2	.5;.5				X						X						X						X						X			X					
			.3;.7				X						X						X						X						X							X	
		.8;.8	.5;.5	X								X				X								X						X				X					
			.3;.7	X								X				X								X				X										X	
	1	.2;.2	.5;.5		X						X						X						X						X					X					
			.3;.7			X					X						X						X						X							X			
		.8;.8	.5;.5			X						X						X						X						X						X			
			.3;.7	X								X						X						X						X				X					
Marginal Totals				5	2	5	4	0	0	1	4	7	4	0	0	4	3	5	4	0	0	1	4	7	4	0	0	2	4	6	4	0	0	6	0	4	0	6	0
Total				11	14	13	9	0	1	2	22	13	11	0	0	10	14	13	10	0	1	2	24	12	10	0	0	4	18	15	20	0	1	22	0	14	0	12	0

Notes: TM: the data-generating model; M2: the MRM without covariates; M3: the MRM with the continuous covariate only; M4: the MRM with the dichotomous covariate only; M5: the over specified model; M6: the MRM with mismatching covariates.

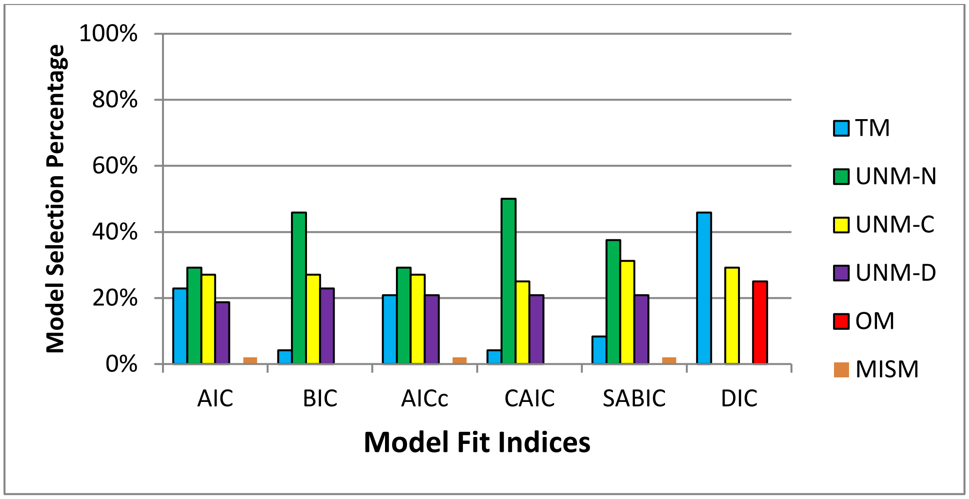

Figure 4.8a. Overall model selection percentage across simulation conditions.

Overall, AIC, BIC, AICc, CAIC and SABIC did not perform very well in identifying the data-generating model. Among them, AIC and AICc performed the worst as they had difficulty differentiating among the true model, the MRM without covariates, the MRM with only the continuous covariate and the MRM with only the dichotomous covariate. They did not have a strong tendency of selecting any of the models. On the other hand, BIC, CAIC and SABIC performed similarly and they all tended to choose the most parsimonious model, the MRM without covariates. The only model fit index that performed well in identifying the MRM with correctly specified covariates was DIC, a Bayesian measure of fit. However, a closer examination of Table 4.14b showed that the performance of model fit indices was different depending on certain levels of manipulated factors. Thus, the selection decisions were analyzed with respect to data completeness, correlation and odds ratio. It was assumed that the missing data and the strength of relations between covariates and model parameters may impact the selection performance of these overall fit indices.

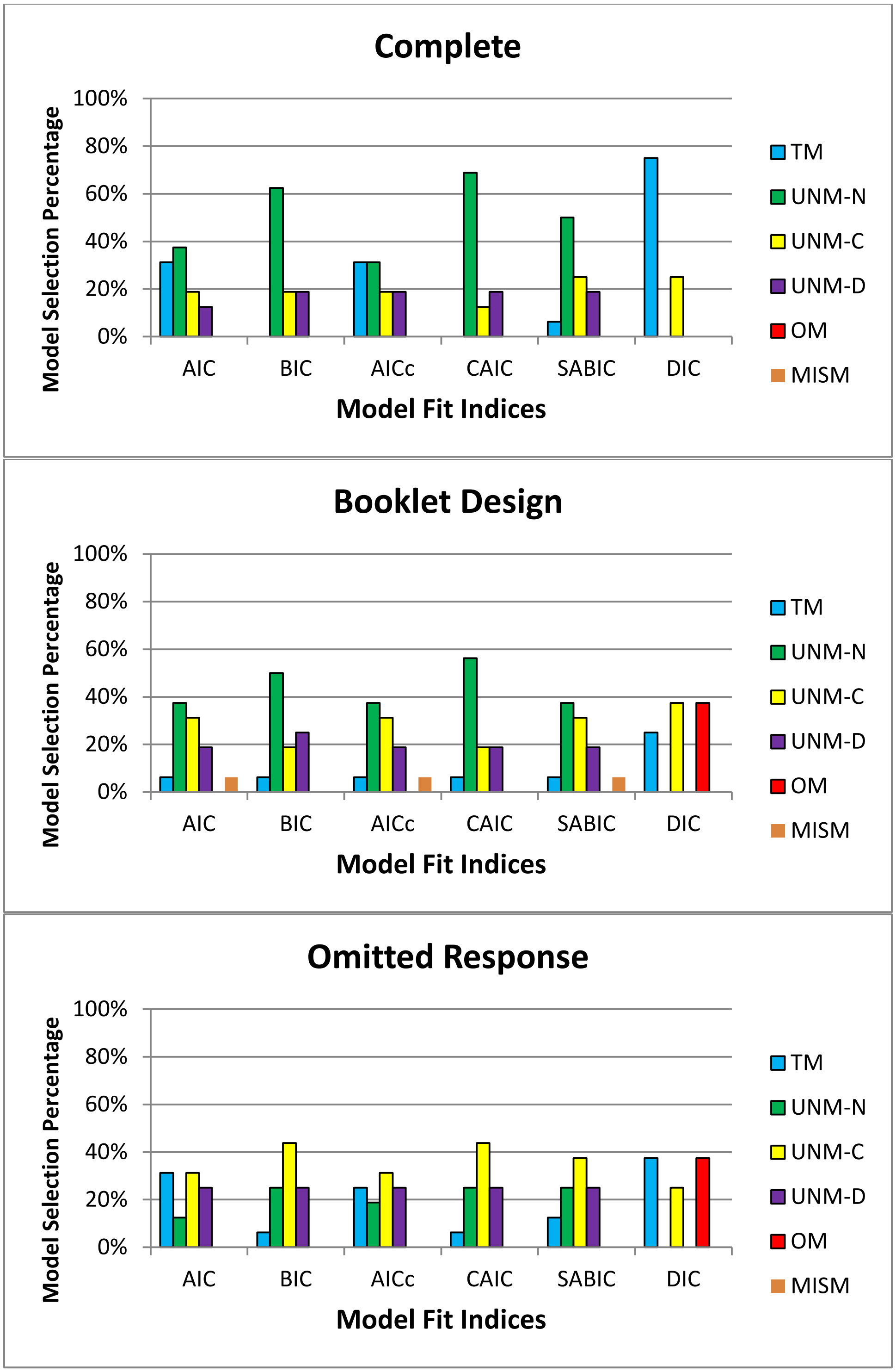

Figure 4.8b. Model selection percentage by data completeness across other manipulated factors.

As shown in Figure 4.8b, the ability of differentiating models was stronger for the six indices when the data was complete. In this scenario, BIC, CAIC and SABIC had a strong tendency of selecting the most parsimonious model; whereas AIC and AICc also had a weak tendency of choosing the MRM without covariates, yet they both had difficulty differentiating between the true model and the MRM without covariates. With regard to DIC, it was highly effective in selecting the data-generating model. As for the few cases that DIC chose the MRM with only the continuous covariate, they all occurred when the odds ratio was weak, so it was reasonable for DIC to select a more parsimonious model without the dichotomous covariate in those cases.

In the booklet design condition, BIC, CAIC and SABIC still tended to choose the MRM without covariates, yet the tendency was relatively weak. AIC and AICc in this condition both had difficulty differentiating between the MRM with only the continuous covariate and the MRM without covariates. Moreover, DIC was no longer effective when the book design was present. It could not make a decision between the overspecified model and the MRM with only the continuous covariate, and its selection percentages of these two models were only slightly higher than that of the true model.

In addition, for the omitted response condition, none of the indices had high percentages of selecting any models. BIC, CAIC and SABIC no longer chose the most parsimonious model; instead, they had a weak tendency of selecting the MRM with only the continuous covariate. AIC and AICc in this condition had difficulty choosing between the true model and the MRM with only the continuous covariate. As for DIC, it had the same selection percentage for the true model and the overspecified model, and its second choice was the MRM with only the continuous covariate.

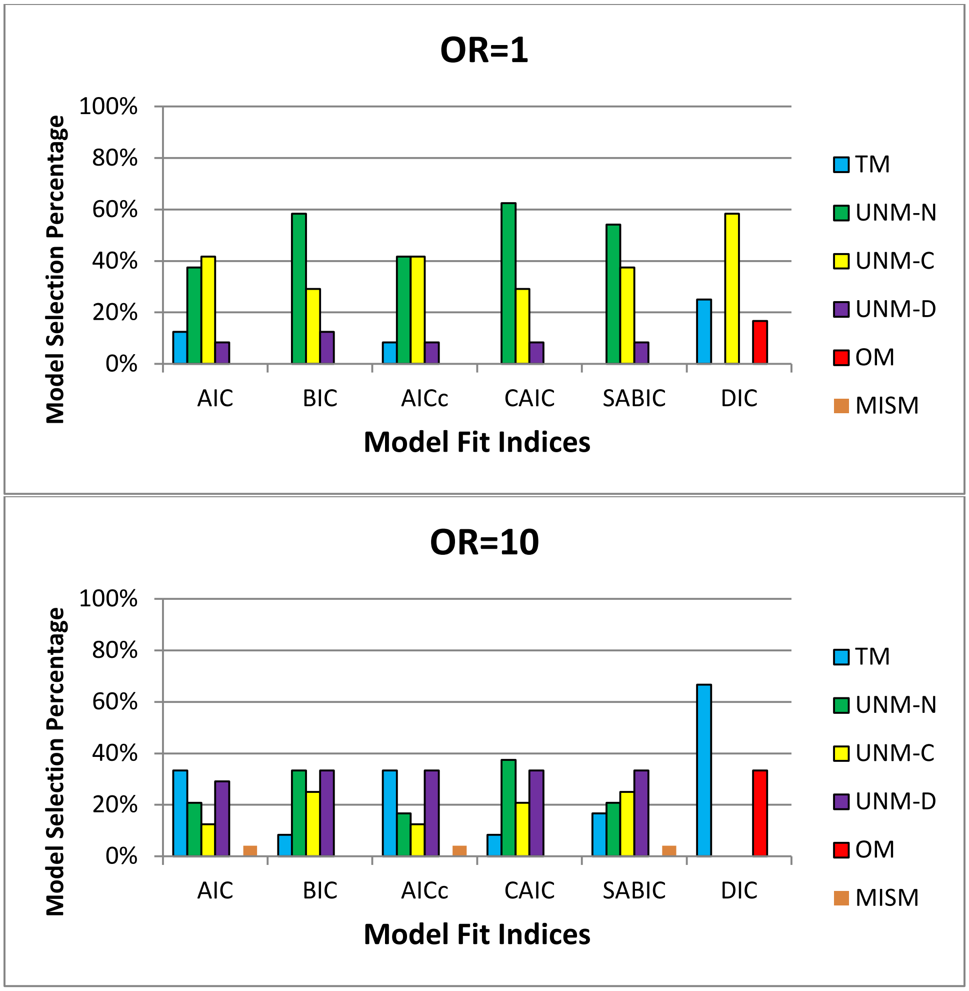

Figure 4.8c. Model selection percentage by the strength of relation between the dichotomous covariate and the latent class membership across other manipulated factors.

Further, Figure 4.8c presented the percentage of selection decision by odds ratio. When the odds ratio was weak, BIC, CAIC and SABIC most frequently selected the MRM without covariates, whereas AIC and AICc again had difficulty differentiating between the MRM with only the continuous covariate and the MRM without covariates. DIC predominantly identified the MRM with only the continuous covariate as the best-fitting model, which was a reasonable choice in consideration of model parsimony when OR=1. However, as the odds ratio was strong, the decisions were made most often between MRM with only the dichotomous covariate and the MRM without covariates for BIC, CAIC and SABIC. AIC and AICc similarly had difficulty discriminating between the true model and the MRM with only the dichotomous covariate. Again, DIC successfully identified the MRM with correct covariate specification with the highest selection percentage.

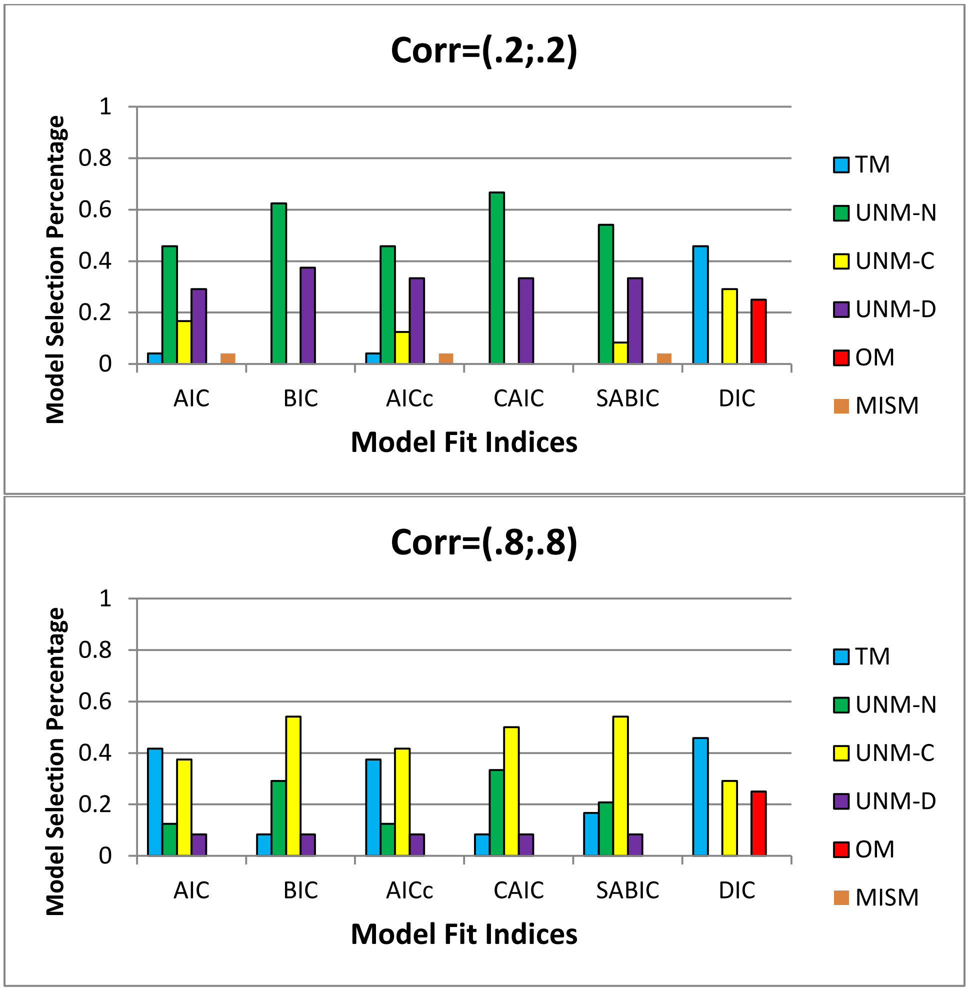

Figure 4.8d. Model selection percentage by the strength of relation between the continuous covariate and the person parameter across other manipulated factors.

Finally, with respect to the correlation between the continuous covariate and the person parameter, Figure 4.8d displayed the selection percentage. When the correlation was as weak as 0.200, AIC, BIC, AICc, CAIC and SABIC all favored the most parsimonious model, the MRM without covariates, and all of their second choices were the MRM with only the dichotomous covariate. Meanwhile, DIC correctly identified the true model as the best-fitting model. However, when the correlation was as strong as 0.800, BIC, CAIC and SABIC most frequently selected the MRM with only the continuous covariate, whereas AIC and AICc had difficulty choosing between the true model and the MRM with only the continuous covariate. Still, DIC predominantly made the correct model selection decision.

In sum, the above results suggested that DIC was the most successful index in selecting the MRM with correct covariate inclusion in the present study. The performances of BIC, CAIC and SABIC were quite similar and they had a consistent tendency of favoring model parsimony; whereas AIC and AICc often reached converging results and their decisions appeared to be highly sensitive to the strength of relations between the covariates and the model parameters. Another important finding about overall model fit indices was that their ability to differentiate among models tended to be strongly compromised when missing data were present. Also, the tendency of selecting the overspecified model occurred in particular for DIC in both the booklet design and omitted response conditions.

4.2 Empirical Examples: PISA 2009 U.S. Reading

The performance of different covariate inclusion approaches in the MRM was further demonstrated in real data applications. Two samples were extracted from the PISA 2009 U.S. students’ reading assessment. There were a total of 131 reading items in the test, distributed in 13 booklets. Among those items, 8 were polytomously scored, and were thus excluded from the current analyses. Sample 1 consisted of 1,525 students responding to 16 dichotomous items from booklet 2, 4, 5 and 7. No missing data were included in this sample. On the other hand, Sample 2 included 4,892 students responding to 123 dichotomous items. Missing data by booklet design were kept in Sample 2. The total percentage of missingness was 76.780%. In both samples, no missing data were present in the covariates. Section 3.4 has described sample information and covariate selection in details.

4.2.1 Sample 1

The results of the analyses based on Sample 1 were summarized in Table 4.15 and 4.16. Given the limited sample size and the number of items, the two-class MRM with both *reading enjoyment time* and *ESCS* as predictors of both the person parameter and the latent class membership and the two-class MRM with *reading enjoyment time* as the predictor of the person parameter and *ESCS* as the predictor of the latent class membership did not converge and thus were not presented here. Since the true values of model parameters were unknown, the accuracy of latent group classification and model parameter recovery could not be investigated as in the simulation study. Additionally, given that no good absolute model fit indices could be used in this context to quantify the discrepancy between the model and the data, only those overall relative model fit indices were provided in Table 4.15 to compare the relative fit of the non-mixture Rasch model and the two-class MRM with different approaches to covariate inclusion. Further, Table 4.16 presented the characteristics of the sample based on the parameter estimates from the two-class MRM with *ESCS* as the predictor of the person parameter and *reading enjoyment time* as the predictor of the latent class membership.

Table 4.15. *Model fit indices based on Sample 1*.

Model	AIC	BIC	AICc	CAIC	SABIC	DIC
RASCH	22556	22652	22557	22670	22595	23794
RASCH-C	22559	22660	22559	22679	22599	23757
RASCH-D	22558	22659	22558	22678	22599	23777
RASCH-CD	22562	22669	22563	22689	22605	23743
UNM-N	21906	22092	21907	22127	21981	23217
UNM-C	**21854**	**22051**	**21856**	**22088**	**21934**	**23130**
UNM-D	21996	22188	21998	22224	22074	23285
TM	21946	22149	21948	22187	22028	23217

Notes: Rasch: the Rasch model; Rasch-C: the Rasch model with *ESCS* as the predictor of the person parameter; Rasch-D: the Rasch model with *reading enjoyment time* as the predictor of the person parameter; Rasch-CD: the Rasch model with both *reading enjoyment time* and *ESCS* as predictors of the person parameter; UNM-N: the two-class MRM without covariates; UNM-C: the two-class MRM with *ESCS* as the predictor of the person parameter (UNM-C); UNM-D: the two-class MRM with *reading enjoyment time* as the predictor of the latent class membership; TM: the two-class MRM with ESCS as the predictor of the person parameter and *reading enjoyment time* as the predictor of the latent class membership.

Table 4.16. *Sample 1 characteristics based on parameter estimates*.

Sample Characteristics	Values
LC1	Mean=0.035; Variance=.963
LC2	Mean=2.208; Variance=.606
Prop	LC1: 0.775; LC2: 0.225
OR	*Reading Enjoyment Time*: OR=3.190
Corr	α_{11}=0.482; α_{12}=0.468

First, as shown in Table 4.15, all of the model fit indices supported two-class mixture model, rather than non-mixture model, in the sample. Also, all six model fit indices unanimously identified the two-class MRM with *ESCS* as the predictor of the person parameter, as the best-fitting model. It was surprising because these fit indices did not provide consistent result in the simulation study. However, as indicated by the simulation results, the model selection decision of AIC, BIC, AICc, CAIC and SABIC could be easily affected by the relative strength of the relations between the covariates and the model parameters. Moreover, DIC was proved in the simulation study to be the best index and frequently select the correct model in the complete data situation. Thus, for Sample1, it could be concluded that the two-class MRM with *ESCS* as the predictor of the person parameter was the best-fitting model based on the consistent results obtained from all six model fit indices.

Further, a closer examination of Sample 1was taken and the summary statistics were provided in Table 4.16. Based on the regression coefficient estimates from the two-class MRM with ESCS as the predictor of the person parameter and *reading enjoyment time* as the predictor of the latent class membership, it was found that *ESCS* had a moderate relation with the person parameter. The odds of being in latent class 1 for students whose reading enjoyment time was less than 30 minutes per day was 3.190 times that for students whose reading enjoyment time was more than 30 minutes per day, indicating a moderate relation between reading enjoyment time and latent class membership. Considering that the average latent reading ability was 0.035 for LC1 and 2.208 for LC2, the relations between the covariates and the model parameters were reasonable. Also, as *reading enjoyment time* was only a moderately informative dichotomous covariate, its odds ratio was not strong enough to make the dichotomous covariate necessary in the MRM. Thus, it is reasonable that the model with only *ESCS* as the predictor of the person parameter performed better than other models in terms of the overall fit.

4.2.2 Sample 2

The model selection results for Sample 2 were presented in Table 4.17. Similar to Sample 1, the two-class MRM with both *reading enjoyment time* and *ESCS* as predictors of both the person parameter and the latent class membership and the two-class MRM with *reading enjoyment time* as the predictor of the person parameter and *ESCS* as the predictor of the latent class membership again did not converge in Sample 2 and were not presented. Table 4.18 presented the characteristics of this sample based on the parameter estimates from the two-class MRM with *ESCS* as the predictor of the person parameter and *reading enjoyment time* as the predictor of the latent class membership.

Table 4.17. *Model fit indices based on Sample 2.*

Model	AIC	BIC	AICc	CAIC	SABIC	DIC
RASCH	130887	131699	130894	131824	131302	134972
RASCH-C	130857	131675	130864	131801	131275	134629
RASCH-D	130852	131670	130859	131796	131270	134608
RASCH-CD	130863	131688	130870	131815	131284	134730
UNM-N	**128907**	**130524**	**128934**	**130773**	**129733**	133054
UNM-C	129116	130746	129143	130997	129949	133917
UNM-D	128986	130610	129013	130860	129815	132639
TM	129104	130741	129131	130993	129940	**132631**

Notes: Rasch: the Rasch model; Rasch-C: the Rasch model with *ESCS* as the predictor of the person parameter; Rasch-D: the Rasch model with *reading enjoyment time* as the predictor of the person parameter; Rasch-CD: the Rasch model with both *reading enjoyment time* and *ESCS* as predictors of the person parameter; UNM-N: the two-class MRM without covariates; UNM-C: the two-class MRM with *ESCS* as the predictor of the person parameter (UNM-C); UNM-D: the two-class MRM with *reading enjoyment time* as the predictor of the latent class membership; TM: the two-class MRM with ESCS as the predictor of the person parameter and *reading enjoyment time* as the predictor of the latent class membership.

Table 4.18. *Sample 2 characteristics based on parameter estimates.*

Sample Characteristics	Values
LC1	Mean=-0.683; Variance=0.416
LC2	Mean=1.216; Variance=0.814
Prop.	LC1: 0.281; LC2: 0.709
OR	*Reading Enjoyment Time*: OR=3.040
Corr.	α_{11}=0.293; α_{12}=0.496

As shown in Table 4.17, all six model fit indices favored two-class mixture Rasch model rather than non-mixture Rasch model in the sample. In this sample, AIC, BIC, AICc, CAIC and SABIC unanimously identified the two-class MRM without covariates, as the best-fitting model, whereas DIC favored the two-class MRM with *reading enjoyment time* and *ESCS* included as covariates. These selection decisions were consistent with what has been observed in the simulation with the booklet design present; namely, DIC tended to choose the most complicated model whereas the other five indices selected the most parsimonious model. Thus, as with the simulation results, no conclusion could be reached for this sample with regard to the best-fitting model simply based on the overall model fit indices.

Further, the summary statistics of Sample 2 were provided in Table 4.18. The relations between the covariates and the model parameters did not differ much between the results based on Sample 1 and Sample 2. The average reading ability difference for the two latent classes was approximately 2 based on the estimates from both samples. Nonetheless, the only major difference between the two samples was the proportion of latent classes. In Sample 1, students who were proficient in reading accounted for about 0.225 of the sample; whereas in Sample 2, proficient students accounted for the majority (i.e., 0.709) of the sample.

Chapter 5 Discussion

Given the potentials of covariate inclusion in IRT models as suggested by literature, the present study explored different approaches to adding covariates into the MRM and the corresponding impacts on model estimation. In the simulation study, the relations between the covariates and the model parameters, DIF and mixing proportion were under manipulation. In addition, three types of data completeness, including complete data, booklet design and missing data by omitted responses, were simulated to approximate practical assessment settings. The effects of covariate inclusion approaches, as well as other manipulated factors, were compared and analyzed in terms of the accuracy of latent group classification, model parameter recovery, and overall model fit. The findings from the current study may shed light on future research and practices, and these findings are summarized and relevant implications are addressed in details in this chapter.

5.1 Discussion of the Simulation Results

In Chapter 4, results were summarized with respect to the accuracy of latent class assignment as indicated by correct classification rate, the recovery of item, person and regression coefficient parameters, and the model selection decisions based on AIC, BIC, AICc, CAIC, SABIC and DIC. In the discussion section, results will be discussed in three perspectives: 1) the impact of different approaches to covariate inclusion, which is the focus of the present study, 2) the effects of other manipulated factors, which are the add-on information about the estimation of mixture IRT models obtained from the current simulation, and 3) the implications regarding the effectiveness of different model selection indices.

5.1.1 Different approaches to covariate inclusion

The data generating model, with both the dichotomous and continuous covariates correctly specified, has the best performance in terms of the accuracy of latent class assignment, as compared with other models. It also has, on average, the smallest SE and RMSE in item parameter recovery, the smallest RMSE in person parameter recovery, the smallest RMSE in regression coefficient parameter recovery (except the slope parameter in the linear regression) and the highest 95% coverage rate for both item and person parameter recovery, although the effect size of model on these parameter recovery measures is not large enough to claim a practical significance.

Previous literature (Smit et al., 1999; 2000) suggested that the latent class assignment may substantially benefit from incorporating dichotomous covariates that are moderately or strongly associated with the latent class variable. In line with their results, the current study also witnesses a substantial increase in the correct classification rate if both dichotomous and continuous covariates are correctly specified in the MRM. Moreover, if only one covariate, dichotomous or continuous, is correctly specified in the MRM (i.e., UNM-D or UNM-C), there is also an improvement in the correct classification rate, but the MRM with only the dichotomous covariate performs slightly better than the MRM with only the continuous covariate. The reason might be that the dichotomous covariate enters the model directly as a predictor of the latent class membership in the UNM-D.

As for the parameter recovery, Mislevy and Sheehan (1989a; 1989b) suggested that the incorporation of covariates associated with the latent trait could compensate for the sparse information in the response data and hence reduce the mean squared error of person parameter estimates and the standard error of item parameter estimates in maximum likelihood estimation.

Later, Smit et al. (1999; 2000) in their study confirmed the results by showing a reduction of SE in the item parameter estimates, and Adams et al. (1997) showed a substantial reduction of mean squared error in the ability estimation. Similarly, in the present study, it appears that the correct covariate inclusion may lead to a reduction in the item parameter SE and RMSE, person parameter RMSE and an increase in the 95% coverage rate for both item and person parameter estimates. Although this pattern has been observed in the descriptive statistics, the model effect is not of practical significance for item or person parameter recovery, as shown in the repeated measures ANOVA results. A plausible explanation for the small effect size is the test length used in the current simulation. Previous studies all used very short tests with no more than 10 items (Mislevy & Sheehan, 1989a; 1989b; Smit et al., 1999; 2000) and indicated that the effects of covariate information on parameter recovery could diminish as test length increases. However, in the present study, in order to guarantee the convergence rate in the missing data scenarios, the sample size is set to be 2000 and the test length to be 30. This combination might be too ideal for the model parameter estimation of the MRM so that no additional information from covariates is necessary. This could be the major reason why the effect of model is not pronounced for parameter recovery in the ANOVA. Additionally, there is an interesting finding that the improvement in person parameter recovery may be exclusively due to the inclusion of the continuous covariate as a predictor of the person parameter, because the MRM with only the dichotomous covariate does not perform any better than the MRM without covariates in terms of the SE and RMSE of person parameter recovery. Thus, it is possible that the covariate information may function differentially in the model estimation and the benefits brought to the MRM may depend on the approach to covariate inclusion.

Further, regarding the regression coefficient parameters, the MRMs with only one covariate correctly specified (i.e., UNM-C and UNM-D) result in the least biased intercept estimates in each corresponding regression function, and the overspecified model leads to the least biased slope estimates in both the logistic and the linear functions. However, in term of the overall quality of regression parameter recovery as indicated by RMSE, the true model still performs better than the other models.

In summary, for the different approaches to covariate inclusion, the results in the present study show that the correct specification of both covariates in the MRM could potentially benefit the model performance in terms of the accuracy of latent group classification and the parameter estimation. Moreover, if only one of the covariates is correctly specified in the MRM, the model performance could still be improved to some extent, and the continuous covariate tends to influence both the latent group classification and the item and person parameter recovery whereas the dichotomous covariate could only improve the latent class identification. Further, based on all the model performance criteria mentioned above, it is found that the true model and the overspecified model are almost indistinguishable from each other, indicating that including redundant covariate information may not necessarily worsen the model performance as long as all the necessary covariates are correctly specified in the model. However, the MRM with mismatching covariates results in the worst model performance in terms of most of the criteria considered in the present study, implying that the mismatch between covariates and model parameters may lead to even worse results than not including any covariates in the model.

5.1.2 Effects of the other manipulated factors

Among the other manipulated factors, DIF, mixing proportion, data completeness, and their interactions tend to strongly impact the accuracy of latent class assignment, as well as item and person parameter recovery. As mixing proportion and DIF have been extensively studied in mixture IRT literature, they are not discussed in details here. Regarding data completeness, the booklet design tends to lead to the worst result in terms of most of the evaluation criteria used in the present study, with the exception that the omitted response condition results in the largest bias and the lowest 95% coverage for the person parameter estimates. The poor performance of booklet design is within expectation, considering the largest amount of missing data involved; however, the even worse performance of omitted response in two person parameter outcome measures is surprising, and one possible reason for that could be the conditional missing data mechanism involved in the omitted response.

Further, as shown in Chapter 4, data completeness tends to frequently interact with other factors in impacting the outcome measures, and the key to the interactions is the booklet design. The effects of model, mixing proportion or DIF, and certain interaction effects, tend to be much stronger particularly in the booklet design. For example, in terms of latent class assignment, the model effect is more pronounced in the booklet design, with the MRM without covariates and the MRM with mismatching covariates performing even worse in this situation. Another example is the bias for person parameter estimates. When the mixing proportion is equal or the DIF size is small, there is on average a slight positive bias in the person parameter estimates; whereas when the mixing proportion is unequal or the DIF size is large, there tends to be a larger negative bias in the person parameter estimates. However, for the complete data and omitted response conditions, the bias is largely unaffected by those factors. Similar results regarding other measures are also found with the booklet design as presented in Chapter 4. These results herein imply that the quality of mixture IRT model estimation may be easily influenced by different factors if booklet design is implemented in the assessment instrument. It may be worth considering further simulation study to fully reveal the impact of different types and amounts of missing data, especially booklet design, on the estimation of mixture IRT models.

5.1.3 Model selection

One aspect of the present study is to provide information about model fit and model selection with respect to covariate inclusion, which has not been discussed by other studies in this line of research. Previous research regarding model fit in the mixture IRT modeling context without covariates (Li et al., 2009) recommended the use of BIC because of its outstanding performance and consistency in detecting latent class enumeration. It was suggested that both AIC and DIC had a tendency to select the most complex model (Li et al., 2009). However, different from the previous study, the current simulation provides unique information about the effectiveness of overall model fit indices in the mixture IRT modeling context with covariate inclusion.

In general, among the six indices reported in the study, DIC is the most effective one in identifying the correct covariate inclusion in the MRM. Moreover, it is also very sensitive to the relation between the dichotomous covariate and the latent class membership. It could identify the more parsimonious option when the dichotomous covariate is not necessary in the model. Regarding the other five indices, they are not found to be useful in the current study, yet it interesting to find that AIC and AICc are highly consistent with each other, and BIC, CAIC and SABIC tend to provide very similar results. This pattern has also been observed in a previous study (Zhu, 2013). In the current simulation, BIC, CAIC and SABIC have a very strong tendency to select the most parsimonious model, whereas AIC and AICc have great difficulty differentiating among models. Thus, different from the message provided by previous research that AIC tends to select more complex model, the current simulation indicates that AIC and

AICc are highly inconsistent in the MRM context when covariates are involved. These two indices may not be good choices for practitioners when mixture IRT models are used. Furthermore, although BIC is proved to be successful in selecting the best latent structure and choosing among 1-, 2- and 3-parameter logistic IRT models, this index may not be sensitive to the fit of covariate inclusion in the mixture IRT models. The reason could be that including covariates and having more complex latent structure respectively complicate the model in different perspectives. Thus, the use of BIC should be implemented with caution as it works well in some contexts but not the others.

However, in most commonly-used commercial software programs for mixture IRT model parameter estimation, the use of AIC and BIC is prevalent, and other model fit indices are usually not provided. For example, in *M*plus, only AIC, BIC and SABIC are available for mixture IRT model estimation. DIC is only implemented in the Bayesian module for two-level models in the most recent version of *M*plus (i.e., version 7.3), and it is not applied in the estimation of mixture models. The overemphasis of BIC in the literature and the ignorance of other model fit indices in the commercial software programs may lead to misfitting models being selected as better-fitting models for practitioners. Thus, it is suggested that the calculation of more model fit indices may be implemented in software programs and DIC could also be included if a Bayesian module exists, so that researchers may choose which index to use, depending on the purpose of the study, the data structure and the effect size of model parameters. In addition, the present study is based on Bayesian estimation and DIC is a model fit index specifically designed for Bayesian posterior estimates of model parameters. Thus, the use of DIC in the current simulation may not be directly applied to maximum likelihood estimation context.

Another important finding in the present study regarding model selection is that the effectiveness of all six indices is highly sensitive to the missing data. Even for DIC, its performance is greatly compromised when missing data are present. To be specific, DIC shows a tendency of selecting the most complicated model no matter the missing data come from omitted responses or booklet design. The other indices also have great difficulty in differentiating the true model and the three underspecified models in missing data conditions. The results of real data applications also confirm the findings in the simulations. Therefore, one important suggestion to come out of this study for practitioners is to be extremely cautious about model selection indices when using them with missing data present. As the effectiveness of model fit indices is sometimes model and design specific and sometimes compromised by missing data, it is recommended that researchers should evaluate the data set and the models from different perspectives, rather than solely relying on information-based fit indices to choose among models.

5.2 Applications of Covariate Inclusion

As for the applications of covariate inclusion approaches, it is hoped that the mixture IRT model with covariates correctly specified may help identify latent DIF, explain latent DIF using manifest grouping variables (e.g., dichotomous covariate), and improve model parameter estimation simultaneously. Previously, covariate inclusion was proved useful in non-mixture IRT models for the purpose of explaining estimated effects (e.g., Wilson & De Boeck, 2004) or improving model parameter estimation (e.g., Adams et al., 1997). The current study incorporates covariates into the MRM via different approaches, and extends the use of covariate information to a broader scenario.

Purely in the perspective of model estimation, covariate inclusion is promising for mixture IRT models with the potential benefit of improving the latent group classification and the estimation of model parameters. However, regarding the real data applications, there exists a theoretical debate with respect to the validity of inference drawn about the population if covariate information is used, because covariate inclusion violates the fundamental of equitable measurement and test fairness; namely, the parameter estimation should be independent of any variables beyond the response data per se (Adams et al., 1997). Thus, as mentioned earlier, it is desirable to use covariates to improve the precision of model parameter estimation, yet it is less desirable to draw inference based on the conditional model, especially when high-stake decisions are involved (Mislevy & Sheehan, 1989a).

Additionally, one important methodology, which is closely related to the covariate inclusion approach and also commonly used in large-scale assessment, is the plausible value imputation method. Plausible values are imputed values drawn from an empirically derived distribution of latent achievement scores that are conditional on the observed values of items responses and respondents' background variables (i.e., covariates). For an in-depth description of the plausible value imputation method, one can review some recent research work by Mislevy (1991; 1993), Rubin (1987), von Davier et al. (2009) and Wu (2005). As mentioned in Adams et al. (1997), to draw plausible values, NAEP uses an approach very similar to two-step estimation with covariates. Item parameters are estimated first without the covariates and the item parameters are fixed in the second phase for the generation of plausible values to better approximate population parameters (Adams et al., 1997). This methodology could be taken as an important extension and practical application based on covariate inclusion approaches.

5.3 Limitations and Future Direction

As with all other studies, certain limitations remain in the present study. First, considering the amount of time required for the model estimation under the Bayesian framework, a number of factors are fixed in the simulation design, so the results are limited to the manipulated factors under investigation. Future research is necessary to further the findings by adding more simulation factors or including more levels based on the current factors, especially for test length, sample size and data completeness. As discussed in the previous section, the test length may be the major reason why the model effect on the parameter recovery is not of practical significance. Also, growth mixture literature (Kohli et al., 2013; Li et al., 2014) suggested that sample size, separations between latent classes and the interaction between the two may affect model parameter estimation and latent class identification. As sample size is not manipulated in the present study, it is not known whether this finding also holds in mixture IRT context. Simulations that simultaneously consider test length, sample size and separation between latent classes could be conducted to further this line of research. Additionally, as for the data completeness, the current study implies that both the amount and the mechanism of missing data (i.e., random or conditional) may impact the parameter recovery and the performance of model fit indices. However, with only one level for each type of missing data, the effects of missing data amount and mechanism are confounded. This is also an issue that is worth further investigation. Second, for the regression parameters, the current study using one-step estimation shows that the linear regression parameters tend to be slightly overestimated and the logistic regression parameters underestimated. Previous research suggested that one-step estimation is favored than two-step estimation due to the fact that the latter one would greatly underestimate the regression parameters (Adam et al., 1997). However, without a direct comparison of one-step versus two-step estimation in the present study, it is not clear that how much one-step estimation is better than two-step estimation in terms of recovering the relations between covariates and model parameters in the MRM context. This issue may be explored in future research. Finally, although Bayesian estimation allows much flexibility in model specification, it requires substantial computing time so that the scope of the simulation and the number of replications per cell are limited in the present study. To better understand the impact of covariate inclusion on mixture IRT models, future research may consider including more replications and broadening the scope of the current simulation by using maximum likelihood method to achieve greater estimation efficiency.

In summary, despite the limitations, the findings from this study definitely add to the literature about different covariate inclusion approaches in mixture IRT modeling. With an ever-increasing use of complicated models in psychometrics, a proper use of covariate information is of theoretical and practical importance for researchers to achieve more accurate model estimation. With a simulation study followed by real data examples, the current study provides important evidence about the impact of covariate inclusion on the accuracy of latent group classification, model parameter recovery, and overall model fit. Empirical information based on real data about the appropriateness of covariate inclusion in practical assessment settings is also included. It complements similar previous studies and lays a good foundation for future explorations.

Appendix A

Table 1

Two Sets of Generating Item Difficulty Parameters b_{i1} and b_{i2}.

Item	Average DIF=1.5			Average DIF=1.0		
	b_{i1}	b_{i2}	$\|b_{i1}-b_{i2}\|$	b_{i1}	b_{i2}	$\|b_{i1}-b_{i2}\|$
1	0.384	-1.838	2.221	0.933	0.396	0.538
2	1.367	-0.154	1.521	-0.525	0.077	0.602
3	-0.345	0.980	1.325	1.814	2.400	0.586
4	1.349	0.403	0.946	0.083	-0.894	0.977
5	0.303	-0.926	1.229	0.396	0.766	0.370
6	0.521	-0.624	1.145	-2.194	1.559	3.753
7	1.143	-0.257	1.401	-0.360	1.132	1.493
8	0.216	-1.367	1.583	0.143	-0.251	0.394
9	1.130	-1.887	3.018	-0.204	-1.107	0.903
10	-0.600	0.342	0.942	0.446	1.185	0.739
11	-0.146	0.685	0.830	-0.322	-1.728	1.406
12	1.420	-0.412	1.832	0.478	0.159	0.320
13	-0.523	0.690	1.213	0.196	-0.669	0.865
14	0.314	-0.327	0.641	0.715	0.973	0.258
15	0.924	-0.586	1.510	-0.960	-1.231	0.270
16	0.580	2.479	1.899	0.671	0.730	0.060
17	-0.655	1.036	1.690	1.655	-0.391	2.046
18	-0.254	-1.392	1.138	1.243	-0.811	2.053
19	0.124	-1.560	1.684	-1.562	-0.481	1.080
20	0.308	1.507	1.199	1.182	1.524	0.342
21	-1.883	0.119	2.002	0.570	-0.663	1.233
22	-0.409	1.285	1.694	1.141	0.729	0.412
23	0.463	-0.764	1.227	-1.241	0.943	2.184
24	-0.879	0.584	1.463	0.684	0.361	0.323
25	-1.619	0.180	1.799	-0.523	1.060	1.583
26	0.042	0.709	0.667	0.362	0.740	0.377
27	0.797	-0.485	1.283	-1.840	-3.606	1.766
28	0.088	0.584	0.496	0.562	0.869	0.307
29	-1.011	0.433	1.444	0.891	-1.315	2.206
30	-3.148	0.565	3.713	-4.436	-2.456	1.979

Table 2. *Simulation conditions in the present study*.

Condition	Data Completeness	OR	Corr.	Prop.	DIF
1	complete	10	.2;.2	.5;.5	1.5
2	complete	10	.2;.2	.3;.7	1.5
3	complete	10	.8;.8	.5;.5	1.5
4	complete	10	.8;.8	.3;.7	1.5
5	complete	1	.2;.2	.5;.5	1.5
6	complete	1	.2;.2	.3;.7	1.5
7	complete	1	.8;.8	.5;.5	1.5
8	complete	1	.8;.8	.3;.7	1.5
9	complete	10	.2;.2	.5;.5	1
10	complete	10	.2;.2	.3;.7	1
11	complete	10	.8;.8	.5;.5	1
12	complete	10	.8;.8	.3;.7	1
13	complete	1	.2;.2	.5;.5	1
14	complete	1	.2;.2	.3;.7	1
15	complete	1	.8;.8	.5;.5	1
16	complete	1	.8;.8	.3;.7	1
17	booklet design	10	.2;.2	.5;.5	1.5
18	booklet design	10	.2;.2	.3;.7	1.5
19	booklet design	10	.8;.8	.5;.5	1.5
20	booklet design	10	.8;.8	.3;.7	1.5
21	booklet design	1	.2;.2	.5;.5	1.5
22	booklet design	1	.2;.2	.3;.7	1.5
23	booklet design	1	.8;.8	.5;.5	1.5
24	booklet design	1	.8;.8	.3;.7	1.5
25	booklet design	10	.2;.2	.5;.5	1
26	booklet design	10	.2;.2	.3;.7	1
27	booklet design	10	.8;.8	.5;.5	1
28	booklet design	10	.8;.8	.3;.7	1
29	booklet design	1	.2;.2	.5;.5	1
30	booklet design	1	.2;.2	.3;.7	1
31	booklet design	1	.8;.8	.5;.5	1
32	booklet design	1	.8;.8	.3;.7	1
33	omitted responses	10	.2;.2	.5;.5	1.5
34	omitted responses	10	.2;.2	.3;.7	1.5
35	omitted responses	10	.8;.8	.5;.5	1.5
36	omitted responses	10	.8;.8	.3;.7	1.5
37	omitted responses	1	.2;.2	.5;.5	1.5
38	omitted responses	1	.2;.2	.3;.7	1.5
39	omitted responses	1	.8;.8	.5;.5	1.5

40	omitted responses	1	.8;.8	.3;.7	1.5
41	omitted responses	10	.2;.2	.5;.5	1
42	omitted responses	10	.2;.2	.3;.7	1
43	omitted responses	10	.8;.8	.5;.5	1
44	omitted responses	10	.8;.8	.3;.7	1
45	omitted responses	1	.2;.2	.5;.5	1
46	omitted responses	1	.2;.2	.3;.7	1
47	omitted responses	1	.8;.8	.5;.5	1
48	omitted responses	1	.8;.8	.3;.7	1

Note: Data generated in each condition are estimated by 6 models.

Table 3. *Average correct latent group classification rate*.

					Estimation Model					
Data	DIF	OR	Corr.	Prop.	TM	UNM-N	UNM-C	UNM-D	OM	MISM
Complete	1.5	10	.2;.2	.5;.5	0.974	0.971	0.970	0.974	0.975	0.970
				.3;.7	0.977	0.973	0.972	0.976	0.976	0.972
			.8;.8	.5;.5	0.978	0.970	0.974	0.974	0.978	0.969
				.3;.7	0.980	0.973	0.976	0.977	0.980	0.972
		1	.2;.2	.5;.5	0.970	0.971	0.970	0.971	0.970	0.970
				.3;.7	0.972	0.972	0.972	0.972	0.972	0.972
			.8;.8	.5;.5	0.974	0.970	0.974	0.969	0.974	0.969
				.3;.7	0.976	0.973	0.976	0.973	0.976	0.972
	1	10	.2;.2	.5;.5	0.951	0.942	0.942	0.951	0.951	0.941
				.3;.7	0.955	0.946	0.945	0.955	0.955	0.944
			.8;.8	.5;.5	0.956	0.940	0.949	0.949	0.956	0.936
				.3;.7	0.961	0.948	0.953	0.956	0.961	0.942
		1	.2;.2	.5;.5	0.943	0.942	0.942	0.942	0.942	0.942
				.3;.7	0.945	0.946	0.945	0.946	0.946	0.945
			.8;.8	.5;.5	0.949	0.940	0.949	0.941	0.949	0.936
				.3;.7	0.952	0.948	0.953	0.947	0.953	0.941
Booklet Design	1.5	10	.2;.2	.5;.5	0.937	0.929	0.930	0.937	0.936	0.923
				.3;.7	0.929	0.912	0.911	0.928	0.927	0.895
			.8;.8	.5;.5	0.941	0.927	0.940	0.935	0.942	0.916
				.3;.7	0.940	0.915	0.929	0.930	0.939	0.877
		1	.2;.2	.5;.5	0.929	0.929	0.930	0.911	0.930	0.928
				.3;.7	0.911	0.912	0.911	0.910	0.910	0.912
			.8;.8	.5;.5	0.939	0.927	0.940	0.928	0.939	0.921
				.3;.7	0.928	0.915	0.929	0.914	0.928	0.887
	1	10	.2;.2	.5;.5	0.913	0.888	0.888	0.912	0.913	0.884
				.3;.7	0.890	0.825	0.836	0.895	0.887	0.797
			.8;.8	.5;.5	0.923	0.888	0.901	0.911	0.923	0.786
				.3;.7	0.906	0.832	0.867	0.897	0.905	0.667
		1	.2;.2	.5;.5	0.887	0.888	0.888	0.886	0.884	0.881
				.3;.7	0.823	0.825	0.837	0.832	0.824	0.832
			.8;.8	.5;.5	0.899	0.887	0.901	0.885	0.897	0.766
				.3;.7	0.858	0.832	0.867	0.837	0.859	0.693
Omitted Responses	1.5	10	.2;.2	.5;.5	0.938	0.926	0.927	0.938	0.937	0.924
				.3;.7	0.930	0.914	0.914	0.930	0.930	0.912
			.8;.8	.5;.5	0.948	0.925	0.938	0.938	0.947	0.922
				.3;.7	0.943	0.915	0.930	0.933	0.942	0.913
		1	.2;.2	.5;.5	0.926	0.926	0.927	0.926	0.926	0.926
				.3;.7	0.914	0.914	0.914	0.914	0.914	0.914
			.8;.8	.5;.5	0.938	0.925	0.938	0.925	0.935	0.925
				.3;.7	0.930	0.915	0.930	0.915	0.929	0.915

1	10	.2;.2	.5;.5	0.936	0.922	0.923	0.935	0.935	0.921
			.3;.7	0.938	0.920	0.920	0.937	0.938	0.917
		.8;.8	.5;.5	0.945	0.920	0.934	0.932	0.946	0.915
			.3;.7	0.946	0.923	0.935	0.939	0.947	0.909
	1	.2;.2	.5;.5	0.923	0.922	0.923	0.922	0.923	0.922
			.3;.7	0.920	0.920	0.920	0.920	0.920	0.920
		.8;.8	.5;.5	0.935	0.920	0.934	0.920	0.934	0.915
			.3;.7	0.935	0.923	0.935	0.923	0.935	0.911

Table 4a. *Average standard error of item parameters.*

					Estimation Model					
Data	DIF	OR	Corr.	Prop.	TM	UNM-N	UNM-C	UNM-D	OM	MISM
Complete	1.5	10	.2;.2	.5;.5	0.073	0.074	0.074	0.073	0.073	0.074
				.3;.7	0.076	0.077	0.077	0.076	0.076	0.077
			.8;.8	.5;.5	0.073	0.074	0.074	0.074	0.073	0.074
				.3;.7	0.078	0.079	0.079	0.079	0.078	0.079
		1	.2;.2	.5;.5	0.074	0.074	0.074	0.074	0.074	0.074
				.3;.7	0.077	0.077	0.077	0.077	0.077	0.077
			.8;.8	.5;.5	0.074	0.074	0.074	0.074	0.074	0.074
				.3;.7	0.079	0.079	0.079	0.079	0.079	0.079
	1	10	.2;.2	.5;.5	0.065	0.066	0.066	0.065	0.065	0.066
				.3;.7	0.070	0.071	0.072	0.070	0.070	0.071
			.8;.8	.5;.5	0.065	0.066	0.065	0.066	0.065	0.066
				.3;.7	0.070	0.071	0.071	0.070	0.070	0.071
		1	.2;.2	.5;.5	0.066	0.066	0.066	0.066	0.066	0.066
				.3;.7	0.072	0.071	0.072	0.071	0.072	0.071
			.8;.8	.5;.5	0.066	0.066	0.065	0.065	0.066	0.066
				.3;.7	0.071	0.071	0.071	0.071	0.071	0.071
Booklet Design	1.5	10	.2;.2	.5;.5	0.106	0.107	0.107	0.106	0.106	0.110
				.3;.7	0.160	0.155	0.154	0.160	0.158	0.152
			.8;.8	.5;.5	0.105	0.109	0.104	0.108	0.106	0.113
				.3;.7	0.158	0.160	0.151	0.166	0.158	0.159
		1	.2;.2	.5;.5	0.107	0.107	0.107	0.193	0.107	0.108
				.3;.7	0.154	0.155	0.154	0.155	0.155	0.156
			.8;.8	.5;.5	0.104	0.109	0.104	0.109	0.105	0.111
				.3;.7	0.151	0.160	0.151	0.160	0.150	0.162
	1	10	.2;.2	.5;.5	0.099	0.101	0.102	0.099	0.099	0.102
				.3;.7	0.102	0.101	0.100	0.104	0.102	0.102
			.8;.8	.5;.5	0.095	0.105	0.098	0.105	0.098	0.160
				.3;.7	0.101	0.099	0.098	0.104	0.103	0.111
		1	.2;.2	.5;.5	0.101	0.101	0.102	0.103	0.104	0.105
				.3;.7	0.101	0.101	0.100	0.101	0.101	0.102
			.8;.8	.5;.5	0.099	0.105	0.097	0.108	0.102	0.166
				.3;.7	0.095	0.099	0.098	0.100	0.096	0.109
Omitted Responses	1.5	10	.2;.2	.5;.5	0.080	0.080	0.080	0.080	0.080	0.080
				.3;.7	0.090	0.091	0.092	0.089	0.090	0.092
			.8;.8	.5;.5	0.079	0.081	0.079	0.080	0.079	0.081
				.3;.7	0.091	0.095	0.093	0.092	0.090	0.096
		1	.2;.2	.5;.5	0.080	0.080	0.080	0.080	0.080	0.080
				.3;.7	0.092	0.091	0.092	0.091	0.091	0.091
			.8;.8	.5;.5	0.080	0.081	0.079	0.081	0.080	0.081
				.3;.7	0.093	0.095	0.093	0.094	0.093	0.096

1	10	.2;.2	.5;.5	0.072	0.073	0.073	0.072	0.072	0.074
			.3;.7	0.082	0.084	0.083	0.082	0.082	0.083
		.8;.8	.5;.5	0.070	0.073	0.071	0.072	0.070	0.074
			.3;.7	0.079	0.081	0.080	0.080	0.079	0.084
	1	.2;.2	.5;.5	0.074	0.073	0.073	0.073	0.074	0.074
			.3;.7	0.084	0.084	0.083	0.084	0.084	0.083
		.8;.8	.5;.5	0.071	0.073	0.071	0.073	0.072	0.074
			.3;.7	0.080	0.081	0.080	0.081	0.080	0.083

Table 4b. *Average RMSE of item parameters.*

					Estimation Model					
Data	DIF	OR	Corr.	Prop.	TM	UNM-N	UNM-C	UNM-D	OM	MISM
Complete	1.5	10	.2;.2	.5;.5	0.074	0.075	0.075	0.074	0.074	0.075
				.3;.7	0.077	0.078	0.078	0.077	0.077	0.078
			.8;.8	.5;.5	0.074	0.075	0.074	0.074	0.074	0.075
				.3;.7	0.079	0.080	0.080	0.079	0.079	0.080
		1	.2;.2	.5;.5	0.075	0.075	0.075	0.075	0.075	0.075
				.3;.7	0.078	0.078	0.078	0.078	0.078	0.078
			.8;.8	.5;.5	0.074	0.075	0.074	0.075	0.074	0.075
				.3;.7	0.080	0.080	0.080	0.080	0.079	0.080
	1	10	.2;.2	.5;.5	0.066	0.066	0.067	0.066	0.066	0.066
				.3;.7	0.071	0.072	0.073	0.071	0.071	0.073
			.8;.8	.5;.5	0.065	0.066	0.066	0.065	0.065	0.066
				.3;.7	0.071	0.071	0.072	0.070	0.070	0.073
		1	.2;.2	.5;.5	0.066	0.066	0.067	0.067	0.067	0.066
				.3;.7	0.072	0.072	0.073	0.072	0.072	0.072
			.8;.8	.5;.5	0.066	0.066	0.066	0.066	0.066	0.066
				.3;.7	0.071	0.071	0.072	0.071	0.071	0.073
Booklet Design	1.5	10	.2;.2	.5;.5	0.119	0.128	0.126	0.120	0.120	0.136
				.3;.7	0.447	0.480	0.479	0.446	0.446	0.509
			.8;.8	.5;.5	0.119	0.131	0.117	0.124	0.119	0.146
				.3;.7	0.389	0.481	0.402	0.446	0.390	0.552
		1	.2;.2	.5;.5	0.127	0.128	0.126	0.205	0.126	0.128
				.3;.7	0.477	0.480	0.479	0.477	0.473	0.483
			.8;.8	.5;.5	0.118	0.131	0.117	0.131	0.118	0.139
				.3;.7	0.402	0.481	0.402	0.478	0.402	0.534
	1	10	.2;.2	.5;.5	0.124	0.137	0.138	0.123	0.124	0.137
				.3;.7	0.247	0.266	0.270	0.247	0.248	0.272
			.8;.8	.5;.5	0.117	0.138	0.130	0.128	0.121	0.233
				.3;.7	0.247	0.263	0.270	0.246	0.247	0.377
		1	.2;.2	.5;.5	0.138	0.137	0.138	0.139	0.148	0.149
				.3;.7	0.266	0.266	0.271	0.265	0.266	0.264
			.8;.8	.5;.5	0.134	0.138	0.130	0.140	0.143	0.258
				.3;.7	0.261	0.263	0.270	0.263	0.262	0.352
Omitted Responses	1.5	10	.2;.2	.5;.5	0.103	0.110	0.110	0.104	0.104	0.112
				.3;.7	0.125	0.138	0.139	0.125	0.125	0.139
			.8;.8	.5;.5	0.097	0.110	0.103	0.103	0.098	0.112
				.3;.7	0.120	0.142	0.132	0.128	0.120	0.142
		1	.2;.2	.5;.5	0.110	0.110	0.110	0.110	0.110	0.110
				.3;.7	0.139	0.138	0.139	0.138	0.137	0.137
			.8;.8	.5;.5	0.103	0.110	0.103	0.110	0.104	0.110
				.3;.7	0.133	0.142	0.132	0.142	0.133	0.140

1	10	.2;.2	.5;.5	0.081	0.083	0.085	0.081	0.081	0.084
			.3;.7	0.094	0.098	0.099	0.094	0.094	0.098
		.8;.8	.5;.5	0.076	0.082	0.079	0.080	0.076	0.083
			.3;.7	0.088	0.095	0.092	0.093	0.087	0.098
	1	.2;.2	.5;.5	0.083	0.083	0.085	0.083	0.083	0.083
			.3;.7	0.098	0.098	0.099	0.098	0.098	0.097
		.8;.8	.5;.5	0.078	0.082	0.079	0.082	0.078	0.082
			.3;.7	0.090	0.095	0.092	0.095	0.089	0.097

Table 4c. *Average 95% coverage rate of item parameters.*

					Estimation Model					
Data	DIF	OR	Corr.	Prop.	TM	UNM-N	UNM-C	UNM-D	OM	MISM
Complete	1.5	10	.2;.2	.5;.5	0.952	0.953	0.954	0.953	0.951	0.954
				.3;.7	0.963	0.965	0.961	0.962	0.961	0.964
			.8;.8	.5;.5	0.953	0.952	0.951	0.955	0.951	0.953
				.3;.7	0.956	0.954	0.953	0.956	0.957	0.955
		1	.2;.2	.5;.5	0.952	0.953	0.954	0.951	0.951	0.951
				.3;.7	0.964	0.965	0.961	0.963	0.965	0.965
			.8;.8	.5;.5	0.953	0.952	0.951	0.951	0.950	0.951
				.3;.7	0.955	0.954	0.953	0.955	0.957	0.955
	1	10	.2;.2	.5;.5	0.947	0.944	0.945	0.949	0.947	0.947
				.3;.7	0.949	0.949	0.953	0.950	0.951	0.945
			.8;.8	.5;.5	0.954	0.952	0.952	0.957	0.955	0.953
				.3;.7	0.953	0.954	0.953	0.951	0.957	0.947
		1	.2;.2	.5;.5	0.946	0.944	0.945	0.944	0.945	0.948
				.3;.7	0.947	0.949	0.953	0.948	0.950	0.949
			.8;.8	.5;.5	0.953	0.952	0.952	0.951	0.951	0.951
				.3;.7	0.953	0.954	0.953	0.953	0.953	0.948
Booklet Design	1.5	10	.2;.2	.5;.5	0.921	0.871	0.887	0.918	0.915	0.849
				.3;.7	0.561	0.527	0.527	0.563	0.557	0.493
			.8;.8	.5;.5	0.907	0.865	0.914	0.901	0.915	0.822
				.3;.7	0.665	0.535	0.624	0.553	0.660	0.499
		1	.2;.2	.5;.5	0.877	0.871	0.887	0.890	0.892	0.872
				.3;.7	0.528	0.527	0.527	0.524	0.529	0.525
			.8;.8	.5;.5	0.911	0.865	0.914	0.877	0.915	0.842
				.3;.7	0.627	0.535	0.624	0.529	0.619	0.511
	1	10	.2;.2	.5;.5	0.884	0.839	0.831	0.885	0.882	0.837
				.3;.7	0.794	0.732	0.750	0.807	0.795	0.712
			.8;.8	.5;.5	0.879	0.842	0.836	0.867	0.873	0.659
				.3;.7	0.771	0.741	0.712	0.811	0.779	0.477
		1	.2;.2	.5;.5	0.835	0.839	0.831	0.833	0.801	0.794
				.3;.7	0.733	0.732	0.751	0.739	0.728	0.740
			.8;.8	.5;.5	0.837	0.845	0.843	0.835	0.808	0.591
				.3;.7	0.721	0.741	0.712	0.750	0.721	0.519
Omitted Responses	1.5	10	.2;.2	.5;.5	0.865	0.849	0.851	0.861	0.866	0.845
				.3;.7	0.853	0.839	0.839	0.851	0.853	0.837
			.8;.8	.5;.5	0.899	0.866	0.892	0.878	0.903	0.865
				.3;.7	0.861	0.829	0.856	0.843	0.867	0.829
		1	.2;.2	.5;.5	0.849	0.849	0.851	0.847	0.848	0.852
				.3;.7	0.837	0.839	0.839	0.833	0.836	0.838
			.8;.8	.5;.5	0.888	0.866	0.892	0.871	0.889	0.865
				.3;.7	0.853	0.829	0.856	0.834	0.855	0.825

1	10	.2;.2	.5;.5	0.870	0.857	0.861	0.869	0.869	0.859
			.3;.7	0.860	0.847	0.845	0.858	0.857	0.849
		.8;.8	.5;.5	0.907	0.872	0.904	0.876	0.912	0.849
			.3;.7	0.900	0.851	0.889	0.863	0.897	0.817
	1	.2;.2	.5;.5	0.860	0.857	0.861	0.858	0.857	0.858
			.3;.7	0.848	0.847	0.845	0.847	0.851	0.846
		.8;.8	.5;.5	0.909	0.872	0.904	0.871	0.907	0.855
			.3;.7	0.890	0.851	0.889	0.850	0.888	0.825

Table 5a. *Average bias of person parameters*.

					Estimation Model					
Data	DIF	OR	Corr.	Prop.	TM	UNM-N	UNM-C	UNM-D	OM	MISM
Complete	1.5	10	.2;.2	.5;.5	0.004	0.007	0.007	0.007	0.007	0.006
				.3;.7	0.007	0.012	0.011	0.010	0.010	0.008
			.8;.8	.5;.5	0.007	0.007	0.010	0.006	0.009	0.007
				.3;.7	0.011	0.012	0.014	0.011	0.013	0.010
		1	.2;.2	.5;.5	0.007	0.007	0.007	0.007	0.007	0.007
				.3;.7	0.011	0.011	0.011	0.011	0.011	0.011
			.8;.8	.5;.5	0.010	0.007	0.010	0.007	0.010	0.007
				.3;.7	0.015	0.012	0.014	0.012	0.015	0.013
	1	10	.2;.2	.5;.5	0.005	0.005	0.009	0.005	0.005	0.008
				.3;.7	0.008	0.010	0.018	0.008	0.008	0.012
			.8;.8	.5;.5	0.010	0.007	0.014	0.007	0.010	0.010
				.3;.7	0.011	0.008	0.020	0.007	0.011	0.010
		1	.2;.2	.5;.5	0.005	0.005	0.009	0.014	0.005	0.008
				.3;.7	0.010	0.010	0.018	0.010	0.010	0.012
			.8;.8	.5;.5	0.011	0.007	0.014	0.015	0.010	0.010
				.3;.7	0.013	0.008	0.020	0.008	0.012	0.010
Booklet Design	1.5	10	.2;.2	.5;.5	0.041	0.064	0.049	0.052	0.053	0.071
				.3;.7	-0.148	-0.183	-0.183	-0.161	-0.163	-0.216
			.8;.8	.5;.5	0.060	0.067	0.042	0.054	0.050	0.084
				.3;.7	-0.090	-0.184	-0.111	-0.161	-0.106	-0.210
		1	.2;.2	.5;.5	0.062	0.064	0.049	0.066	0.051	0.065
				.3;.7	-0.183	-0.183	-0.183	-0.196	-0.193	-0.182
			.8;.8	.5;.5	0.054	0.067	0.042	0.052	0.045	0.077
				.3;.7	-0.109	-0.184	-0.111	-0.197	-0.122	-0.192
	1	10	.2;.2	.5;.5	0.019	0.018	0.034	0.031	0.019	0.031
				.3;.7	-0.004	-0.011	0.025	0.014	-0.006	-0.024
			.8;.8	.5;.5	0.020	0.018	0.039	0.033	0.026	0.059
				.3;.7	0.025	-0.012	0.051	0.012	0.023	0.074
		1	.2;.2	.5;.5	0.018	0.018	0.034	0.027	0.022	0.039
				.3;.7	-0.010	-0.011	0.028	0.003	-0.010	-0.010
			.8;.8	.5;.5	0.024	0.018	0.038	0.027	0.030	0.069
				.3;.7	0.022	-0.012	0.051	0.004	0.022	0.074
Omitted Responses	1.5	10	.2;.2	.5;.5	0.100	0.112	0.108	0.097	0.096	0.114
				.3;.7	0.108	0.117	0.118	0.104	0.104	0.117
			.8;.8	.5;.5	0.092	0.113	0.100	0.097	0.090	0.116
				.3;.7	0.105	0.121	0.117	0.106	0.102	0.116
		1	.2;.2	.5;.5	0.112	0.112	0.108	0.108	0.108	0.112
				.3;.7	0.118	0.117	0.118	0.114	0.113	0.116
			.8;.8	.5;.5	0.104	0.113	0.100	0.109	0.103	0.112
				.3;.7	0.117	0.121	0.117	0.118	0.114	0.115

1	10	.2;.2	.5;.5	0.047	0.050	0.058	0.048	0.047	0.049
			.3;.7	0.055	0.058	0.069	0.055	0.055	0.057
		.8;.8	.5;.5	0.046	0.051	0.058	0.049	0.045	0.049
			.3;.7	0.052	0.057	0.066	0.053	0.051	0.050
	1	.2;.2	.5;.5	0.049	0.050	0.058	0.050	0.049	0.050
			.3;.7	0.059	0.058	0.069	0.058	0.059	0.058
		.8;.8	.5;.5	0.049	0.051	0.058	0.052	0.049	0.049
			.3;.7	0.057	0.057	0.066	0.058	0.056	0.051

Table 5b. *Average standard error of person parameters.*

					Estimation Model					
Data	DIF	OR	Corr.	Prop.	TM	UNM-N	UNM-C	UNM-D	OM	MISM
Complete	1.5	10	.2;.2	.5;.5	0.185	0.186	0.185	0.185	0.185	0.187
				.3;.7	0.119	0.118	0.119	0.119	0.119	0.120
			.8;.8	.5;.5	0.151	0.186	0.151	0.185	0.151	0.187
				.3;.7	0.097	0.119	0.098	0.118	0.097	0.120
		1	.2;.2	.5;.5	0.185	0.186	0.185	0.186	0.185	0.186
				.3;.7	0.119	0.120	0.119	0.120	0.119	0.120
			.8;.8	.5;.5	0.151	0.186	0.151	0.186	0.151	0.186
				.3;.7	0.098	0.119	0.098	0.119	0.098	0.119
	1	10	.2;.2	.5;.5	0.153	0.155	0.154	0.154	0.153	0.155
				.3;.7	0.098	0.099	0.097	0.098	0.098	0.099
			.8;.8	.5;.5	0.130	0.158	0.130	0.157	0.130	0.158
				.3;.7	0.084	0.101	0.084	0.100	0.084	0.102
		1	.2;.2	.5;.5	0.154	0.155	0.154	0.153	0.154	0.154
				.3;.7	0.098	0.099	0.097	0.099	0.098	0.099
			.8;.8	.5;.5	0.132	0.158	0.130	0.155	0.132	0.157
				.3;.7	0.086	0.101	0.084	0.101	0.086	0.101
Booklet Design	1.5	10	.2;.2	.5;.5	0.228	0.231	0.232	0.226	0.225	0.240
				.3;.7	0.215	0.224	0.223	0.221	0.221	0.231
			.8;.8	.5;.5	0.177	0.231	0.183	0.227	0.181	0.239
				.3;.7	0.161	0.222	0.170	0.219	0.165	0.225
		1	.2;.2	.5;.5	0.229	0.231	0.232	0.320	0.231	0.230
				.3;.7	0.223	0.224	0.223	0.230	0.228	0.224
			.8;.8	.5;.5	0.181	0.231	0.183	0.235	0.182	0.234
				.3;.7	0.169	0.222	0.170	0.227	0.173	0.224
	1	10	.2;.2	.5;.5	0.176	0.185	0.179	0.173	0.177	0.181
				.3;.7	0.127	0.129	0.115	0.123	0.127	0.130
			.8;.8	.5;.5	0.145	0.189	0.147	0.176	0.145	0.200
				.3;.7	0.096	0.127	0.086	0.122	0.096	0.136
		1	.2;.2	.5;.5	0.185	0.185	0.179	0.184	0.183	0.178
				.3;.7	0.128	0.129	0.114	0.132	0.128	0.129
			.8;.8	.5;.5	0.152	0.189	0.147	0.190	0.152	0.196
				.3;.7	0.096	0.127	0.086	0.129	0.097	0.134
Omitted Responses	1.5	10	.2;.2	.5;.5	0.196	0.195	0.195	0.198	0.197	0.193
				.3;.7	0.129	0.126	0.125	0.131	0.130	0.125
			.8;.8	.5;.5	0.158	0.195	0.158	0.199	0.158	0.193
				.3;.7	0.103	0.125	0.101	0.131	0.104	0.127
		1	.2;.2	.5;.5	0.194	0.195	0.195	0.196	0.195	0.195
				.3;.7	0.125	0.126	0.125	0.127	0.126	0.127
			.8;.8	.5;.5	0.158	0.195	0.158	0.196	0.158	0.195
				.3;.7	0.101	0.125	0.101	0.126	0.102	0.129

1	10	.2;.2	.5;.5	0.167	0.170	0.167	0.168	0.167	0.170
			.3;.7	0.112	0.115	0.111	0.113	0.112	0.116
		.8;.8	.5;.5	0.136	0.173	0.136	0.171	0.137	0.174
			.3;.7	0.092	0.118	0.094	0.116	0.092	0.121
	1	.2;.2	.5;.5	0.169	0.170	0.167	0.170	0.169	0.170
			.3;.7	0.114	0.115	0.111	0.116	0.114	0.116
		.8;.8	.5;.5	0.140	0.173	0.136	0.173	0.140	0.174
			.3;.7	0.095	0.118	0.094	0.118	0.095	0.122

Table 5c. *Average RMSE of person parameters.*

					Estimation Model					
Data	DIF	OR	Corr.	Prop.	TM	UNM-N	UNM-C	UNM-D	OM	MISM
Complete	1.5	10	.2;.2	.5;.5	0.199	0.199	0.199	0.199	0.199	0.200
				.3;.7	0.126	0.128	0.128	0.128	0.127	0.127
			.8;.8	.5;.5	0.176	0.199	0.177	0.199	0.177	0.199
				.3;.7	0.112	0.129	0.114	0.128	0.114	0.128
		1	.2;.2	.5;.5	0.199	0.199	0.199	0.199	0.199	0.199
				.3;.7	0.127	0.128	0.128	0.128	0.128	0.128
			.8;.8	.5;.5	0.177	0.199	0.177	0.199	0.177	0.200
				.3;.7	0.114	0.129	0.114	0.129	0.115	0.130
	1	10	.2;.2	.5;.5	0.164	0.165	0.164	0.164	0.164	0.164
				.3;.7	0.104	0.104	0.108	0.104	0.104	0.105
			.8;.8	.5;.5	0.150	0.168	0.150	0.167	0.150	0.168
				.3;.7	0.096	0.106	0.099	0.106	0.096	0.107
		1	.2;.2	.5;.5	0.165	0.165	0.164	0.166	0.165	0.164
				.3;.7	0.104	0.104	0.108	0.104	0.104	0.105
			.8;.8	.5;.5	0.151	0.168	0.150	0.168	0.151	0.168
				.3;.7	0.097	0.106	0.099	0.106	0.097	0.107
Booklet Design	1.5	10	.2;.2	.5;.5	0.271	0.307	0.292	0.280	0.279	0.326
				.3;.7	0.276	0.312	0.313	0.294	0.297	0.344
			.8;.8	.5;.5	0.245	0.309	0.237	0.282	0.236	0.345
				.3;.7	0.220	0.309	0.248	0.290	0.239	0.326
		1	.2;.2	.5;.5	0.303	0.307	0.292	0.363	0.292	0.307
				.3;.7	0.313	0.312	0.313	0.332	0.331	0.310
			.8;.8	.5;.5	0.246	0.309	0.237	0.298	0.238	0.327
				.3;.7	0.247	0.309	0.248	0.328	0.264	0.316
	1	10	.2;.2	.5;.5	0.196	0.201	0.206	0.202	0.196	0.206
				.3;.7	0.149	0.167	0.131	0.132	0.151	0.173
			.8;.8	.5;.5	0.173	0.204	0.178	0.204	0.172	0.248
				.3;.7	0.134	0.168	0.119	0.132	0.135	0.177
		1	.2;.2	.5;.5	0.201	0.201	0.206	0.204	0.201	0.211
				.3;.7	0.165	0.167	0.129	0.155	0.165	0.164
			.8;.8	.5;.5	0.180	0.204	0.177	0.208	0.179	0.254
				.3;.7	0.146	0.168	0.119	0.153	0.147	0.173
Omitted Responses	1.5	10	.2;.2	.5;.5	0.267	0.276	0.270	0.265	0.263	0.277
				.3;.7	0.206	0.213	0.213	0.203	0.202	0.213
			.8;.8	.5;.5	0.236	0.277	0.239	0.265	0.232	0.278
				.3;.7	0.186	0.217	0.196	0.204	0.183	0.212
		1	.2;.2	.5;.5	0.275	0.276	0.270	0.272	0.271	0.276
				.3;.7	0.213	0.213	0.213	0.209	0.208	0.212
			.8;.8	.5;.5	0.243	0.277	0.239	0.273	0.241	0.276
				.3;.7	0.197	0.217	0.196	0.212	0.193	0.211

1	10	.2;.2	.5;.5	0.194	0.196	0.202	0.194	0.194	0.196
			.3;.7	0.140	0.142	0.152	0.140	0.140	0.142
		.8;.8	.5;.5	0.170	0.199	0.178	0.197	0.170	0.199
			.3;.7	0.124	0.144	0.136	0.141	0.123	0.141
	1	.2;.2	.5;.5	0.195	0.196	0.202	0.196	0.195	0.196
			.3;.7	0.142	0.142	0.152	0.143	0.142	0.142
		.8;.8	.5;.5	0.173	0.199	0.178	0.199	0.173	0.199
			.3;.7	0.128	0.144	0.136	0.144	0.127	0.142

Table 5d. *Average 95% coverage rate of person parameters*.

					Estimation Model					
Data	DIF	OR	Corr.	Prop.	TM	UNM-N	UNM-C	UNM-D	OM	MISM
Complete	1.5	10	.2;.2	.5;.5	0.955	0.955	0.955	0.955	0.955	0.955
				.3;.7	0.955	0.955	0.955	0.956	0.955	0.956
			.8;.8	.5;.5	0.956	0.957	0.957	0.956	0.956	0.956
				.3;.7	0.955	0.955	0.955	0.955	0.955	0.955
		1	.2;.2	.5;.5	0.955	0.955	0.955	0.955	0.955	0.955
				.3;.7	0.955	0.955	0.955	0.955	0.955	0.956
			.8;.8	.5;.5	0.956	0.957	0.957	0.957	0.956	0.956
				.3;.7	0.955	0.955	0.955	0.955	0.955	0.954
	1	10	.2;.2	.5;.5	0.955	0.955	0.955	0.955	0.955	0.955
				.3;.7	0.955	0.954	0.954	0.954	0.954	0.954
			.8;.8	.5;.5	0.956	0.955	0.955	0.956	0.956	0.955
				.3;.7	0.955	0.955	0.955	0.955	0.955	0.954
		1	.2;.2	.5;.5	0.955	0.955	0.955	0.954	0.955	0.955
				.3;.7	0.954	0.954	0.954	0.954	0.954	0.954
			.8;.8	.5;.5	0.955	0.955	0.955	0.956	0.956	0.955
				.3;.7	0.955	0.955	0.955	0.955	0.955	0.954
Booklet Design	1.5	10	.2;.2	.5;.5	0.955	0.955	0.955	0.954	0.955	0.954
				.3;.7	0.944	0.943	0.942	0.945	0.945	0.939
			.8;.8	.5;.5	0.955	0.954	0.957	0.954	0.955	0.953
				.3;.7	0.953	0.942	0.953	0.945	0.953	0.924
		1	.2;.2	.5;.5	0.955	0.955	0.955	0.955	0.956	0.955
				.3;.7	0.943	0.943	0.942	0.944	0.943	0.942
			.8;.8	.5;.5	0.957	0.954	0.957	0.955	0.957	0.954
				.3;.7	0.953	0.942	0.953	0.943	0.952	0.927
	1	10	.2;.2	.5;.5	0.953	0.953	0.951	0.953	0.953	0.951
				.3;.7	0.956	0.954	0.952	0.955	0.956	0.955
			.8;.8	.5;.5	0.956	0.954	0.954	0.955	0.955	0.938
				.3;.7	0.953	0.955	0.946	0.956	0.954	0.928
		1	.2;.2	.5;.5	0.951	0.953	0.951	0.951	0.948	0.947
				.3;.7	0.955	0.954	0.953	0.954	0.955	0.955
			.8;.8	.5;.5	0.954	0.954	0.954	0.953	0.951	0.929
				.3;.7	0.951	0.955	0.946	0.954	0.951	0.927
Omitted Responses	1.5	10	.2;.2	.5;.5	0.932	0.928	0.928	0.932	0.932	0.927
				.3;.7	0.924	0.916	0.916	0.925	0.924	0.916
			.8;.8	.5;.5	0.940	0.930	0.938	0.934	0.939	0.929
				.3;.7	0.928	0.917	0.922	0.925	0.928	0.915
		1	.2;.2	.5;.5	0.928	0.928	0.928	0.928	0.928	0.929
				.3;.7	0.916	0.916	0.916	0.917	0.917	0.917
			.8;.8	.5;.5	0.938	0.930	0.938	0.930	0.937	0.930
				.3;.7	0.921	0.917	0.922	0.918	0.921	0.917

1	10	.2;.2	.5;.5	0.940	0.939	0.939	0.940	0.940	0.939
			.3;.7	0.939	0.936	0.936	0.939	0.939	0.936
		.8;.8	.5;.5	0.948	0.940	0.947	0.942	0.948	0.938
			.3;.7	0.943	0.936	0.942	0.939	0.943	0.932
	1	.2;.2	.5;.5	0.939	0.939	0.939	0.939	0.939	0.939
			.3;.7	0.936	0.936	0.936	0.936	0.936	0.936
		.8;.8	.5;.5	0.947	0.940	0.947	0.940	0.947	0.938
			.3;.7	0.942	0.936	0.942	0.936	0.942	0.932

Appendix B

Recall that in Equation 3.7:

$$\text{odds ratio (OR)} = \frac{(P(g=1 \mid D_i = 0)/P(g=2 \mid D_i = 0))}{(P(g=1 \mid D_i = 1)/P(g=2 \mid D_i = 1))}.$$

Given the two-way table:

	$D_i = 1$	$D_i = 0$	
$g = 2$	p_{21}	p_{20}	$p_{2\cdot}$
$g = 1$	p_{11}	p_{10}	$p_{1\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 0}$	

Then the OR may be expressed as:

$$\text{odds ratio (OR)} = \frac{p_{10}/p_{\cdot 0}}{p_{20}/p_{\cdot 0}} \Bigg/ \frac{p_{11}/p_{\cdot 1}}{p_{21}/p_{\cdot 1}} = \frac{p_{21}p_{10}}{p_{11}p_{20}}.$$

Given the values of the marginal probabilities $p_{\cdot 1}$ and $p_{2\cdot}$, when OR ≠ 1,

$$p_{21} = \frac{1 + (p_{2\cdot} + p_{\cdot 1})(OR - 1) - S}{2(OR - 1)},$$

where

$$S = \sqrt{(1 + (p_{2\cdot} + p_{\cdot 1})(OR - 1))^2 - 4OR(1 - OR)p_{2\cdot}p_{\cdot 1}}.$$

When OR = 1,

$$p_{21} = p_{2\cdot}p_{\cdot 1}.$$

The other three cell probabilities can be obtained based on the marginal probabilities and p_{21}.

Appendix C

Sample WinBUGS Code for the Estimation of the Data-generating Model (TM)

```
Model
{
# For mixing proportion
for (j in 1:J) { P.tot[j] <- sum(P[,j])/N;}

# Item Parameters
for (m in 1:M-1) {
for (j in 1:J) {
b[m,j] ~ dnorm(0, 1);
}
}
b[M,1] <- -1*sum(b[1:(M-1),1]);
b[M,2] <- -1*sum(b[1:(M-1),2]);

# Person Parameter and the linear function with covariates
for (i in 1:N) { G[i]~dcat( P[i,] )}
for (j in 1:J){a0[j] ~ dnorm(0,1);a1[j] ~ dnorm(0,1);}
taue[1] ~ dgamma(.5,1)
taue[2] ~ dgamma(.5,1)
for (i in 1:N){
thhat[i] <- a0[G[i]]+a1[G[i]]*con[i,1]
theta[i]~dnorm(thhat[i],taue[G[i]])
}

# Response model and the logit function with covariates
for (j in 2:J) { int[j] ~ dnorm(0,1); sl1[j] ~ dnorm(0,1);}
int[1]<-0; sl1[1] <- 0;

for (i in 1:N) {
for (j in 1:J) { P[i,j]<- PHI[i,j] / sum(PHI[i,]);
log(PHI[i,j]) <- int[j] + sl1[j]*dich[i,1];
}

for (m in 1:M) {
logit(p[i,m])<- theta[i]-b[m, G[i]];
resp[i,m] ~ dbern(p[i,m]);
}
}
}
```

References

Ackerman, T. A. (1992). An explanation of differential item functioning from a multidimensional perspective. *Journal of Educational Measurement*, *24*, 67-91.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistics*, *22*, 47-76.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716-723.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39,* 331-348.

Boughton, K., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 147-156). New York, NY: Springer.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In O. Opitz, B. Lausen & R. Klar (Eds.), Information and classification: Concepts, methods and applications (pp. 40-54). Berlin, Germany: Springer.

Brooks, S., & Roberts, G. O. (1998). Convergence assessments of Markov chain Monte Carlo algorithms. *Statistics and Computing, 8,* 319-335.

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach (2nd edition). New York, NY: Springer.

Celeux, G. (1998). Bayesian inference for mixtures: The label-switching problem. In R. Payne & P. Green (Eds.), COMPSTAT 98 (pp. 227-232). Heidelberg, Germany: Physica.

Celeux, G., Hurn, M., & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association, 95,* 957-970.

Chen, Y.-F., & Jiao, H. (2012). *The impact of missing responses on parameter estimation and classification accuracy in a mixture Rasch model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada.

Chen, Y.-F., & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of the PISA 2009 reading assessment. *Educational Assessment, 19*, 77-96.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics, 35,* 336-370.

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2006). *An investigation of priors on the probabilities of mixtures in the mixture Rasch model*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Canada.

Cho, S. J., Cohen, A. S., & Kim, S.-H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation, 83*, 278-306.

Chung, H., Loken, E., & Schafer, J. L. (2004). Difficulties in drawing inferences with finite mixture models: A simple example with a simple solution. *The American Statistician, 58,* 152-158.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133-148.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91*, 883-904.

Croon, M. A. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43,* 171-192.

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology, 44*, 315-331.

Dai, Y. (2009). *A mixture Rasch model with a covariate: A simulation study via Bayesian Markov Chain Monte Carlo estimation*. (Unpublished doctoral dissertation). University of Maryland, College Park.

Dai, Y. (2013). A mixture Rasch model with a covariate: A simulation study via Bayesian Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 37*, 375-396.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association, 83*, 173-178.

Dayton, C. M., & Macready, G. B. (1989). A latent class covariate model with applications to criterion referenced testing. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 129-143). New York, NY: Plenum.

Dayton, C. M., & Macready, G. B. (2007). Latent class analysis in psychometrics. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics, volume 26: Psychometrics* (pp. 421-446). Amsterdam, the Netherlands: Elsevier.

De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2*, 243-276.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38*, 213-234.

Diebolt, J. & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, 56*, 363-375.

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

Finch, W. H. (2011). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement, 71*, 663-683.

Finch, W. H., & Pierson, E. E. (2011). A mixture IRT analysis of risky youth behavior. *Frontiers in Quantitative Psychology, 2,* 1-10.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3-26.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87*, 476-486.

Fox, J.-P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68*, 169-191.

Fruhwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association, 96*, 194-209.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. New York, NY: Chapman & Hall.

Gilks, W. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith (Eds.), *Bayesian Statistics 4* (pp. 641-665). Oxford, UK: Oxford University Press.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*, 237-288.

Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable: Part I a modified latent structure approach. *American Journal of Sociology, 79*, 1179-1259.

Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215-231.

Haberman, S. J. (1979). *Analysis of qualitative data, volume 2: New developments*. New York, NY: Academic Press.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear, panel, trend and cohort analysis*. London, UK: Sage.

Hagenaars, J. A. (1993). *Loglinear models with latent variables*. London, UK: Sage.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

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

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, NY: Springer.

Jiao, H., & Chen, Y.-F. (2014). Differential item and testlet functioning. In A. Kunnan (Ed.), *The companion to language assessments* (pp.1282-1300). New York, NY: John Wiley & Sons, Inc.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*, 82-100.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79-93.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27,* 307-327.

Kim, J., & Bolt, D. M. (2007). An NCME instructional module on estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26*, 38-51.

Kohli, N., Harring, J. R., & Hancock, G. R. (2013). Piecewise linear–linear latent growth mixture models with unknown knots. *Educational and Psychological Measurement, 73*, 935-955.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for dichotomous mixture IRT models. *Applied Psychological Measurement, 33*, 353-373.

Li, L., & Hser, Y. (2011). On inclusion of covariates for class enumeration of growth mixture models. *Multivariate Behavioral Research, 46*, 266-302.

Li, M., Harring, J. R., Macready, G. B. (2014). Investigating the feasibility of using Mplus in the estimation of growth mixture models. *Journal of Modern Applied Statistical Methods, 13*, 484-513.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66,* 579-619.

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

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

Lu, R., & Jiao, H. (2009). *Detecting DIF using mixture Rasch model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21-39.

Lubke, G. H., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 26-47.

Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement, 59,* 615-630.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337.

Magidson, J., & Vermunt, J. (2004). Latent class models. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 175-198). Newbury Park, CA: Sage.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32,* 611-631.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement, 34*, 521-538.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11,* 81-91.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R. J. (1993). Should "multiple imputations" be treated as "multiple indicators"? *Psychometrika, 58,* 79–85.

Mislevy, R. J., Levy, R., Kroopnick, M., & Wise, D. (2006). *Evidentiary foundations of mixture item response theory models*. Paper presented at the Center for Integrated Latent Variable Research Conference, College Park, MD.

Mislevy, R. J., & Sheehan, K. M. (1989a). Information matrices in latent-variable models. *Journal of Educational Statistics, 14*, 335-350.

Mislevy, R. J., & Sheehan, K. M. (1989b). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*, 661–679.

Mislevy, R. J., & Verhelst, N. D. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215.

Mokken, R.J. (1996). Nonparametric models for dichotomous responses. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-367). New York, NY: Springer.

Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1-33). Mahwah, NJ: Erlbaum.

Muthén, B. O. (2004). Latent variable analysis: Growth modeling mixture and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345-368). Newbury Park, CA: Sage.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide (7th edition)*. Los Angeles, CA: Muthén & Muthén.

NCES. (2009). *The nation's report card: An overview of procedures for the NAEP assessment* (NCES 2009–493). Washington, DC: U.S. Government Printing Office.

OECD (2009). *PISA data analysis manual SAS second edition*. Retrieved from http://www.oecd-ilibrary.org/the-rasch-model_5kskx0xcmp8s.pdf;jsessionid=218v6ptg4012f.x-oecd-live-01?contentType=/ns/Chapter&itemId=/content/chapter/9789264056251-6-en&containerItemId=/content/serial/19963777&accessItemIds=&mimeType=application/pdf.

Petras, H., & Masyn, K. (2010). General growth mixture analysis with antecedents and consequences of change. In A. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 69-100). New York, NY: Springer.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Reckase, M. D. (1972). *Development and application of a multivariate logistic latent trait model*. Unpublished doctoral dissertation, Syracuse University, Syracuse, NY.

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

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

Rijmen, F., & de Boeck, P. (2003). A latent class model for individual differences in the interpretation of conditionals. *Psychological Research, 67*, 219-231.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Rost, J. & Langeheine, R. (1997). A guide through latent structure models for categorical data. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 13-37). Muenster, Germany: Waxmann.

Rost, J., & von Davier, M. (1993). Measuring different traits in different populations with the same items. 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. 446-450). Stuttgart, Germany: Gustav Fischer Verlag.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective*. Unpublished doctoral dissertation. University of Maryland, College Park.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333-343.

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

Smit, A., Kelderman, H., & van der Flier, H. (1999). Collateral information and mixed Rasch models. *Methods of Psychological Research Online, 4*, 19-32.

Smit, A., Kelderman, H., & van der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. *Methods of Psychological Research Online, 5*, 31-43.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linden, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 583-616.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS user manual (Version 1.4.3)*. Cambridge, UK: MRC Biostatistics Unit.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62*, 795-809.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Tay, L., Newman, D. A., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods, 14*, 147-176.

van der Heijden, P. G. M., Dressens, J., & Bockenholt, U. (1996). Estimating the concomitant variable latent-class model with the EM algorithm. *Journal of Educational and Behavioral Statistics, 21*, 215-229.

van der Heijden, P. G. M., Mooijaart, A., & de Leeuw, J. (1992). Constrained latent budget analysis. In C. C. Clogg (Ed.), *Sociological methodology 1992* (pp. 279-320). Cambridge, MA: Blackwell.

Verhelst, N. D., & Eggen, T. J. H. M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek* [Psychometric and statistical aspects of measurement research] (PPON rapport 4). Arnhem, the Netherlands: Cito.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology, 33*, 213-239.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*, 450-469.

Vermunt, J. K. & Magidson, J. (2000-2013). *Latent GOLD User's Guide*. Belmont, Massachusetts: Statistical Innovations Inc.

von Davier, M. (2001). *WINMIRA* [Computer software]. Kiel, Germany: Institute for Science Education.

von Davier, M. (2005a). *A general diagnostic model applied to language testing data* (ETS Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2005b). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models* [Computer software]. Princeton, NJ: Educational Testing Service.

von Davier, M. (2007). *Hierarchical general diagnostic models* (ETS Research Report No. RR-07-19). Princeton, NJ: Educational Testing Service.

von Davier, M. (2008a). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287-307.

von Davier, M. (2008b). The mixture general diagnostic model. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 255-275). Charlotte, NC: Information Age Publishing.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement: Interdisciplinary Research and Perspectives, 7*, 67-74.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8-28.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI monograph series, 2*, 9-36.

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C.R. Rao, & S. Sinharay (Eds.), *Handbook of statistics, volume 26: Psychometrics* (pp. 643-661). Amsterdam, the Netherlands: Elsevier.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement, 28*, 389-406.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43-74). New York, NY: Springer.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*, 114-128.

Yamamoto, K. (1987). *A hybrid model for item responses*. Unpublished doctoral dissertation, University of Illinois at Chicago.

Yamamoto, K. (1989). *A Hybrid model of IRT and latent class models* (ETS Research Report No. RR-89-41). Princeton, NJ: Educational Testing Service.

Yamamoto, K. Y., & Everson, H. T. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost, & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89-98). Muenster, Germany: Waxmann.

Zhu, X. (2013). *Distinguishing continuous and discrete approaches to multilevel mixture IRT models: A model comparison perspective*. Unpublished doctoral dissertation. University of Maryland, College Park

Zwinderman, A. H. (1997). Response models with manifest predictors. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 245-256). New York, NY: Springer.