ABSTRACT

Title of dissertation: LATENT GROWTH CURVE ANALYSIS WITH ITEM RESPONSE DATA: MODEL SPECIFICATION, ESTIMATION, AND PANEL ATTRITION

Xiaying Zheng, Doctor of Philosophy, 2017

Dissertation directed by: Professor Ji Seung Yang
Measurement, Statistics and Evaluation
Department of Human Development
and Quantitative Methodology

Measuring change in a construct over time in educational or psychological research is often achieved by administering the same items to the same respondents repeatedly over time. When item response data are categorical, a second-order latent growth model (LGM) can be used by incorporating an item response theory (IRT) model as the measurement model (referred to as LGM-IRT). Common item effects can be specified as orthogonal specific factors in the measurement model. This study investigated three issues in using LGM-IRT with common item effects, namely model parameterization, estimation of model parameters, and sample attrition. Selected longitudinal IRT models were first reviewed. The Schmid-Leiman transformation was used to transform the second-order model to first-order formulation so that the model could be estimated in common multidimensional IRT

software packages. Simulation studies were carried out to examine different methods of estimating the model, namely using different estimation methods (diagonally weighted least square estimator, Monte Carlo expectation-maximization algorithm and Metropolis-Hastings Robbins-Monro algorithm) and using reduced models. The estimation methods were examined under different test lengths, sample sizes, and panel attrition mechanisms. The reduced models were examined under complete data situation. One empirical analysis was conducted to compare and contrast the different methods using data from the "Multistate Study of Pre-Kindergarten 2001–2003" by the National Center for Early Development and Learning. The results of this research can provide provide guidelines on the utility of the model using aforementioned three estimation methods and the reduced models. The research combines modeling techniques of structural equation modeling and IRT and can make contribution to the literature of this unified general framework.

Latent Growth Curve Analysis with Item Response Data: Model Specification, Estimation, and Panel Attrition

by

Xiaying Zheng

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
2017

Advisory Committee:
Professor Ji Seung Yang, Chair/Advisor
Professor Jeffrey Harring, Co-Chair
Professor Gregory Hancock
Professor Parthasarathi Lahiri
Professor Laura Stapleton

To my parents

Acknowledgments

First and foremost I would like to thank my advisor, Dr. Ji Seung Yang for her unwavering supports over the past five years. I am fortunate to have the opportunity to work on challenging and stimulating projects with Dr. Yang. And I am immensely grateful for the energy and time she devoted to this project. I would also like to thank Dr. Laura Stapleton for her mentorship and supports during my academic career in EDMS program. I would also like to acknowledge my Co-Chair, Dr. Jeffrey Harring. Without his suggestions and comments, this thesis would not have been possible. I thank all the members in my committee for the many insightful comments that led to a much improved manuscript.

I would like to acknowledge Dr. Li Cai for providing a flexMIRT licence for the EDMS program. I would also like to acknowledge the administrative support from the staff members, including Jannitta Graham, Charm Mudd, Cornelia Snowden, and Maria Rawlings.

I thank all my friends and colleagues in EDMS program for their friendship, help, and laughters.

Lastly, I owe my deepest thanks to my family - my mother, father, uncles, and aunts, who love me unconditionally and have always been supportive of my life and career choices.

I apologize to those I've inadvertently left out. Thank you all!

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Measuring change in an educational or psychological construct of interest over time has been an active area in the field of educational and psychological research. The measurement of the change is often achieved by a repeated measures design, where the same items or the same subset of items are administered to the same examinees repeatedly over time. When item response data are categorical, item response theory (IRT) models are often adopted by researchers who have two purposes, namely item analysis and scoring. IRT provides the benefit of characterizing the items and examinees separately via item analysis and scoring (Lord, 1980). In order to utilize these features of IRT models in longitudinal data analyses, various longitudinal IRT models (e.g., Cai, 2010b; Embretson, 1991; Liu & Hedeker, 2006; te Marvelde, Glas, Van Landeghem, & Van Damme, 2006) have been proposed to calibrate items and measure examinees' latent change as a special application of multilevel IRT models and multidimensional IRT models (MIRT).

Under the framework of structural equation modeling (SEM), latent growth modeling (LGM; see e.g., Meredith & Tisak, 1990) has been a popular approach to longitudinal data analysis. The second-order LGM (See, Section 2.1.3 for more details) was proposed and widely used to analyze longitudinal data with observed

continuous indicators (Duncan & Duncan, 1996; Hancock & Buehl, 2008; Hancock, Kuo, & Lawrence, 2001; McArdle, 1988; Sayer & Cumsille, 2001). When the response data are categorical, summed or average scores have traditionally been used as the measures of the construct of interest. Using IRT model as the measurement model in a second-order LGM (referred to as LGM-IRT in this study) to give fair attention to both measurement and structural models is a relatively recent development in the field of IRT modeling (e.g., Jeon & Rabe-Hesketh, 2015; Paek, Li, & Park, 2016; Wang, Kohli, & Henn, 2016).

The LGM-IRT model has several advantages over their MIRT counterparts such as a simple structure model in which the full variance-covariance matrix of time specific latent abilities is estimated (See, section 2.12 for more details). First, the LGM-IRT models are more parsimonious than as the time-specific latent factors can be summarized with latent initial status and growth variables. Second, the LGM structure enables researchers to examine individual differences with a smaller number of latent growth factors corresponding to a theorized trajectory. Third, the covariance between examinees' initial status and their growth rates can be specified and estimated. Fourth, time-invariant and time-specific covariates could be incorporated into the structural part of the model to explain individual differences in the growth factors.

While the LGM-IRT modeling approach provides aforementioned benefits, application of the model in applied studies is somewhat limited due to several methodological challenges. The goals of this study are to identify and investigate the issues in model specification, estimation, and missing data in LGM-IRT models. The

background, purposes, and contributions of the current study are provided in the following sections of the chapter.

1.1 Background

In educational or psychological research, the same items are sometimes repeatedly administered to the same examinees across time in order to measure the examinees' change in the construct of interest. The responses to the items are often scored categorically due to the popularity of the Likert-scale format, which stems from wanting to understand the degree, not simply whether or non respondents agree or disagree. For example, in the "Multistate Study of Pre-Kindergarten 2001–2003" by the National Center for Early Development and Learning (Clifford, Bryant, Burchinal, & Barbarin, 2005), the academic skills of young children from pre-kindergarten trough kindergarten were evaluated across four semesters in two years. The teachers were asked to rate the students' Language and Literacy skills as well as Mathematical Thinking skills every semester, using nine five-category Likert-scale items per scale. In longitudinal psychological studies, such practice is also common. For example, in the "Korean Youth Panel Survey", 50 five-category Likert-scale items were repeatedly administered to the junior high school cohort over 6 waves to track changes in students' self-identity related constructs, such as self-esteem (measured by 12 items) and stress (measured by 17 items).

When item response data are categorical, using IRT models (see, Lord, 1980, among others for the fundamentals of IRT models) has been a popular method to

estimate item characteristics as well as examinees' latent abilities, traits or tendencies. As opposed to observed summed scores, IRT models have the advantage of extracting more information from item response patterns rather than aggregating observed scores. Since IRT takes the nonparallel items into account, measurement error can be reduced as long as the measurement model fits reasonably well. A wide variety of longitudinal IRT models have been proposed in the past three decades. These models were developed under different frameworks to answer a wide range of research questions and are distinguished, in part, because of the assumptions made about the models such as inclusion of local item dependence assumption. This section provides an overview of selected longitudinal IRT models and methodological issues associated with them. Of particular interest is the second-order LGM with IRT as the measurement model, where the issues of item local dependence and model parameterization can cause complications in application.

1.1.1 Overview of Selected Longitudinal IRT Models

Conventional unidimensional IRT models such as the graded response model (Samejima, 1969) were developed to estimate person and item parameters in a single test administration. However, under repeated measures designs, the new aspect of time adds complexities that conventional practices could no longer address properly. Longitudinal IRT models have been around and utilized since the 1970s (see e.g., Fischer, 1976, for early application of IRT to longitudinal data) to analyze response data collected under repeated measures designs. More recently, the advancements in

multilevel IRT (e.g., Adams, Wilson, & Wu, 1997; Fox & Glas, 2001; Jiao, Kamata, Wang, & Jin, 2012; Kamata, 2001) and multidimensional IRT (MIRT; e.g., Reckase, 1985, 2009) have provided new perspectives and methods to model longitudinal item response data. A brief overview of selected multilevel/multidimensional IRT models for repeated measures data (Cai, 2010b; Embretson, 1991; Jeon & Rabe-Hesketh, 2015; Liu & Hedeker, 2006; Paek et al., 2016; te Marvelde et al., 2006; Wang et al., 2016) is presented in this section.

The conceptual path diagrams of selected models reviewed in this research are presented in Figure 1.1 to illustrate the methods. More detailed reviews and path diagrams of these models can be found in Chapter 2.

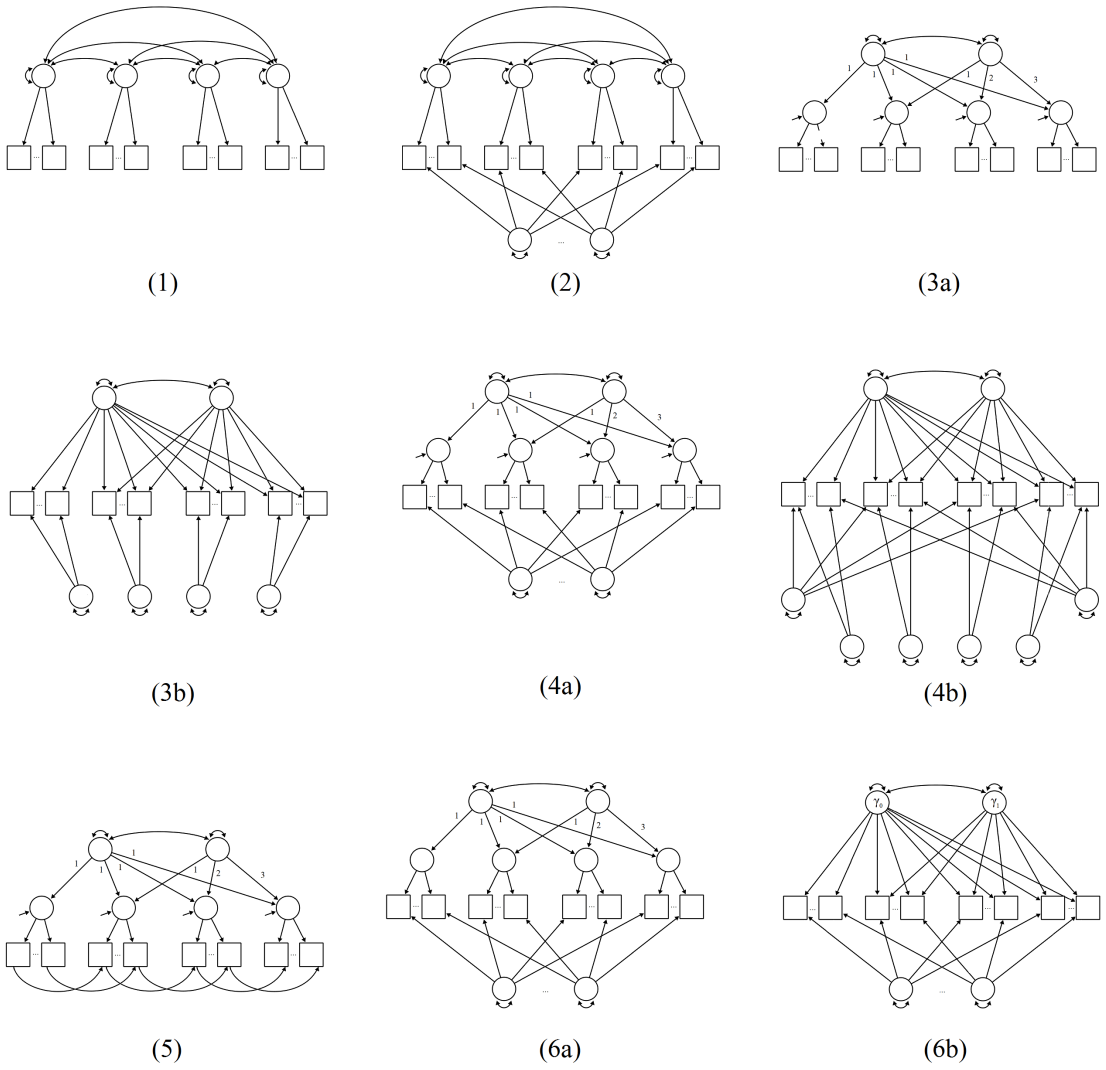

Figure 1.1. Conceptual path diagrams for selected longitudinal IRT models. Diagram 1 is the simple structure correlated-factor model by te Marvelde et al. (2006). Diagram 2 is the two-tier item factor model by Cai (2010b). Diagram 3a is the second-order LGM-IRT model without local dependence consideration by Paek et al. (2016). Diagram 3b is the first-order formation of Diagram 3a. Diagram 4a is the second-order LGM-IRT model with common item effects. Diagram 4b is the first-order formation of Diagram 4a. Diagram 5 is the second-order LGM-IRT model with order local dependence by Jeon and Rabe-Hesketh (2015). Diagram 6a is the second-order LGM-IRT model with common item effects but without time-specific disturbances. Diagram 6b is the first-order formation of Diagram 6a.

Under the multilevel modeling framework (Raudenbush & Bryk, 2002), longi-

tudinal data have been analyzed by treating time points as the level-1 units nested within level-2 persons (see e.g., Verbeke & Molenberghs, 2000, among others for the fundamentals of this approach). The change in the outcome variable can be assessed by using the time variable as a covariate at level one. In IRT modeling, similar techniques have been used to analyze longitudinal item response data (e.g., Liu & Hedeker, 2006). The item response data from all time points could be incorporated into a two-level IRT model by treating the latent scores at different time points as the level-1 units nested within persons. The changes in examinees' latent abilities are modeled by regressing the time-specific abilities on time. The examinees' growth curves can then be interpreted using the random intercepts and random coefficients for time (see, Curran, Edwards, Wirth, Hussong, & Chassin, 2007; Curran et al., 2008; Liu, Hedeker, & Mermelstein, 2013, for applied examples of this method).

In the framework of MIRT, several models have been proposed to analyze longitudinal item response data. For example, Embretson (1991) proposed a within-item longitudinal MIRT model, where the items could load on multiple dimensions. The latent variable at the first time point represented the baseline abilities, while the dimensions at other time points were parameterized as the changes from previous time points (see Figure 2.1 for graphical representation of this model and Section 2.1.2 for more details). Other researchers approached the question from a between-item dimensionality perspective, where each item could only load on one dimension. For example, te Marvelde et al. (2006) proposed using simple-structure correlated factors MIRT to analyze longitudinal item response data. Instead of using baseline and difference parameters as in the within-item MIRT model, the simple-structure

correlated factor MIRT model estimates the latent abilities of examinees at all time points (see Diagram 1, Figure 1.1). Hill (2006) and Cai (2010b) proposed using another type of between-item MIRT, namely the two-tier full-information item factor model to analyze longitudinal item response data (see Diagram 2, Figure 1.1). In addition to the correlated factors as in te Marvelde et al. (2006), the two-tier model introduces a series of orthogonal specific dimensions to address the residual dependence among repeated categorical items even after the examinees' latent scores are taken into account (i.e., item local dependence; Crocker & Algina, 1986).

Recently, some researchers have extended the between-item MIRT models to the second-order LGM model by imposing a growth structure on the time-specific latent variables. Instead of modeling multiple correlated factors to represent the latent abilities at each time point, the LGM-IRT model characterizes the growth curve of examinees with a mathematical function while allowing individual differences in the change parameters (e.g., random slopes and intercepts for linear change). For example, the simple-structure correlated factor MIRT could be extended to a second-order LGM without local item dependence (Paek et al., 2016, Diagram 3a, Figure 1.1). The two-tier model could be extended to a second-order LGM with common item effects in the measurement model (Wang et al., 2016, Diagram 4a, Figure 1.1). Viewing the local item dependence as an order effect, Jeon and Rabe-Hesketh (2015) developed a second-order LGM with "order local dependence" where the response to an item in later measurement waves was regressed on the response to the same item in the previous wave (see Diagram 5, Figure 1.1).

Although these longitudinal IRT models are have been around, practical ap-

plications of these models are limited due to several methodological issues. The methodological issues in applying these longitudinal IRT models are presented in the following section.

1.1.2 Methodological Issues in Longitudinal IRT Models

In this section, the selected longitudinal IRT models are briefly introduced in light of three methodological issues, namely the utilization of available response data, consideration of item local dependence, and estimation.

Utilization of all available item response data. As Hill (2006) pointed out, even with longitudinal surveys, researchers sometimes only use the data from the first administration for item calibration. Response data at subsequent time points are then scored with item parameters obtained from the first time point. This practice leaves out the additional information from subsequent measurements, which could be used for more stable and accurate calibration. The loss of information can be a severe issue especially when sample sizes are small, as IRT models require sufficient sample size to be properly estimated. Thus, using only the first-wave data for calibration under small sample size could result in unstable item parameter estimates (i.e., estimates with large standard errors). The uncertainty in these estimates could be further carried over in subsequent scoring processes (Cheng & Yuan, 2010; Liu & Yang, 2017; Patton, Cheng, Yuan, & Diao, 2013, 2014; Thissen & Wainer, 1990; Yang, Hansen, & Cai, 2012).

All the longitudinal IRT models reviewed in this research allow for joint cal-

ibration of data from all time points. The multi-wave data are modeled either as unidimensional IRT conditioned on time (e.g., multilevel IRT for longitudinal data), or as MIRT models with each primary dimension representing the latent scores at a time point. It is worth noting that, by correlating the main factors in the between-item MIRTs, the main factors could "borrow strength" from each other to improve the precision of item parameter estimates as well as latent ability score estimates (Cai, 2010b).

Consideration of item local dependence. Item local independence is an important assumption of IRT models, which requires that the responses to different items should not be related to each other after the respondent's latent score is taken into account (Crocker & Algina, 1986). However, under repeated measures designs, the local independence assumption hardly holds. After all, the examinees actually respond to the same items or subset of items multiple times. As shown in a simulation study by Tuerlinckx and De Boeck (2001), IRT models are typically not robust to violation of the local independence assumption as it can cause biased item parameter estimates. Thus, using IRT models to measure examinees' latent change requires mechanisms that account for the local dependence introduced by using the same items multiple times.

Three of the aforementioned models explicitly address the issue of local item dependence, namely the two-tier model, the LGM-IRT extended from the two-tier model, and the LGM-IRT with order local dependence. Among them, the two-tier model and its LGM-IRT extension parameterize the residual item dependency as a series of orthogonal specific factors. Local item dependence is treated as a

random effect, which captures the unique interaction of each examinee with the repeated items (Cai, 2010b; Maydeu-Olivares & Coffman, 2006). The same modeling techniques have also been used to address local dependence among items that share the same stimulus, which is referred to as the "testlet effect" (Lee & Frisbie, 1999). In order to avoid confusion with the "testlet effect," this research adopts the term "common item effects" to refer to local item dependence parameterized as specific factors in a longitudinal data analysis context.

In contrast to the common item effect, the LGM-IRT with order local dependence by Jeon and Rabe-Hesketh (2015) treats item local dependence as an item-specific fixed effect. It can be seen as a fixed adjustment to an item intercept depending on the response to the same item in the previous wave. The two types of local item dependence are suitable for different testing contexts. The common item effect is a type of symmetrical "combination dependence" caused by common items, common item stimulus or content domain, whereas the order local dependence is most appropriate for items with ordering effects (e.g., learning from the previous items could make the subsequent items easier) (Hoskens & De Boeck, 1997).

Estimation. Despite the advancements in multilevel and multidimensional IRT models for analyzing repeated measures data, the application of these models in research could be hindered by the estimation difficulty associated with these models. The major challenge is the "curse of dimensionality" (Bellman, 1957) when IRT models are estimated using quadrature-based procedure such as the popular Bock-Aitkin expectation-maximization algorithm for full-information maximum likelihood (FIML-BAEM; Bock & Aitkin, 1981). The computational challenge occurs when the

number of latent variables increases in a model because the number of quadrature points to be evaluated to integrate out the latent variables would increase exponentially. Take a model with three latent dimensions for example. If 15 quadrature points per dimension are utilized for numerical integration, which is the default of Mplus 7.4 (Muthén & Muthén, 1998-2012), the total number of quadrature points that is required to evaluate the response pattern likelihood is $15^3 = 3375$. As the number of response patterns increases (e.g., sample size), the required time to evaluate the likelihood also increases considerably. Typically, estimating a model with more than three dimensions becomes considerably burdensome and impractical. While Adaptive quadrature points (Rabe-Hesketh, Skrondal, Pickles, et al., 2002) can be another option, the methods (e.g., mean and variance; mode and curvature) are multiple and the performance is less examined for multidimensional IRT models. For example, Yang & Zheng (accepted) reported that Mean method in Stata IRT package did not properly estimate a unidimeisonal model with mixed format items.

Of the aforementioned models, the multilevel IRT model for longitudinal data would require three-dimension integration when the growth curve is assumed to be linear (i.e., level-1 time-specific latent variable, level-2 person-specific latent variable, and random slopes of time). The number of dimensions in the between-item MIRT model (Embretson, 1991) and the simple-structure correlated factor model (te Marvelde et al., 2006) equals the number of time points N in an assessment, since each dimension represents the latent scores at one time point. The LGM-IRT without local dependence (Paek et al., 2016), an extension of the simple-structure correlated factor model, would also have N dimensions. The two-tier model would

have $N + I$ dimensions, where I is the number of repeated items. Similarly, the LGM-IRT with common item effects, an extension of the two-tier model, would also have $N + I$ dimensions. The LGM-IRT with order local dependence Jeon and Rabe-Hesketh (2015) would require N dimensional integrations if the conventional BAEM algorithm is used to obtain FIML estimates. .

It would appear that the two-tier model has the highest number of dimensions that need to be integrated over when using FIML estimation. However, the two-tier structure allows analytical dimension reduction techniques (Gibbons & Hedeker, 1992) to be implemented for estimation efficiency. Cai (2010b) has shown that, when the specific dimensions are mutually orthogonal and are uncorrelated with the general factors, the dimensionality of the model could be analytically reduced from $N+I$ to $N+1$ by breaking the integral over the I specific dimensions into a product of I one-dimensional integrals. This means that the number of dimensions in the two-tier model after dimension reduction is always $N + 1$ regardless of the number of repeated items. The dimension reduction technique in IRT models is similar to the method by Cudeck, Harring, and du Toit (2009), where three-dimensional integration is reduced to a one-dimensional problem by utilizing conditional linearity in the context of nonlinear SEM.

As will be demonstrated in Chapter 2, The LGM-IRT model without local dependence can be transformed to a first-order model using the Schmid-Leiman transformation (Schmid & Leiman, 1957) so that it has a two-tier structure (see Diagram 3b Figure 1.1), which also allows for only three-dimensional integration after the dimension reduction technique is implemented. However, when common

item effects are included in the LGM-IRT, the re-parameterized first-order formation cannot be reduced to three dimensions. This is because the specific factors for common item effects and the specific factors for time-specific disturbances are crossed with each other when parameterized as a first-order model (see Diagram 4b Figure 1.1), making the same dimension reduction technique inapplicable. For example, if an assessment has four waves and two repeated items, the model would have six dimensions. Often longitudinal assessments can have more than four time points and two items. Due to the extremely high dimensionality of the LGM-IRT with common item effects, estimation with FIML-BAEM becomes impractical in applied studies.

Methodological issues that require further investigation. As discussed above, all aforementioned IRT models for repeated measures data allow for joint calibration of data from all time points. When item local dependence is ignored, the second-order LGM-IRT can be re-parameterized as a first-order model to reduce the dimensionality to only three. However, when item local dependence is modeled as common item effects, the same dimension reduction technique cannot be utilized. Thus, the estimation of the LGM-IRT becomes very challenging due to the involvement of high-dimensional latent variables. Alternative methods are needed to handle the estimation difficulty in the LGM-IRT model with common item effects. Two such approaches are presented in the following section.

1.1.3 Approaches to Handling Estimation Difficulty in LGM-IRT with Common Item Effects

This section describes two potential approaches to addressing the estimation difficulty of the LGM-IRT model with common item effects. The first method is to explore alternative estimation methods that are more computationally efficient than FIML-BAEM. Within the framework of FIML estimation, potential candidates include Monte Carlo expectation-maximization (FIML-MCEM; Wei & Tanner, 1990) and the Metropolis-Hastings Robbins-Monro algorithm (FIML-MH-RM; Cai, 2010a), both of which were specifically developed to handle models with high dimensionality. As far as the author is aware of, only one comparison study of the two algorithms (see, Han & Paek, 2014) has been done in IRT models with no more than three latent variables. The results suggested that they are comparable in terms of parameter accuracy in three-dimensional IRT models. However, in the LGM-IRT model with common item effects, the number of dimensions is typically much larger than three. It is not yet clear whether the full model could be stably estimated with either algorithm.

Aside from estimation methods under the FIML framework, methods within the limited information estimation framework can also be considered, such as the diagonally weighted least squares (DWLS) estimation for categorical data (e.g., Christoffersson, 1975; Muthén, du Toit, & Spisic, 1997). The limited information estimation method utilizes only the univariate and bivariate margins in the contingency table for estimation, which greatly reduces the computational burden.

However, as Maydeu-Olivares and Joe (2005) pointed out, the computational ease of limited-information estimation is achieved by ignoring higher-order associations among items. In theory, limited-information estimation is less ideal than the FIML method due to the omission of higher-order information. But it is not clear how the limited-information solutions will compare with the FIML methods when the LGM-IRT model is fitted under different data conditions.

An issue that could further complicate the performance of FIML and limited information estimation methods in the longitudinal data context is sample attrition (i.e., examinees permanently dropping out of the study). Attrition is a common problem in longitudinal surveys, which poses two potential challenges for longitudinal analysis. First, attrition makes the matrix of response data more sparse, which could have a negative effect on the performance of the estimation methods. More importantly, when the missing mechanisms (Rubin, 1976) are not handled properly, the parameter estimates could be biased and the inference could be misleading. For example, attrition not at random can lead to biased inference. Even when the attrition occurs at random, different missing at random (MAR) mechanisms could have different effects on the performance of aforementioned estimation methods. Previous research showed that FIML estimation methods could produce consistent estimates under MAR with respect to both covariates and observed outcomes (see e.g., Enders, 2001, among others), while limited information estimators such as DWLS were only consistent under MAR with respect to covariates (MAR-X; Asparouhov & Muthén, 2010). Comparison studies of FIML and limited information estimation have mostly been conducted without missing data (e.g., Forero & Maydeu-Olivares, 2009). The

effects of MAR attrition on the performance of the three estimation methods have not yet been comprehensively compared.

The second potential method to circumvent the high-dimensionality issue is to simplify the model so that the dimension-reduction technique (Gibbons & Hedeker, 1992) could be utilized to reduce the dimensionality of the model to only three. There are two types of nuisance factors in the LGM-IRT, namely the common item effects in the measurement part of the model and the time-specific disturbances in the structural part. When either type of nuisance factor is dropped from the model, the model can be transformed into a three-dimensional integration problem. As shown in Diagrams 3b and 5b in Figure 1.1, the two reduced models can be transformed from the second-order formulations to their first-order equivalents. Using the first-order formulation of the reduced models, it is easy to find that, assuming a linear growth trajectory, the total number of dimensions in either model is always three no matter how many items or time points are present. It is assumed that mis-specifying the model by dropping one type of nuisance could cause bias in the parameters of the measurement part and/or the structural part of the model. However, literature in this regard is limited. Wang et al. (2016) indicated that omitting the common item effects had little influence in the parameter estimates of the model in their empirical example, citing a pilot test they conducted. But the sensitivity of the parameter estimates to the omissions of nuisance factors has not been comprehensively investigated with simulation studies. If the item or structural parameters were indeed robust to the omission of either kind of nuisance factors, it would be advantageous to use the simplified models for computational efficiency.

1.2 Purposes of the Study

This study considers a LGM-IRT model that includes random common item effects in its measurement model. Motivated by the estimation complications caused by high dimensionality of the model, this research seeks to achieve five specific goals.

The first goal of the research is to provide a review of selected longitudinal IRT models. Special attention will paid to the parameterizations of these models under both SEM and MIRT frameworks. The interrelations among these models will also be explained. Detailed reviews of selected longitudinal IRT models are presented in Chapter 2.

The second goal of the research is to compare, via a simulation study, the performance of the aforementioned three estimation methods (namely FIML-MCEM, FIML-MH-RM, and DWLS) in estimating the full model when no attrition is present. The three estimation methods will be compared in terms of estimation time and parameter accuracy under different data conditions such as test length, sample size and so on.

The third goal of the research is to assess the performance of the two reduced models by dropping the common item effects or the time-specific disturbances. The estimated item and structural parameters from the reduced models will be compared to the true generating values to gauge the effects of the misspecifications.

The fourth goal of the research is to compare the performance of the three estimation methods in estimating the full model when attrition occurs under MAR. Two types of MAR mechanisms are considered, namely, MAR with respect to both

covariates and observed outcomes, and MAR with respect to covariates only (MAR-X). It should be noted that the missing-completely-at-random (MCAR) mechanism is not examined in this study since it is less common in longitudinal studies (Young & Johnson, 2015). The missing not at random (MNAR) mechanism is also not included in the study, since there is another line of research that specifically deals with MNAR in LGM models. Investigating these methods is out of the scope of the current study.

The fifth goal of the research is to provide an empirical illustration of applying the LGM-IRT model to real-world data with all three estimation methods. The feasibility of using these methods are examined. And the results from the three methods will be compared and contrasted.

This research is to answer three main research questions listed below.

- How does the performance of the aforementioned estimation methods compare with each other under complete data?
- What is the impact of dropping common item effects or time-specific disturbances on model parameter estimates under complete data?
- How does the performance of the aforementioned estimation methods compare with each other under MAR and MAR-X panel attrition?

1.3 Contributions of the Study

Although the LGM-IRT model with common item effects has been presented previously (Wang et al., 2016), it has not yet been used in applied studies due to

estimation difficulties. The estimation of the LGM-IRT model is less explored. The performance of different estimation methods within the context of LGM-IRT was not fully examined before as far as the author is aware of. The current study can provide guidelines on the utility of the model using aforementioned three estimation methods.

The study can also provide information on the effects of ignoring the nuisance factors. Due to the estimation difficulties of the full model, researchers sometimes omit components of the full model for computational ease. As mentioned before, when either of the two types of nuisance factors is dropped, the dimensionality of the model could be reduced to only three. Even though dimension reduction lifts the computational burden greatly, its effect has not been comprehensively investigated. The findings of the current study can provide new insights into the issue.

The current study combines modeling techniques of the two-tier item factor model and latent growth model. The equivalence of second-order latent growth models and MIRT models is demonstrated with re-parameterization. The study is situated in the unified framework of IRT and structural equation modeling (SEM) and makes contribution to the literature under this unified general framework.

Chapter 2: Literature Review

To achieve the first goal of the research, three classes of IRT models for longitudinal analysis are first reviewed. The parameterizations of the models and the interrelations between the models are explained by numeric equations as well as path diagrams. The Candidate estimation methods for high-dimensionality IRT models are then described, namely FIML-MCEM and FIML-MH-RM, and limited-information estimators. Lastly, sample attrition in longitudinal studies and its effects on the performance of these estimation methods are also introduced.

2.1 Longitudinal Item Response Theory Models

Three classes of longitudinal IRT models are identified in this research based on the frameworks under which they were developed. The three frameworks and the specific models are presented in Table 2.1. Each model is described in detail, and the interrelations between these models are discussed.

It should be noted that some early longitudinal IRT models are not reviewed in this study, such as the linear logistic test model (Fischer, 1973, 1983, 1995) and the linear logistic model with relaxed assumptions (Fischer, 1976, 1983, 1989), because they require stronger assumptions and are less used nowadays. The focus of this

Table 2.1

Classification of Longitudinal IRT Models Reviewed in This Study

Framework	**Models**
Multilevel IRT models	Multilevel IRT model that decomposes person parameters
Single-level MIRTs	Within-item multidimensional model
	Simple structure correlated factor model
	Two-tier item factor model
LGM-IRT models	LGM-IRT without local dependence consideration
	LGM-IRT with order local dependence
	LGM-IRT with common item effects

study is on longitudinal IRT models that measure latent change at the test level. Therefore this research does not include models that measure item-level change, such as the three-level multilevel IRT model that decomposes the item intercepts (Liu & Hedeker, 2006) and the item-level growth curve model (Paek et al., 2016).

For the purpose of model description, the following indices are used throughout the review:

- $n = 1, \ldots, N$ is the index for time points.
- t_{nj} is the time value associated with the n^{th} time point for person j. If the testing timings are the same for all examinees, the time value for all examinees at the n^{th} time point is t_n
- $i = 1, \ldots, I$ is the index for items that are repeated at each time point.
- $j = 1, \ldots, J$ is the index for examinees.

For simplicity of model description, all the models reviewed in this chapter are described using the two-parameter logistic (2-PL) model (Birnbaum, 1968), assuming that the tests are made up of dichotomous items. They can be easily generalized

to polytomous items when the graded response model is adopted (Samejima, 1969) or mixed-format tests. The 2-PL model can be written as:

$$P(Y_{ij} = 1 \mid \theta_j) = \frac{1}{1 + \exp(-\alpha_i \theta_j - \lambda_i)}, \tag{2.1}$$

where Y_{ij} is person j's response to item i, θ_j is the latent score of the person, and α_i and λ_i are the slope (i.e., discrimination) and intercept of item i respectively. The difficulty of the item is $-\lambda_i/\alpha_i$. The term $-\lambda_i$ is referred to as the "threshold" in categorical confirmative factor analysis. For discussions on the equivalence of IRT and categorical confirmative factor analysis, readers are referred to the works of Takane and De Leeuw (1987) and Kamata and Bauer (2008).

2.1.1 Multilevel IRT Models for Assessing Change

Multilevel modeling (Raudenbush & Bryk, 2002) has been a popular method to account for nested data structures, which are common in education settings (e.g., students are nested within classes within schools). Under the multilevel modeling framework, multilevel IRT models (see e.g., Adams et al., 1997; Fox & Glas, 2001; Jiao et al., 2012; Kamata, 2001) have been developed to handle the clustering effect of examinees. This is typically achieved by decomposing the person parameters into level-1 (e.g., students) and level-2 (e.g., schools) components. In a repeated measures design, the data structure can be viewed as items nested within time points within examinees. When multilevel IRT models are used to assess test-level change in repeated measures context (see, Curran et al., 2007, 2008; Liu et al., 2013,

for applied examples of this method), the students become the level-2 units, while time points are the level-1 units. A smooth growth curve (e.g., linear or polynomial) function is fitted by entering the time value variable into the multilevel structure as a level-1 covariate. The latent scores are represented as a combination of fixed and random effects. Take the 2-PL IRT model for example, examinee j's probability of answering item i at the nth time point can be written as:

$$P(Y_{inj} = 1 \mid \theta_{nj}) = \frac{1}{1 + \exp(-\alpha_i \theta_{nj} - \lambda_i)}, \tag{2.2}$$

where α_i and λ_i are the item slope and intercept respectively. The subscripts of the item parameters do not include time point index n, assuming that item parameters are time-invariant. When the growth curve is assumed to be linear, the time-specific latent ability θ_{nj} for person j at time point n can be decomposed as a combination of fixed and random person effects and time-specific residuals:

$$\theta_{nj} = (\gamma_0 + u_{0j}) + (\gamma_1 + u_{1j}) t_{nj} + \epsilon_{nj}, \tag{2.3}$$

where γ_0 is the population-level intercept; γ_1 is the population linear change rate; u_{0j} and u_{1j} are the random intercept and random slope effects for examinee j; and ϵ_{nj} is the level-1 disturbance term for person j's latent score at the n^{th} time point. Thus, each examinee's growth curve could be quantified with his/her unique intercept and growth rate. The multilevel structure in Equation (2.3) could also be modified to accommodate other person-level or occasion-level covariates and other trajectory

function forms (e.g., quadratic or exponential curve).

Using multilevel IRT models to measure change in latent scores has several advantages. First, time-invariant and time-variant covariates could be entered into level two and level one of the model, respectively. One could use standard likelihood ratio tests or fit indices to decide what covariates should be included. Second, since the change is modeled as some smooth function of time, the timings of the tests do not have to be the same for each examinee. Note that in Equation (2.3), the time value t_{nj} could be person-specific. This means that not all examinees have to be measured at the same occasion, although this is less common in educational testing. Third, by allowing person-level random effects, the multilevel IRT models could account for the heterogeneity in examinees' growth curves. The covariance of the random intercepts and the random slopes could also be estimated.

In repeated measures designs, it is common that the time-specific disturbances ϵ_{nj} could have a time series structure (Hedeker & Gibbons, 2006). It is worth noting that such covariance structures are used to model the dependence of respondent's disturbances over time. Thus they do not account for item-level conditional dependence. Lack of local dependence consideration is the major disadvantage of this model.

Multilevel longitudinal IRT models are typically estimated by the method of maximum likelihood. Ideally, it is more statistically efficient to simultaneously fit and test both the measurement model in Equation (2.2) and the random coefficient model in Equation (2.3), since no information is lost in the process when all parameters are jointly estimated. However, as Curran, Edwards, Wirth, Hussong,

and Chassin (2007) pointed out, empirical data in developmental research are often characterized with significant complexity. A simultaneous estimation strategy might not be feasible in empirical studies due to convergence problems stemming from estimation difficulties. Jointly testing covariates and the time series covariance structure can be time consuming. When convergence was an issue, Curran et al. (2007) advocated a two-stage estimation strategy, where item calibration was done in stage 1 and the growth curve was fitted and tested in stage 2. Such an approach would lead to loss of statistical efficiency. But it could be a practical compromise with empirical data.

2.1.2 Multidimensional IRT Models for Longitudinal Data

The second class of longitudinal IRT models were developed under the general framework of MIRT (Reckase, 1985, 2009). Three MIRT models for longitudinal data are reviewed in this section, namely the within-item multidimensional model (Embretson, 1991), the simple structure correlated-factor model (te Marvelde et al., 2006), and the two-tier model (Cai, 2010b; Hill, 2006). The details of the three models are presented below.

Embretson (1991) presented a within-item multidimensional Rasch model (Rasch, 1960) where the change from one time point to the next was explicitly parameterized. Within-item multidimensionality means that each item could load on more than one latent dimension. In the within-item longitudinal MIRT model, each time

point represents a dimension. The model is expressed as:

$$P(Y_{inj} = 1 \mid \boldsymbol{\theta}_j) = \frac{1}{1 + \exp(-\sum_{n=1}^{N} \theta_{nj} - \lambda_i)}. \tag{2.4}$$

At the first time point, latent score θ_{1j} is the baseline ability. When $n > 1$, θ_{nj} is interpreted as the change in person j's latent ability from time point $(n-1)$ to time point n. For example, θ_{2j} was the change in latent score for person j between the first and the second testing occasion. The item intercept λ_i was assumed to be time-invariant, while the item loadings were all set to "1" following the tradition of Rasch models. The change in latent scores are not modeled as a smooth function. Rather, a specific change parameter is estimated after the first time point for each person in a piece-wise manner.

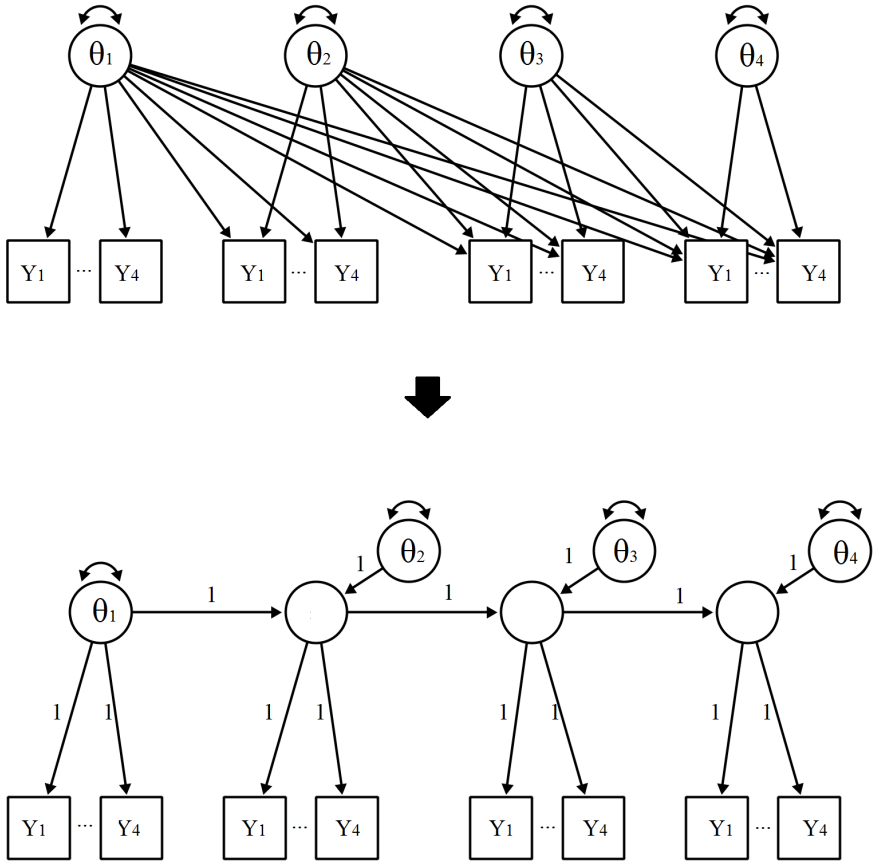

Figure 2.1. Within-item multidimensional IRT Model for longitudinal data by Embretson (1991). This example includes four time points and four repeated items. The upper panel is based on the original parameterization of Rasch model, where all loadings are 1. The model can be transformed to a second-order structure as shown in the lower panel.

The path diagram of the within-item multidimensionality model is presented in the upper panel of Figure 2.1 based on the original parameterization of Embretson (1991). The path diagram could be transformed to the second-order structure as shown in Figure 2.1 using a Schmid-Leiman transformation (Schmid & Leiman, 1957), which is more common in the SEM framework. Even though the original model was developed as a Rasch model, it could be easily generalized to a 2-PL model by relaxing the equal-discrimination constraint (see Koran, 2009). It can be

observed from the lower panel of Figure 2.1 that this model is equivalent to the latent change score model (McArdle, 2001, 2009) with categorical indicators. The potential weakness of this model is that, as demonstrated in the lower panel of Figure 2.1, the variance of latent scores has to increase constantly along time, which might not be realistic in empirical research.

Instead of using a within-item MIRT model, some researchers proposed using between-item-dimensionality MIRT models for longitudinal data, where items administered at a time point only loaded on an occasion-specific dimension. One such example is the simple structure correlated-factor model proposed by te Marvelde et al. (2006), the measurement part of which can be written as Equation 2.2. The structural part is a correlated structure. The path diagram for this model is shown in Figure 2.2. te Marvelde et al. (2006) claimed that the correlations between factors should account for the conditional dependence of repeated items. However, some other researchers observed that the correlations between the factors should be viewed as a measure for the stability of the measured construct (Cai, 2010b; Cudeck & Harring, n.d.). The dependence of repeated items between occasions could persist even after the examinees' latent scores at different time points are correlated. Other techniques are needed to address the conditional dependence.

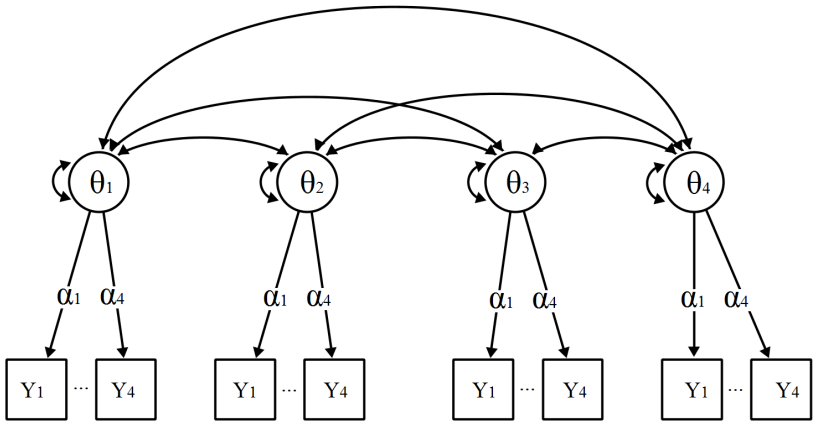

Figure 2.2. Simple structure correlated-factor model for longitudinal data by te Marvelde et al. (2006). This example includes four time points and four repeated items. Measurement invariance across time is assumed.

When items are scored on a continuous rating scale, conditional dependence is typically addressed by using correlated residuals (see e.g., Hancock et al., 2001; McArdle, 1988). With categorical items, bifactor-type models are often used to account for conditional dependence with orthogonal specific factors (e.g., Cai, Yang, & Hansen, 2011; Gibbons & Hedeker, 1992; Hill, 2006; Rijmen, 2009) in the framework of full-information item factor analysis (IFA) (Gibbons & Hedeker, 1992; Takane & De Leeuw, 1987). Hill (2006) first proposed a MIRT model with common item effects to address the issue of item local dependence with repeated measures data. However, Hill (2006) found that this model was too complex to be practically useful due to its high dimensionality. Building on the work of Hill (2006), Cai (2010b) formalized this model as the "two-tier full-information item factor model." Cai (2010b) further found that the two-tier structure allows analytical dimension reduction technique (Gibbons & Hedeker, 1992) to be implemented, which greatly alleviated the

estimation burden of the model. The model is described in detail here.

In the two-tier model, each item and its repeated duplicates at different time points are seen as a bundle. If a test with I items is repeated over time, all the items across time can be sorted into I bundles. In addition to the correlated main factors, the two-tier IFA model introduced I orthogonal specific factors (one for each bundle). Each bundle of items can only load on their specific factor and the main factor. Measurement invariance across occasions is assumed. The loadings of items within a bundle on their specific factor can be constrained to be equal to those on the main factor. No cross loading is allowed. The two-tier model could be represented as:

$$P(Y_{inj} = 1 \mid \theta_{nj}, \xi_{ij}) = \frac{1}{1 + \exp(-\alpha_i \theta_{nj} - \alpha_i \xi_{ij} - \lambda_i)}, \tag{2.5}$$

where ξ_{ij} is the common item effect for person j on item i. A vector of the common item effects $\boldsymbol{\xi}_j$ are from I independent normal distributions. The term $\alpha_i \xi_{ij}$ is the common item effect for item i and person j. It has also been referred to as an "item-specific random effect" (Jeon & Rabe-Hesketh, 2015). In the two-tier model, each item would have one intercept and two equal slopes, one for the main factor and the other for its specific factor. By comparing Equation (2.8) and (2.5), it is easy to see that the two-tier IFA model is a more general form of the simple structure MIRT model by te Marvelde et al. (2006). The two-tier model not only accounts for the correlation of latent scores at different time points, but also the conditional dependence of the repeated items. The path diagram of the two-tier

model is presented in Figure 2.3.

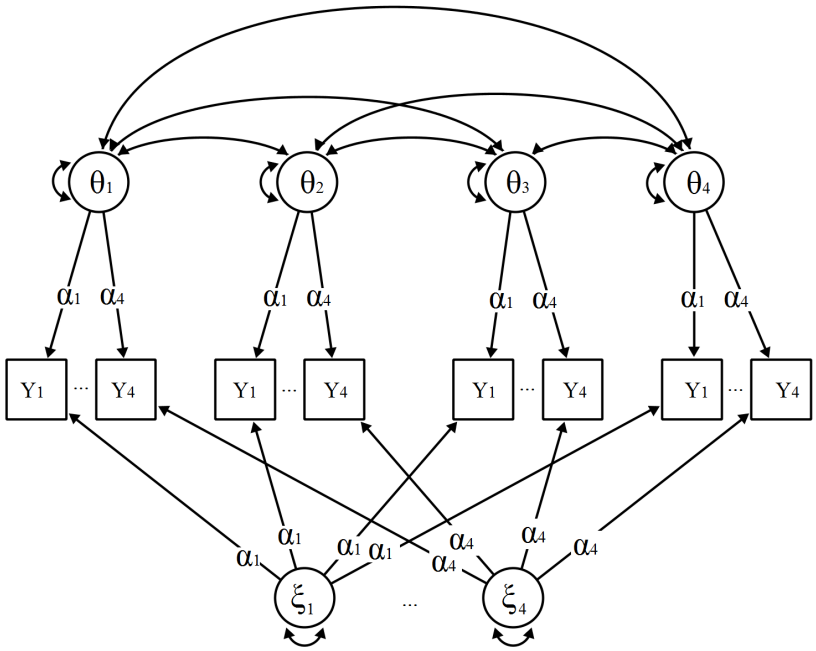

Figure 2.3. Two-tier item factor model for longitudinal data by Cai (2010b). This example includes four time points and four repeated items. Measurement invariance across time is assumed.

One advantage of the two-tier model is that it can utilize the dimension reduction technique (Gibbons & Hedeker, 1992) to greatly reduce estimation burden during FIML estimation via the FIML-BAEM algorithm. Using the dimension reduction techniques, the conditional distribution of an examinee's response pattern given the model parameters and his factor scores can be written as:

$$f(\mathbf{Y}_j|\boldsymbol{\omega},\boldsymbol{\theta}_j,\boldsymbol{\xi}_j) = \int_{\mathbb{R}^N}\int_{\mathbb{R}^S} f(\mathbf{Y}_j,\boldsymbol{\theta}_j,\boldsymbol{\xi}_j)d\boldsymbol{\xi}_j d\boldsymbol{\theta} = \int_{\mathbb{R}^N}\prod_{s=1}^{S}\int_{\mathbb{R}}\left[\prod_{i\in\mathcal{I}_s} f(Y_{ij}|\boldsymbol{\theta}_j,\xi_{js},\boldsymbol{\omega})f(\xi_{js}|\boldsymbol{\omega})d\xi_{js}\right] f(\boldsymbol{\theta}_j|\boldsymbol{\omega})d\boldsymbol{\theta}_j, \tag{2.6}$$

where $\mathbf{Y}_j$ is the observed response data for person j, $\boldsymbol{\theta}_j$ is a length N vector of primary factors, $\boldsymbol{\xi}_j$ is a length S vector of specific factors, $\boldsymbol{\omega}$ represents the model

parameters, and $\mathcal{I}_s$ is the index for items that load on specific dimension s. The likelihood of the model can then be written as:

$$L(\mathbf{Y}|\boldsymbol{\omega},\boldsymbol{\theta},\boldsymbol{\xi}) = \int_{\mathbb{R}^N} \prod_{s=1}^{S} \int_{\mathbb{R}} \left[\prod_{j=1}^{J} \prod_{i \in \mathcal{I}_s} f(Y_{ij}|\boldsymbol{\theta}_j, \xi_{js}, \boldsymbol{\omega}) f(\xi_{js}|\boldsymbol{\omega}) d\xi_{js} \right] f(\boldsymbol{\theta}_j|\boldsymbol{\omega}) d\boldsymbol{\theta}_j. \quad (2.7)$$

As shown in the above two equations, the $(N + S)$-dimensional integration is reduced to $(N + 1)$-dimensional integration. For example, a two-tier model with two time points only requires three-dimensional integration regardless of the number of repeated item. Even though dimension reduction technique can be used in the two-tier model to alleviate some estimation burden, estimating the two-tier model via FIML-BAEM algorithm could still be a major issue when the number of testing occasions N increases. Therefore two-tier full-information IFA model might not be suitable for situations where there are a large number of repeated measurements. Estimation methods with alternative computation algorithms can be explored in situations with a large number of time points.

Despite their differences, one common feature of the three longitudinal MIRT models is that they do not assume any smooth growth curve function. Instead, the latent abilities are estimated individually for each person at each time point. The advantage is that the model could accommodate complex and atypical growth processes.

2.1.3 Second-Order Latent Growth Curve Models

More recently, some researchers began to explore applying latent growth modeling (LGM) techniques to model the change in examinees' latent construct, using IRT as the measurement model (Jeon & Rabe-Hesketh, 2015; Paek et al., 2016; Wang et al., 2016). Although combining LGM and IRT is a relatively new strategy, the theory of the second-order LGM has been utilized for some time to model observed continuous indicators in the literature of SEM (Duncan & Duncan, 1996; Hancock & Buehl, 2008; Hancock et al., 2001; McArdle, 1988; Sayer & Cumsille, 2001). In a second-order LGM, a latent variable (i.e., the first-order factor) is measured by multiple continuous indicators at each time point (Sayer & Cumsille, 2001). A second-order factor structure is then imposed on the first-order factors to separate the variances that are associated with growth from the time-specific disturbance variances that are not related to growth (Sayer & Cumsille, 2001). Change can be estimated for the first-order latent variables instead of the individual observed continuous indicators to reduce measurement errors.

An LGM-IRT can be seen as an extension of a second-order LGM to categorical observed indicators in the SEM framework. The first-order variable are measured by categorical indicators with an IRT model. The variances/covariances of the latent variables are then decomposed with a second-order structure. In the IRT framework, the LGM-IRT can also be seen as an MIRT model, whose correlated factors structure is replaced by a higher-order growth structure to capture the change process of the time-specific factors. For example, if the change process of examinees is assumed to

be linear, the common structure model can be expressed as:

$$\theta_{nj} = \gamma_{0j} + \gamma_{1j} t_{nj} + \epsilon_{nj}, \tag{2.8}$$

where θ_{nj} is the person- and time-specific latent ability, γ_{0j} and γ_{1j} are the random intercept and slope of the person j respectively, and ϵ_{nj} is the person-specific disturbance. It can be observed that the structural model under the SEM framework in Equation (2.8) is mathematically equivalent to Equation (2.3) under multilevel modeling framework. The LGM model can be seen as a multilevel model formulated as a single-level model. For discussion on the equivalence of two types of models, readers are referred to the works of Chou, Bentler, and Pentz (1998), and Bauer (2003) among others.

There are several advantages of the LGM-IRT model as opposed to the MIRT model that it is extended from. First, the LGM-IRT model is more parsimonious than its corresponding MIRT model. In a longitudinal study with N time points ($N \geq 3$), the N variances, $N \times (N-1)/2$ covariances and N means in the corresponding MIRT model are summarized with only five parameters in the LGM-IRT (i.e., the means of the intercepts and slopes, the variances of the intercepts and slopes, and the covariance between the intercepts and slopes). Second, the random intercept and slope allows researchers to examine individual differences in initial status and growth rate as they are modeled as random effects. Third, by estimating the covariance between the random slopes and intercepts, the LGM-IRT model could answer the specific question of "what is the correlation between initial status

and growth rates?" This can be an important piece of information for applied researchers, because whether a person with lower initial status changes faster or slower than one with higher initial status may have very different policy implications. It is worth noting that the latent intercept of the LGM does not necessarily have to be the first time point. It can be parameterized at any time point arbitrarily based on the purpose of the research. Fourth, time-invariant or time-specific covariates can be incorporated into the model.

Depending on the type of IRT model used in the measurement part, this study identifies three types of second-order LGM-IRT models, namely LGM-IRT without local dependence considerations, LGM-IRT with order local dependence, and LGM-IRT with common item effects. The three models are presented here respectively. Paek et al. (2016) proposed a second-order LGM-IRT using a Rasch model (Rasch, 1960). If the equal-discrimination constraint is relaxed, the more general form of the model is presented in the upper panel of Figure 2.4. Even though the original model was formulated as a first-order model, the model could be transformed to its second-order equivalent as shown in the lower panel of Figure 2.4. It can also be observed from the first-order path diagram that the model only requires three-dimensional integration after dimension reduction technique (Gibbons & Hedeker, 1992) is implemented, as there are no cross loadings of the disturbances. If the growth curve is modeled as a quadratic function, a quadratic growth factor will be added to the structural part of the model. The number of dimensions would increase by 1. The model would then require four-dimensional integration after dimension reduction.

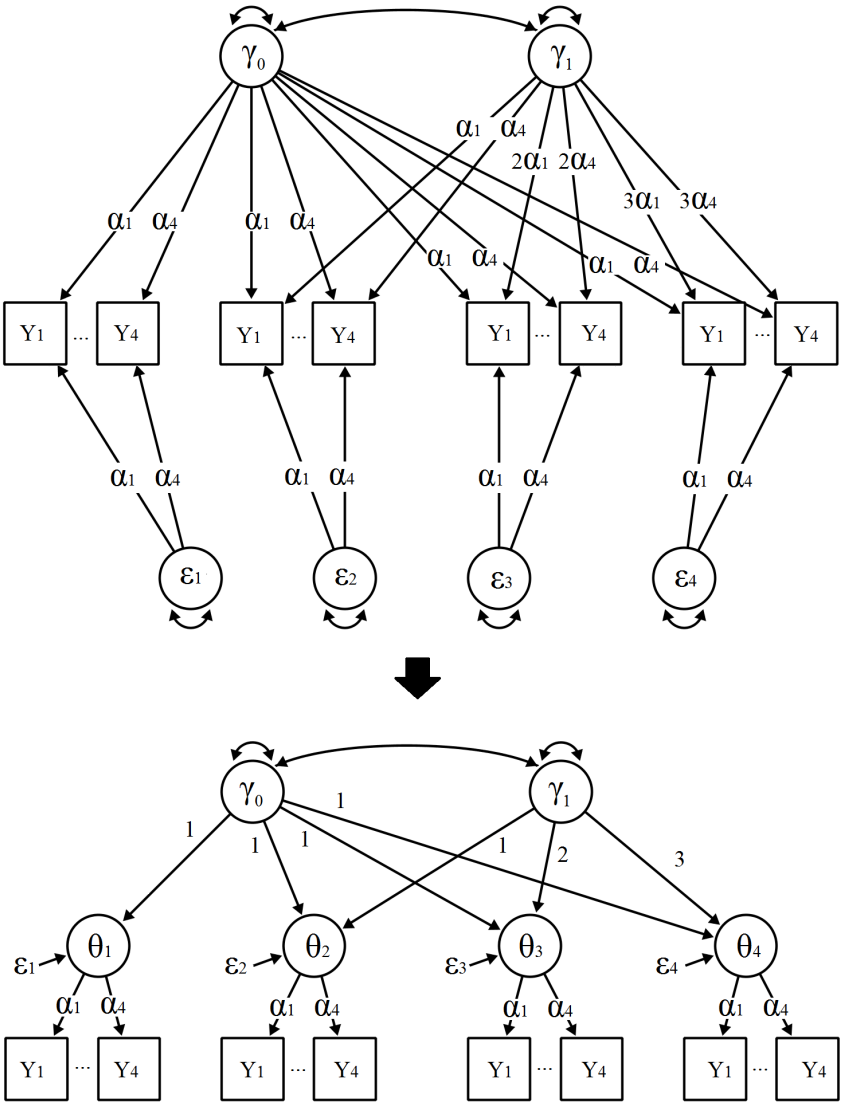

Figure 2.4. Second-order LGM-IRT model without local dependence consideration by Paek et al. (2016). This example includes four time points and four repeated items. The upper panel is based on the original parameterization in MIRT framework. γ_0 and γ_1 are the latent intercept and slope respectively. ϵ_n is the time-specific disturbance. The model can be transformed to the second-order structure as shown in the lower panel.

Although developed under a different framework, the LGM-IRT by Paek et al. (2016) is in fact mathematically equivalent to the longitudinal IRT model developed under the multilevel modeling framework described earlier, as both the models do not have local dependence considerations. The model could also be seen as imposing the latent growth structure on the correlated factors of the simple structure MIRT model by te Marvelde et al. (2006). Similar to the MIRT by te Marvelde et al. (2006), the major weakness of the model is that it lacks local dependence consideration, which may lead to bias in item and/or structural parameters.

Wang et al. (2016) extended the two-tier model to a LGM-IRT model where the local dependence is modeled as common item effects. The measurement part of the model is the same with the aforementioned two-tier model (see Equation (2.5)) and is not repeated here. In contrast to order local dependence, the item local dependence in this model is modeled with random effects, which means that the dependence among items could vary across examinees. Thus, common item effects could be seen as the unique interactions between examinees and items. According to Hoskens and De Boeck (1997), local item dependence parameterized as random effects can be conceptualized as a type of symmetrical "combination dependence" as opposed to order local dependence. Hoskens and De Boeck (1997) argued that combination dependence is most suitable for those items that tap on partial aspects of the same knowledge or items that share the same stimulus content.

The path diagram of the second-order LGM-IRT with common item effects is presented in the upper panel of Figure 2.5. The second-order structure can also be transformed to first-order formation as shown in the lower panel of Figure 2.5.

As we can see in the first-order path diagram, the nuisance factors for the common item effects and the time-specific disturbances are crossing with each other. As a result, the dimension reduction technique can only reduce the dimensionality of the model to four, which would still be computationally demanding using FIML-BAEM. Moreover, currently there is no statistical package that implement dimension reduction for this model. Without dimension reduction, the number of dimensions equals $I + N$ in the second-order parameterization and $I + N + 2$ using the first-order formation. Therefore in a study with many repeated items and time points, the application of this model is greatly limited. In fact, the full LGM-IRT model with common item effects was not used in the empirical study by Wang et al. (2016). The authors resorted to dropping all the specific factors in the model to reduce the estimation difficulty. Wang et al. (2016) argued that the simplification in model did not cause significant bias, citing a pilot test they had conducted. However, the effects of ignoring item local dependence on item and structural parameters have not yet been comprehensively examined through simulation studies.

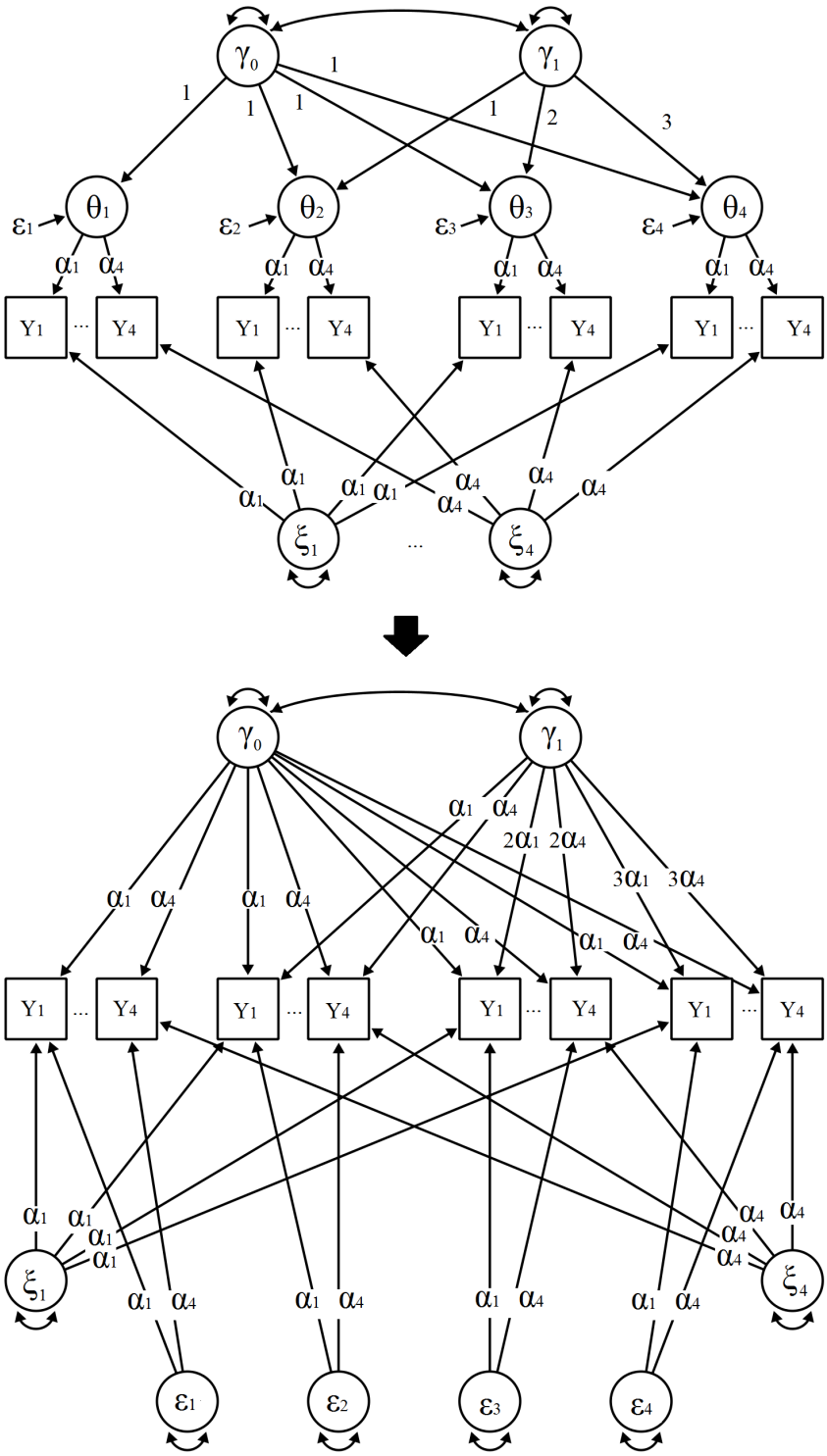

Figure 2.5. Second-order LGM-IRT model with common item effects to address local dependence. This example includes four time points and four repeated items. ϵ_n and ξ_i are the disturbance and common effect respectively. The second-order structure is shown in the upper panel while the first-order structure is shown in the lower panel.

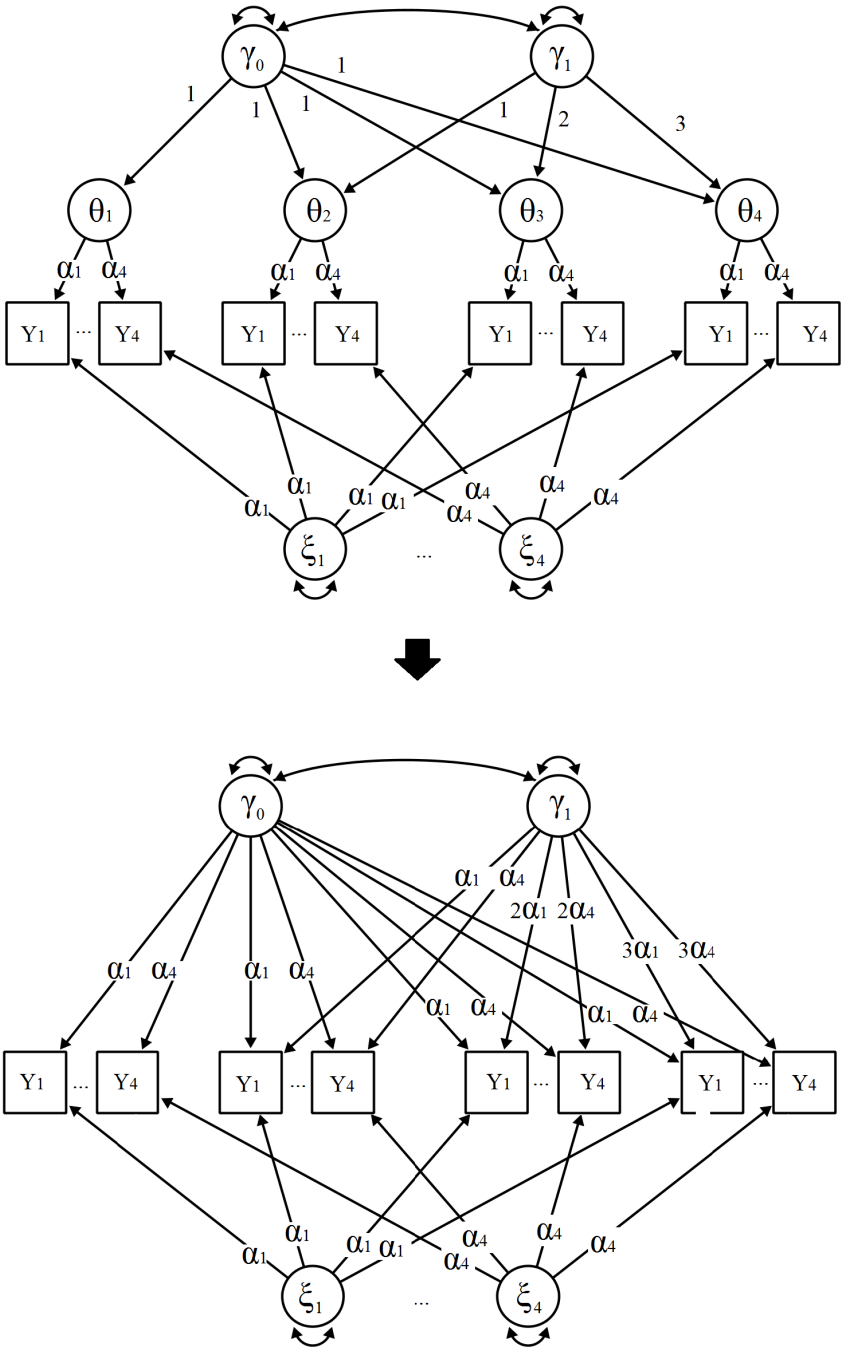

Figure 2.6. Second-order LGM-IRT model with common item effects but without time-specific disturbances. This example includes four time points and four repeated items. ξ_i is the common item effect. The second-order structure is shown in the upper panel while the first-order structure is shown in the lower panel.

Similar to dropping the common item effects as shown in Figure 2.4, the

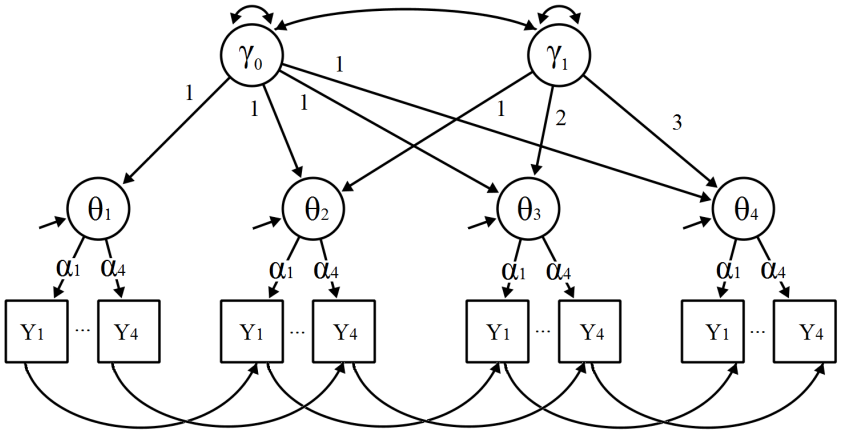

Figure 2.7. Second-order LGM-IRT model with order local dependence by Jeon and Rabe-Hesketh (2015). This example includes four time points and four repeated items. γ_0 and γ_1 are the latent intercept and slope respectively. The intercept of an item is regressed on the score of the item in previous administration.

cross-loading issue in the LGM-IRT with common item effects can also be resolved by leaving out the time-specific disturbances. As shown in Figure 2.6, the model without the time-specific disturbances requires only three-dimensional integration as well. The structural part of this reduced model is:

$$\theta_{nj} = \gamma_{0j} + \gamma_{1j} t_{nj}. \tag{2.9}$$

However, it is uncommon to ignore time-specific disturbances in the structural part of a LGM model. This reduced model without disturbances has not been used by scholars as far as the author is aware of. The effects of eliminating disturbances are not yet examined.

Approaching the issue of local dependence from a different perspective, Jeon

and Rabe-Hesketh (2015) proposed adding "order local dependence" in the measurement part of the LGM-IRT to address the local dependence issue. The order local dependence IRT model was first proposed by Hoskens and De Boeck (1997), where a person's score of a previous item could impact the difficulty of the subsequent item. The measurement model of the LGM-IRT with order local dependence for can be written as:

$$P(Y_{inj} = 1 \mid \theta_{nj}) = \frac{1}{1 + \exp(-\alpha_i \theta_{nj} + \lambda_i + \delta_i Y_{i(n-1)j})}, \tag{2.10}$$

where $\delta_i Y_{i(n-1)j}$ is an adjustment of the item intercept based on previous response, with δ_i being the fixed regression coefficient for item i. The path diagram of this model is presented in Figure 2.7. Intuitively, order local dependence means that when the score for item i in the $(n-1)^{th}$ wave is 1, the difficulty of the same item at the n^{th} is reduced by $-\delta_i/\alpha_i$ accordingly. If the score for item i in the $(n-1)^{th}$ wave is 0, the difficulty of the item stays the same in the n^{th} wave. It should be noted that, the order local dependence is a feature of the item itself, which is modeled as a fixed regression coefficient for each item. The regression coefficient δ_i is assumed to be exactly the same for all examinees.

As Hoskens and De Boeck (1997) pointed out, order local dependence is especially suitable for modeling the ordering effects of items. The "order" here can be factual (i.e., the administration order of the items), historical (i.e., order of the learning process) or conceptual (i.e., mastering the more difficult knowledge implies knowing the easier contents).

2.2 Estimation of LGM-IRT

Three estimation methods in the FIML and limited-information estimation frameworks are reviewed in this section. Under the FIML estimation framework, two methods (FIML-MCEM and FIML-MH-RM) are introduced. Under the limited-information estimation framework, various weighted least squares (WLS) estimation methods including weighted WLS, diagonal-weighted WLS, and unweighted WLS are described.

It is worth noting that Bayesian estimation method is not included in this research. The Bayesian methods avoid using quadratures by drawing samples of a target distribution from single chains of events. Edwards (2010) examined the feasibility of using Bayesian methods for high-dimensional IRT models. It was found that the difficulty in estimating a high-dimensional model cannot be avoided with Bayesian methods. Various Bayesian estimation algorithms were even slower than maximum-likelihood estimation methods in complex high-dimensional IRT models (Edwards, 2010).

2.2.1 Full Information Maximum Likelihood Estimation

Under the FIML estimation framework, two estimation methods have been specifically designed to handle high-dimensional integration problems, namely the Monte Carlo EM algorithm (Wei & Tanner, 1990) and the Metropolis Hastings-Robbins Monro algorithm (Cai, 2010a).

Monte Carlo EM algorithm. The FIML-MCEM algorithm (Wei & Tanner,

1990) is an implementation of the EM algorithm pioneered by Dempster, Laird, and Rubin (1977). The EM algorithm is an iterative procedure to find maximum-likelihood estimates in the presence of unobserved data and parameters (e.g., random effects). It is comprised of the expectation (E) step and the maximization (M) step. The general EM algorithm is first described here, followed by details on the implementation of the MCEM algorithm.

For the purpose of description, the following notation are used. $\mathbf{Z}_o$ and $\mathbf{Z}_m$ are the observed and unobserved data respectively. In the context of IRT, $\mathbf{Z}_o$ would be the observed item responses, and $\mathbf{Z}_m$ would be the unobserved latent variables. Together they form the complete data $\mathbf{Z}$. $\boldsymbol{\omega}$ is a vector of the model parameters to be estimated. $l(\boldsymbol{\omega}|\mathbf{Z}_o, \mathbf{Z}_m)$ and $l(\boldsymbol{\omega}|\mathbf{Z}_o)$ are the complete-data and observed-data log-likelihood for the IRT model. $\boldsymbol{\omega}^{(k)}$ are item parameter estimates at the k^{th} iteration of the EM algorithm. In order to obtain estimates that maximize $l(\boldsymbol{\omega}|\mathbf{Z}_o)$, the expectation of $l(\boldsymbol{\omega}|\mathbf{Z}_o, \mathbf{Z}_m)$ must be first computed in the E step. Further define $L(\boldsymbol{\omega}, \boldsymbol{\omega}^{(k)})$ as the expected complete-data log-likelihood conditioned on item estimates at the k^{th} iteration. The E step of the $(k+1)^{th}$ iteration can then be expressed as:

$$L_{k+1}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(k)}) = \int l(\boldsymbol{\omega}|\mathbf{Z}_o, \mathbf{Z}_m) f(\mathbf{Z}_m|\boldsymbol{\omega}^{(k)}, \mathbf{Z}_o) d\mathbf{Z}_m, \tag{2.11}$$

where $f(\mathbf{Z}_m|\boldsymbol{\omega}^{(k)}, \mathbf{Z}_o)$ is the posterior distribution of $\mathbf{Z}_m$. Parameters are updated by maximizing the function 2.11 The E step and the M step are updated iteratively until some arbitrary convergence criterion is met.

The FIML-MCEM algorithm implements the EM procedure by reformulating the E step for models requiring high-dimensional integrations. In the k^{th} iteration, the MCEM algorithm randomly draws a sample of Q_k vectors of quadrature points from the current conditional distribution $f(\mathbf{Z}_m|\boldsymbol{\omega}^{(k)}, \mathbf{Z}_o)$. The sample is denoted as $(\mathbf{Z}_1^{(k)} \dots \mathbf{Z}_{Q_k}^{(k)})$. In the context of MIRT, each $\mathbf{Z}_q^{(k)}$ is a vector of quadrature points sampled from a multivariate Gaussian density. Using the random draws, the current expectation of complete-data log-likelihood can be approximated as the same average of complete data log-likelihood as:

$$\widetilde{L}_{k+1}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(k)}) = \frac{1}{Q_k} \sum_{q=1}^{Q_k} l(\boldsymbol{\omega}|\mathbf{Z}_o, \mathbf{Z}_q^{(k+1)}). \tag{2.12}$$

The FIML-MCEM updates the reformulated E step and the M step until the change in parameters falls below the convergence criterion. In each iteration of the E step, a new sample of vectors will be drawn.

As Cai (2010a) pointed out, despite being designed for solving high-dimensional problems, there are two weaknesses associated with the FIML-MCEM algorithm that can hinder the application of the algorithm. First, in order to achieve stable parameter estimates, the total number of random draws must increase tremendously as the log-likelihood approaches a maximum. Furthermore, the use of random draws is not efficient in FIML-MCEM, as the algorithm discards all the random draws from previous E step. Due to these weaknesses, the FIML-MCEM algorithm in practice is often very computationally expensive when estimating high-dimensional models.

Metropolis Hastings-Robbins Monro algorithm. FIML-MH-RM (Cai,

2010a) is anther algorithm to handle high-dimensional problems to obtain FIML estimates. One motivation for MHRM is Fisher's Identity (Fisher, 1925). Fisher (1925) proved that the gradient of the observed-data log-likelihood could be obtained from the conditional expectation of the complete-data gradient. Let $\nabla(\boldsymbol{\omega}|\mathbf{Z}_o, \mathbf{Z}_m)$ and $\nabla(\boldsymbol{\omega}|\mathbf{Z}_o)$ be the gradients of the complete-data log-likelihood and observed-data log-likelihood, respectively. Following the previous notation, the Fisher's Identity can be written as:

$$\nabla(\boldsymbol{\omega}|\mathbf{Z}_o) = \int \nabla(\boldsymbol{\omega}|\mathbf{Z}_o, \mathbf{Z}_m) f(\mathbf{Z}_m|\boldsymbol{\omega}, \mathbf{Z}_o) d\mathbf{Z}_m, \tag{2.13}$$

where $f(\mathbf{Z}_m|\boldsymbol{\omega}, \mathbf{Z}_o)$ is the posterior-predictive distribution of $\mathbf{Z}_m$. MHRM seeks to find the FIML estimates that make the right-hand side of Equation (2.13) equal zero. The algorithm has three stages, namely stochastic imputation, stochastic approximation and Robbins-Monro update.

In the stochastic imputation stage, Q_k sets of missing data are imputed from the conditional distribution of the missing data $f(\mathbf{Z}_m|\boldsymbol{\omega}^{(k)}, \mathbf{Z}_o)$ via a Metropolis-Hasting sampler (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). The imputed data are denoted as $(\mathbf{Z}_1^{(k)} \dots \mathbf{Z}_{Q_k}^{(k)})$. In the stochastic approximation stage, the gradient and the information of the complete-data log-likelihood are approximated. Let $\mathbf{s}$ stand for the complete-data gradient. $\mathbf{s}$ at the $(k+1)^{th}$ iteration can be approximated as:

$$\widetilde{\mathbf{s}}_{k+1} = \frac{1}{Q_k} \sum_{q=1}^{Q_k} \nabla(\boldsymbol{\omega}^{(k)}|\mathbf{Z}_o, \mathbf{Z}_q^{(k+1)}) \tag{2.14}$$

Let $\mathbf{H}(\boldsymbol{\omega}|\mathbf{Z})$ be the complete-data information matrix, which is the negative Hessian of the complete-data log-likelihood. The conditional expectation of the complete information $\boldsymbol{\Gamma}$ at the $(k+1)^{th}$ iteration can be written as:

$$\boldsymbol{\Gamma}_{k+1} = \boldsymbol{\Gamma}_k + \iota_k \left\{ \frac{1}{Q_k} \sum_{q=1}^{Q_k} \mathbf{H}(\boldsymbol{\omega}^{(k)}|\mathbf{Z}_o, \mathbf{Z}_q^{(k+1)}) - \boldsymbol{\Gamma}_k \right\}, \tag{2.15}$$

where ι_k is the gain constant that slowly decreases to zero over iterations. In the Robbins-Monro update stage, new parameters are obtained using Robbins-Monro algorithm (Robbins & Monro, 1951):

$$\boldsymbol{\omega^{k+1}} = \boldsymbol{\omega^{k}} + \iota_k(\boldsymbol{\Gamma}_{k+1}^{-1} \widetilde{\mathbf{s}}_{k+1}). \tag{2.16}$$

The FIML-MH-RM algorithm stops when some convergence criterion is met for the parameter estimates.

The standard error estimates can be obtained using the approximated observed-data information matrix. The standard errors can be approximated either as a byproduct of the algorithm (i.e., recursive approximation) or by using the method of Louis (1982), where the observed information matrix is approximated via Monte Carlo simulation. Using the same notation, the observed-data information can be written using Louis' method as:

$$\begin{aligned} \mathbf{I}_{\mathbf{Z}o} = E\ \mathbf{H}(\boldsymbol{\omega}, \mathbf{Z})|\mathbf{Z}] - E\left[\mathbf{s}(\boldsymbol{\omega}, \mathbf{Z})\mathbf{s}^T(\boldsymbol{\omega}, \mathbf{Z})|\mathbf{Z}\right] \\ + E\left[\mathbf{s}(\boldsymbol{\omega}, \mathbf{Z})|\mathbf{Z}\right] E\left[\mathbf{s}^\intercal(\boldsymbol{\omega}, \mathbf{Z})|\mathbf{Z}\right]. \end{aligned} \tag{2.17}$$

$E[\mathbf{s}(\boldsymbol{\omega}, \mathbf{Z})|\mathbf{Z}]$ is 0 when $\boldsymbol{\omega}$ is evaluated at its maximum-likelihood estimates. The remaining two terms can be approximated by drawing samples of imputed missing data.

As for the performance of FIML-MH-RM, Cai (2010a) demonstrated via a simulation study that FIML-MH-RM and FIML-BAEM are comparable in terms of accuracy in a two-dimensional IRT model, while FIML-MH-RM is considerably more computationally efficient. Bashkov (2015) showed that the FIML-MH-RM algorithm yielded unbiased parameter estimates in two-level MIRT models with 10 dimensions. Han and Paek (2014) showed that FIML-MH-RM yielded comparable results with FIML-MCEM and FIML-BAEM, while FIML-MH-RM required much less estimation time than FIML-MCEM and FIML-BAEM in three-dimensional IRT models. However, the number of dimensions is typically larger than ten in the LGM-IRT model with common item effects. It is not yet clear whether the full model could be stably estimated with either algorithm.

2.2.2 Multiple-Step Limited Information Estimation

WLS estimators for categorical data were developed under the general framework of limited-information estimation for categorical data (e.g., Christoffersson, 1975; McDonald, 1982; Muthén, 1978; Muthén et al., 1997). The rationale for limited-information estimation is to obtain parameter estimates that minimize the differences between the model-implied and the observed sample thresholds and tetrachoric/polychoric correlations. WLS estimators under this framework only use first-

and second-order associations among observed categorical data for estimation via a multi-stage procedure. The steps of WLS estimation are presented below.

The ordered response categories are assumed to be discrete indicators of a continuous underlying latent construct. let $x_i^* j$ be the latent score of respondent j for item i. Let x_{ij} denote the respondent's score. let k_i ($k_i = 0, 1, \ldots, K_i$)denote the response categories of the item. The thresholds for the item, $\boldsymbol{\tau}_i$, are defined as:

$$x_{ij} = k_i, \text{if } \tau_{ik_i} < x_i^* j < \tau_{i(k_i+1)}, \tag{2.18}$$

where $\tau_{i0} = -\infty$ and $\tau_{iK_i} = \infty$. In the first step of limited-information estimation, the thresholds $\boldsymbol{\tau}$ for the categories are estimated one item at a time using a probit regression via maximum-likelihood estimation. Only univariate information is used in this stage. In the second stage, bivariate information is utilized. The tetrachoric/polychoric correlations $\boldsymbol{\rho}$ are estimated with maximum likelihood for each pair of items using $\hat{\boldsymbol{\tau}}$ from the previous stage. The estimated thresholds and correlations are stored in a vector $\mathbf{s}$. In the third stage, a least squares function $\mathbf{F}$ is computed. The general form of $\mathbf{F}$ can be expressed as:

$$\mathbf{F} = (\mathbf{s} - \boldsymbol{\sigma})^{\intercal} \mathbf{W} (\mathbf{s} - \boldsymbol{\sigma}), \tag{2.19}$$

where $\boldsymbol{\sigma}$ is the model-implied intercepts and correlations, the $^{\intercal}$ superscript is the transpose operator, and $\mathbf{W}$ is a weight matrix. The model parameters and standard errors can be obtained by minimizing the least squares function.

As Forero and Maydeu-Olivares (2009) pointed out, there have been some confusion about the different types of least squares estimators. They observed that existing limited-information estimators are typically differentiated by the type of weight matrix, $\mathbf{W}$, used in the third stage. Let $\mathbf{\Sigma}$ be the covariance matrix of the estimated thresholds and correlations. When $\mathbf{W} = \mathbf{\Sigma}^{-1}$, the estimator is the weighted least squares method by Muthén (1978, 1984). The diagonal matrix is used here in order to ease computational burden and circumvent the issue of non-invertibility of $\mathbf{\Sigma}$. When $\mathbf{W} = [diag(\mathbf{\Sigma})]^{-1/2}$, the estimator is the diagonally weighted least squares (DWLS) by (Muthén et al., 1997). When $\mathbf{W}$ is an identity matrix, the estimator is the unweighted least squares in Muthén (1993). The popular WLSMV implemented in Mplus 7.4 (Muthén & Muthén, 1998-2012), which will be used in the simulation studies, is actually the DWLS (Forero & Maydeu-Olivares, 2009) for point estimates. The χ^2 test statistic is adjusted to approximate the mean and variance of the expected χ^2 when the model is correctly specified.

The limited-information estimators circumvent numerical integration associated with maximum likelihood estimation to achieve computational efficiency. However, as Wirth and Edwards (2007) argued, the computational ease of limited information estimation is achieved by ignoring higher-order associations among items. In theory, limited information estimation is less ideal than FIML methods due to the omission of high-order information. Forero and Maydeu-Olivares (2009) conducted a comprehensive comparison study of FIML and limited-information estimation methods for categorical data under various data conditions. They found that the differences between the performance of these methods were very small, and that

FIML was slightly less biased when the sample size was small (200). However, the study did not consider missing data, which might influence the performance of the estimation methods. The missing data issue is discussed in the following section.

2.3 Longitudinal Studies and Sample Attrition

Longitudinal analyses, in general, often suffer from some degree of panel attrition, when respondents fail to participate in certain waves of the survey or permanently drop out of the study. In a review of the attrition issue in 11 large-scale surveys in developed countries, Lee (2003) reported that the second-wave attrition rate could range from 4.3% to 15.3%, while the total attrition rate at the end of the survey varied from 14.8% to 51% depending on the survey and the number of waves.

2.3.1 Effects of Attrition on Estimation

In general, a considerable attrition rate can cause two methodological challenges for longitudinal analysis. The most apparent effect of attrition is the loss of statistical power due to the decrease in sample size. Moreover, attrition under different missing mechanisms could further complicate the statistical inference of the model and the performance of the estimation methods used. This section first describes three types of attrition mechanisms. The effects of the attrition mechanisms on the performance of the three estimation methods are then discussed.

In the literature of missing data (see: Rubin, 1976), distinctions are made be-

tween three kinds of attrition mechanisms, namely missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The three mechanisms are briefly presented here. Let $P(\mathbf{A})$ be the probabilities of the realization of attrition patterns $\mathbf{A}$, $\mathbf{Y}_o$ be the values of the observed dependent variables, $\mathbf{Y}_m$ be the values of the missing dependent variables, and $\mathbf{X}$ be the observed covariates. $\mathbf{Y}_o$ and $\mathbf{Y}_m$ together form the complete item response matrix $\mathbf{Y}$. When the attrition occurs under MCAR, the missing is independent of all observed and unobserved data. MCAR can be expressed as: $P(\mathbf{A}|\mathbf{Y}_o, \mathbf{Y}_m, \mathbf{X}) = P(\mathbf{A})$. With sufficient sample size after missing, MCAR is usually not a threat to the validity of the model's the statistical inference and can be ignored. Unfortunately, attrition under MCAR is not common in longitudinal studies (Young & Johnson, 2015). Most attrition in longitudinal surveys is due to a mix of MAR or MNAR.

When the missing mechanism is MAR, the attrition is only dependent on observed data: $P(\mathbf{A}|\mathbf{Y}_o, \mathbf{Y}_m, \mathbf{X}) = P(\mathbf{A}|\mathbf{Y}_o, \mathbf{X})$. It has been well established that FIML estimation methods (e.g., FIML-MCEM and FIML MHRM) are consistent under MAR assuming all relevant covariates are included in the model (see e.g., Enders, 2001). Under limited-information estimation, estimates are obtained by pairwise deletion in the first and second stages of the WLS estimation when missing data are present (Asparouhov & Muthén, 2010). According to Asparouhov and Muthén (2010), a limited information estimator such as DWLS was not consistent under MAR with respect to both covariates and observed outcomes. Asparouhov and Muthén (2010) defined a more restrictive MAR mechanism, where missing is conditioned on the observed covariates $\mathbf{X}$ but not observed out-

comes $\mathbf{Y}_o$ (referred to as MAR-X per the authors). MAR-X can be represented as: $P(\mathbf{A}|\mathbf{Y}_o, \mathbf{Y}_m, \mathbf{X}) = P(\mathbf{A}|\mathbf{X})$. Through a simulation study, Asparouhov and Muthén (2010) found that the DWLS estimator was consistent under MAR-X. Although Asparouhov and Muthén (2010) asserted that DWLS was not consistent under the more general MAR, its performance under the broader MAR was not examined in their simulations. The specific effect of MAR on the performance of WLSMV is not yet clear.

The third mechanism, MNAR, occurs when the attrition depends on the unobserved data. MNAR can be expressed as: $P(\mathbf{A}|\mathbf{Y}_o, \mathbf{Y}_m, \mathbf{X}) \neq P(\mathbf{A}|\mathbf{Y}_o, \mathbf{X})$, which indicates that the missingness is associated with some unobserved variables. MNAR is not ignorable could cause invalid statistical inference and is usually not ignorable. In the literature of LGM models, methods have been developed to explicitly model the missing mechanisms when MNAR attrition is suspected to be a potential threat in the longitudinal study (e.g., Diggle & Kenward, 1994; Enders, 2011; Wu & Carroll, 1988). The major problem with these modeling approaches is that they usually rely on untestable assumptions.

As mentioned before, attrition can be addressed using FIML estimation by including covariates as well as relevant auxiliary variables in the model. The auxiliary variables are related to the missingness but not part of the main analysis. The auxiliary variables can be incorporated through the extra dependent variables approach or the saturated correlates approach (Graham, 2003; Stapleton, Harring, & Lee, 2015). FIML estimation methods are consistent under general MAR attrition. Besides using FIML, multiple imputation (Rubin, 1987) methods can also be used to

handle attrition. The method draws multiple sets of plausible values for the missing data from the posterior distribution of the target variables after observed outcomes and auxiliary information are taken into account. The analysis results from the multiple datasets are then pooled to obtain the point and variance estimates. The method is also consistent under MAR. The multiple imputation approach handles the missing data in a separate stage and can use many variables for imputation, while FIML accounts for missing data in a single stage and can only incorporate a limited number of auxiliary variables (Stapleton et al., 2015). Another method to combat attrition is to use sampling weight adjustment, where the sample weights are adjusted based on background variables so that the weighted sample after attrition would still be representative of the target population. For discussion on methods of sample weight adjustment, readers are referred to the work of Stapleton et al. (2015).

Of all the models reviewed above, this research focuses on the LGM-IRT with common item effects. In the following section, the methods for evaluating the estimation methods of the LGM-IRT model are introduced.

Chapter 3: Methods

The methods used to achieve the second to fifth research goals of the research are introduced in this chapter. The first simulation study is to compare the performance of the three aforementioned estimation methods (i.e., FIML-MCEM, FIML-MH-RM, and DWLS) in estimating the full model without sample attrition. The second simulation study is to assess the performance of two reduced models in terms of model parameter bias and confidence interval coverage under complete data. The third simulation study is to compare the aforementioned estimation methods under the MAR-X and under MAR with respect to both observed covariates and outcomes. Finally the data and models used for the empirical illustration are also described.

3.1 Comparison of the estimation methods without Sample Attrition: Simulation I

This section focuses on the simulation design for the comparison of the estimation methods under complete data. The data generation model and the manipulated factors are explained in detail. The rationales for choosing the specific levels of the manipulated factors are also provided.

3.1.1 The Data Generation Model

The second-order LGM model with common item effects (see Figure 2.5 and Equation (2.6)) was used to generate the item response data for the simulation study. To be consistent with Equation (2.6), let $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ be the vectors of latent intercepts and slopes, $\boldsymbol{\theta}_n$ be the vector of the latent abilities at the n^{th} time point, and $\boldsymbol{\epsilon}_n$ be the vector of the disturbances at the n^{th} time point. First, the latent intercepts and slopes were generated from a bivariate normal distribution with covariance matrix

$$\boldsymbol{\Psi} = \begin{bmatrix} \sigma^2_{\gamma_0} & \\ \sigma_{\gamma_0\gamma_1} & \sigma^2_{\gamma_1} \end{bmatrix}, \tag{3.1}$$

and mean vector $\boldsymbol{\mu} = (\mu_{\gamma_0}, \mu_{\gamma_1})^\intercal$. Then the disturbances at each time point were generated from a normal distribution with mean 0 and variance $\sigma^2_{\boldsymbol{\epsilon}_n}$ one time point at a time. The time-specific disturbances were set to be uncorrelated. Thus, the latent abilities $\boldsymbol{\theta}_n$ can be generated using Equation (2.6). In total, four waves of $\boldsymbol{\theta}_n$ were generated in this study. With this setup, the implied variance and covariance matrix of the latent abilities at the four time points were presented at Table 3.1. Growth trajectories of a random sample of 100 people using these generating parameters were plotted in Figure 3.1.

Table 3.1

Implied Variance-Covariance Matrix of Latent Abilities at the Four Time Points using the Generating Structural Parameters

	Time1	Time2	Time3	Time4
Time1	1.5			
Time2	1.2	2.1		
Time3	1.4	2.0	3.1	
Time4	1.6	2.4	3.2	4.5

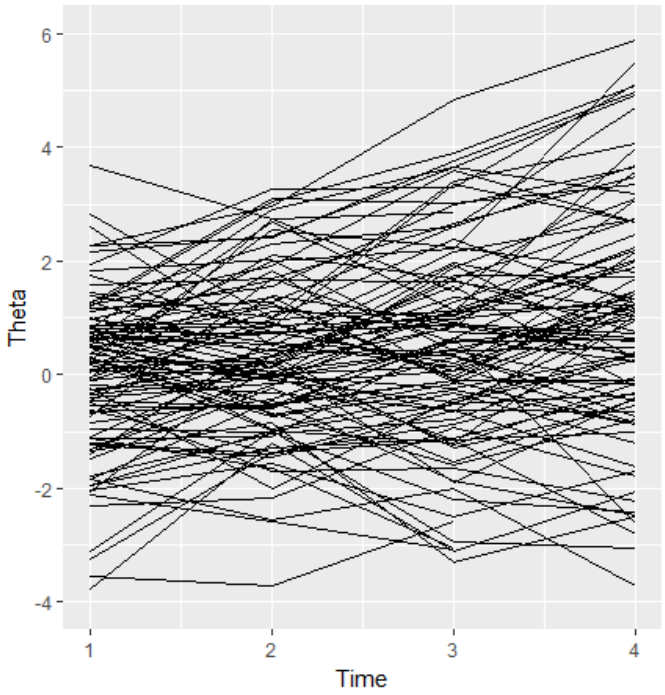

Figure 3.1. Plot of the growth trajectories of a random sample of 100 people using the generating structural parameters. The intercept mean and variance are 0 and 1, respectively. The slope mean and variance are 0.2 and 0.2, respectively. The covariance between the intercepts and slopes is 0.2.

It should be noted that the disturbances could be correlated or have an elaborate covariance matrix. More time points can also be used in practice. However, the covariance matrix and the number of time points were not manipulated in this study due to the scale of the simulation. Uncorrelated disturbance structure and four time points were used in this study as a first step in the investigation of LGM-IRT.

Further, let $\boldsymbol{\xi}_i$ be the common item effects of all examinees for item i generated from a normal distribution with mean 0 and variance $\sigma^2_{\boldsymbol{\xi}_i}$. Taking $\boldsymbol{\theta}_n$ and $\boldsymbol{\xi}_i$ together, the item response data are generated using a five-category graded response model (Samejima, 1969). Let F^*_{injx} be the probability of examinee j scoring x $(x = 0, 1, \ldots, M)$ or above on item i at the nth time point. The model is defined as:

$$F^*_{injx} = F^*(Y_{inj} \geq x \mid \theta_{nj}, \xi_{ij}) = \frac{1}{1 + \exp(-\alpha_i\theta_{nj} - \alpha_i\xi_{ij} - \lambda_{ix})}. \tag{3.2}$$

The examinee's probability of scoring x can be expressed as:

$$F_{injx} = F^*_{injx} - F^*_{inj(x+1)}, \tag{3.3}$$

where the probability for getting a score of 0 or higher is:

$$F^*_{inj0} = 1, \tag{3.4}$$

and the probability for getting a score higher then the maximum point is 0:

$$F^*_{injM} = 0. \tag{3.5}$$

The specific means and variances of the latent variables used for data generation are presented in Table 3.2. The latent intercept was set to follow a normal distribution with mean 0 and variance 1. The average growth rate was set to be 0.2. In choosing the generating latent slope variance value, a few empirical examples

Table 3.2

Generating Values for Latent Variables in Simulation I.

Latent variable parameters	**Values**
Distribution of latent intercepts	$\boldsymbol{\gamma}_0 \sim$ Normal(0, 1)
Distribution of latent slopes	$\boldsymbol{\gamma}_1 \sim$ Normal(0.2, 0.2)
Covariance of latent intercept and slope	$\sigma_{\gamma_0\gamma_1} = 0.2$
Distribution of time-specific disturbances	$\boldsymbol{\epsilon}_n \sim$ Normal(0, $\sqrt{0.5}$)
Distribution of common item effects	$\boldsymbol{\xi}_i \sim$Normal(0,$\sqrt{0.5}$)

in IRT framework are reviewed. The conclusions varied depending on the specific test. Paek et al. (2016) showed that the latent slope variance is about 1/2 of the intercept variance in a cognitive test. Jeon and Rabe-Hesketh (2015) found that the latent slope variance was about 1/4 of latent intercept variance in a psychological survey. Wang et al. (2016) found that the slope variance was almost 0 in a cognitive test. In this research, the variance of the latent growth was set to 1/5 of the intercept variance as this is typically observed in empirical data following the practice of previous simulation studies (Depaoli, 2013; Li, 2015; Liu, 2012). The covariance of latent intercepts and slopes was set to 0.2, which was meant to represent a moderate positive correlation of approximately 0.447 between latent intercepts and slopes. In terms of nuisance factors, the disturbance variance was fixed at 0.50 for all time points. The common item effect variance was also set at 0.50 for all items.

The item parameters of eight items used in this simulation are presented in Table 3.3. These are also the items used in Simulation II and III. The item parameters were selected from the items used in the simulation study of Cai (2010a), which were meant to represent items in real-life situations. These items all have positive and strong discriminations (i.e., slopes) between 1 and 3. The skewness of the items

Table 3.3

Generating Values of Item Parameters in Simulation I, II & III

Item	**Slope**	**Interept 1**	**Interept 2**	**Interept 3**	**Interept 4**	**Skewness**
1	1.18	1.92	1.20	0.55	-0.28	-0.74
2	1.29	1.13	0.99	-0.05	-0.67	-0.27
3	2.17	2.45	0.56	-0.8	-2.53	0.09
4	2.57	1.72	0.18	-1.85	-4.80	-0.05
5	1.64	1.92	1.18	0.93	-0.35	-0.91
6	1.97	0.62	-0.51	-0.68	-1.56	0.51
7	2.41	5.17	3.44	2.21	1.48	-2.50
8	1.47	1.80	1.23	1.07	0.33	-1.11

is also calculated and presented in Table 3.3 (see e.g., Maydeu-Olivares, Fairchild, & Hall, 2017, for details on the equation). Using the standard of $|skewness| > 1.5$ (Forero & Maydeu-Olivares, 2009), all the items were fairly symmetric except for Item 7.

3.1.2 Manipulated Factors

Three factors are manipulated in simulation study I, which are listed in Table 3.4. The number of repeated items was set to four (first four items in Table 3.3) or eight (all eight items in Table 3.3). With eight item coupled with four time points, the dimensionality of the model would be 12 for second-order model and 14 for first-order model. It is recognized that longer tests could be used in empirical studies. However, more repeated items will further increase the dimensionality of the model, which makes estimation under FIML impractical. As a result, the length of the test was not further considered. Three levels of sample size (200, 500, 2000) were selected following the work of Forero and Maydeu-Olivares (2009), corresponding to very small, sufficient and large sample sizes, respectively. The last manipulated factor

Table 3.4

Manipulated Factors in Simulation I

Manipulated factors	**Levels**
Number of repeated items	(4, 8)
Sample size	(200, 500, 2000)
estimation method	(MCEM, MHRM, WLSMV)

was the estimation method used for the model. As mentioned before, the MCEM and MHRM have been developed under the FIML framework, while WLSMV is a popular limited-information estimator that implements DWLS for point estimates. MCEM and WLSMV were implemented with Mplus^{TM} 7.4 (Muthén & Muthén, 1998-2012) and MHRM is implemented with flexMIRT^{TM} 3.4 (Cai, 2017). In total, there were 18 conditions in Simulation I.

3.1.3 Identification of Data Analysis Model in Simulation I

The generated data were fitted with the generating model in the data analysis. In order for the analysis model to be identified, the distribution of the latent intercept was set to its true generating value so that $\boldsymbol{\gamma}_0 \sim$Normal(0, 1). This way, the model implied total variance for the first-wave latent factor was $\sigma^2_{\boldsymbol{\theta}_1} = 1 + \sigma^2_{\boldsymbol{\epsilon}_1} = 1.5$ in Simulation I. It is recognized that a variance of 1.5 does not exactly follow the convention of setting latent variance to 1 in IRT framework. However, this should not be an issue of concern. The structural and item parameter estimates obtained under this parameterization can be easily converted to any other desired scales in practice. If some of the item parameters are known a priori in empirical analysis, the model can also be identified by fixing the parameters of these items. For the

details of the conversion method, readers are referred to the works of Kamata and Bauer (2008) and Wang et al. (2016).

3.1.4 Evaluation Criteria

A pilot test was conducted to examine the number of replications needed in the simulation. The running means of the point estimates are calculated to check when the running mean stabilized. It was found that the running means for all estimates produced by the three methods stabilized before the first 50 replications. An example of the running mean plot is shown in Figure 3.2. The results for other parameter estimates were similar to Figure 3.2. In order to yield more accurate estimates of confidence interval converage rates, 250 replications were conducted under each condition.

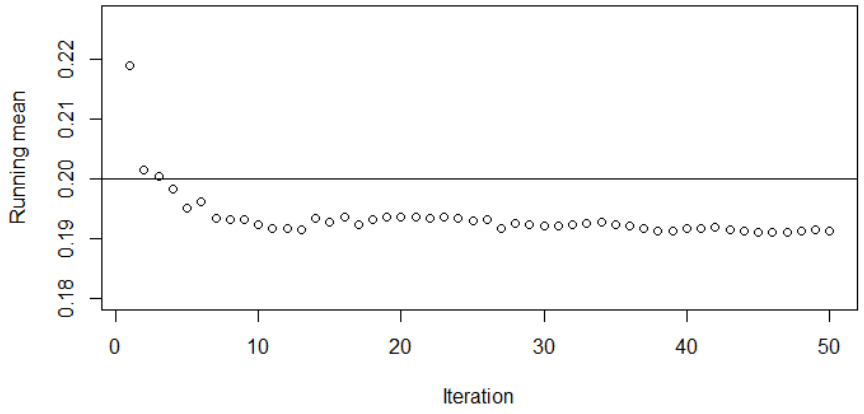

Figure 3.2. Running means of estimates of a model parameter in pilot test. The estimation method is DWLS. The horizontal line represented the true value. The running means stopped fluctuating greatly before the 50th replication.

Five outcome measures were used to evaluate the performance of the estima-

tion methods in Simulation I, including (1) estimation time (i.e., computer run time under same number of cores and processes), (2) convergence rate, (3) percent relative bias of structural and item parameters,(4) root mean squared errors (RMSE) of point estimates, (5) adequacy of the standard errors and (6) coverage of parameter confidence intervals. As for person parameters, this study only considered structural parameters (i.e., latent intercepts, slopes, and the variance-covariance structure of the latent intercepts and slopes). Ability estimates for individual persons were not examined in this simulation.

Convergence for FIML methods was decided by a tolerance of 1e-4 in the E step and a positive definite Fisher information matrix. Convergence for DWLS was decided by a tolerance of 1e-4 in the parameter estimates and a positive definite Fisher information matrix. The convergence rate was calculated as the percentage of converged replications among the original 250 replications. For outcome measures of model parameter estimates and their intervals, a total of 250 replications was used, including the original datasets that have converged for FIML-MCEM and the replacement datasets. Mean relative bias for each model parameter across replications was used to gauge the magnitude of the bias in parameter estimates, which was calculated by dividing the mean difference between the parameter estimate and the true parameter by the true parameter. The RMSE was calculated as the square root of the mean squared difference between the estimate and the true value of a parameter. RMSE is the bifurcation of bias and sampling error. The adequacy of the standard errors was examined by comparing the estimated average standard errors of a parameter over 250 replications to the Monte Carlo standard deviations

of the 250 parameter estimates. The adequacy of standard errors could also be examined by comparing the empirical variances of point estimates to mean squared standard errors. The pilot study indicated that the two methods produced similar results. Therefore only the results of the former method were reported. The coverage rates of the confidence intervals were obtained by taking the percentage of the confidence intervals that included the true parameters over 250 replications. The empirical coverage rates were then compared to the nominal level (95%) to assess the performance of each estimation method. A coverage rate lower than 95% would indicate that the confidence interval is too liberal in terms of type-I error. Whereas, a coverage rate higher than 95% would indicate that the confidence interval is too conservative in terms of type-I error.

3.2 Performance Assessment of Misspecified Models: Simulation II

The second simulation was to examine the performance of the reduced models when the full model was simplified in order to achieve computational efficiency. As mentioned before, either the time-specific disturbances or the common item effects could be omitted for the model to be estimated with three-dimensional integration. The setup for the simulation study is detailed below.

3.2.1 The Data Generation Model

The item response data were generated using the full model as described in Simulation I. The details are not repeated here. After the data are generated under

Table 3.5

Manipulated Factors in Sensitivity Analysis in Simulation II

Manipulated factors	**Levels**
Number of repeated items	(4, 8)
Sample size	(200, 500, 2000)
Generated disturbance variance	(0.25, 0.5, 1)
Generated common item effect variance	(0.25, 0.50, 0.75,1.00)

the full model, the two reduced model as shown in Figures 2.4 and 2.6 were used to analyze the data.

3.2.2 Manipulated Factors

In order to test the sensitivity of the parameter estimates to the misspecification of the model, different levels of disturbance variances and common item effect variances were generated. Time-specific disturbances were set to be smaller than intercept variance, following the work of Enders and Tofighi (2008). Specifically, the disturbance variances were set at 0.25, 0.5 and 1, so that the ratio of the intercept variance to the disturbance variance was 4:1, 2:1 and 1:1 respectively. The common item effect variances were varied between 0.25 and 1.00. It should be noted that the estimation method was not a manipulated factor in the simulation study since the reduced models could be estimated with BAEM in flexMIRT^{TM} with three-dimensional integrations efficiently.

The manipulated factors in Simulation II are summarized in Table 3.5 for the two sensitivity analyses. In total, both sensitivity analyses would have 72 conditions.

3.2.3 Identification of Data Analysis Model in Simulation II

The analysis model was identified in the same manner as in Simulation I. The details are not repeated here.

3.2.4 Evaluation Criteria

For Simulation II, two outcome measures are used to evaluate the performance of the two reduced models, namely the percent relative bias of structural and item parameters and the confidence interval coverage rates of the true structural and item parameters.

3.3 Comparison of the estimation methods with Sample Attrition: Simulation III

The third simulation was to evaluate the performance of the estimation methods with attritions under the MAR-X (i.e., missing at random with respect to covariates only) and the general MAR (i.e., missing at random with respect to both observed outcomes and covariates) missing mechanisms. In this simulation, only the missingness of the response data was simulated. That is to say, the missingness on the covariate was not considered in the current study. The details of the simulation design are explained below.

3.3.1 The Data Generation Model for MAR-X

First, complete item response data were generated with a LGM-IRT model with one time-invariant covariate. The covariate value of a person X_j was generated from a normal distribution with mean 0 and variance 1. Instead of generating the latent intercepts and slopes directly as in Simulation I, the disturbances of the latent intercepts and slopes $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}_1$ were generated from a bivariate normal distribution. The covariate and the disturbances were used to generate the latent slopes and intercepts with the following equations:

$$\gamma_{0j} = 0 + \beta_0 X_j + \delta_{0j}, \tag{3.6}$$

and

$$\gamma_{1j} = 0.2 + \beta_1 X_j + \delta_{1j}, \tag{3.7}$$

where β_0 and β_1 are the regression coefficients of the covariates for the intercept and slope respectively. The parameters for other latent variables are the same as in Simulation I. The configurations of covariate and the latent variables are shown in Table 3.6. With this setup, the model implied variances and covariance of the latent intercept and slope were the same as in Simulation I. After the latent slopes and intercepts wre generated, the same steps as Simulation I were used to generate the time specific latent scores, common item effects, and finally the item response data using the item parameters in Table 3.3. The path diagram for the LGM-IRT model with covariate is shown in Figure 3.3.

After the complete item response data were created, a MAR mechanism was used to create attrition of the examinees. For simplicity, only permanent attrition was generated in this simulation, meaning when an examinee dropped out of a wave, he/she never returned. The probability of examinee j dropping out at time point n ($n > 1$) was set to be proportional to the exponential of its negative covariate value:

$$P(M_{nj} = 1) \propto \exp(-X_j), \tag{3.8}$$

where $M_{nj} = 1$ indicates dropping out permanently at the n^{th} wave. Thus, the permanent attrition of examinees was generated in the order of wave 1 to wave 4 based on the covariate. With four waves, there were four missing data patterns in total. Using this MAR-X mechanism, people with higher covariate values were less likely to drop out at any given wave. Two attrition rates (i.e., 10% per wave and 20% per wave) were considered. The 20% per wave attrition rate was meant to represent a case with severe attrition problem based on the review of Lee (2003), where only about half of the original examinees were retained in the last wave. The empirical point-biserial correlations between the covariate and the missing indicators at each wave was approximately -0.40 in conditions with 20% attrition per wave and -0.28 in conditions with 10% attrition per wave, which represent moderately weak correlations between the covariate and the attrition indicator.

Table 3.6

Configuration of Covariate and Latent Variables in Simulation III.

Latent variable parameters	**Values**
Variance of covariate	$\sigma_X^2 = 1$
Regression coefficient of latent intercept on covariate	$\beta_0 = 0.8$
Regression coefficient of latent intercept on covariate	$\beta_0 = 0.2$
Distribution of latent intercept disturbances	$\boldsymbol{\delta}_0 \sim$ Normal(0, 0.36)
Distribution of latent slope disturbances	$\boldsymbol{\delta}_1 \sim$ Normal(0, 0.16)
Covariance of latent intercept and slope disturbances	$\sigma_{\boldsymbol{\delta}_0\boldsymbol{\delta}_1} = 0.04$
Distribution of time-specific disturbances	$\boldsymbol{\epsilon}_n \sim$ Normal(0, $\sqrt{0.5}$)
Distribution of same item effects	$\boldsymbol{\xi}_i \sim (0, \sqrt{0.5})$

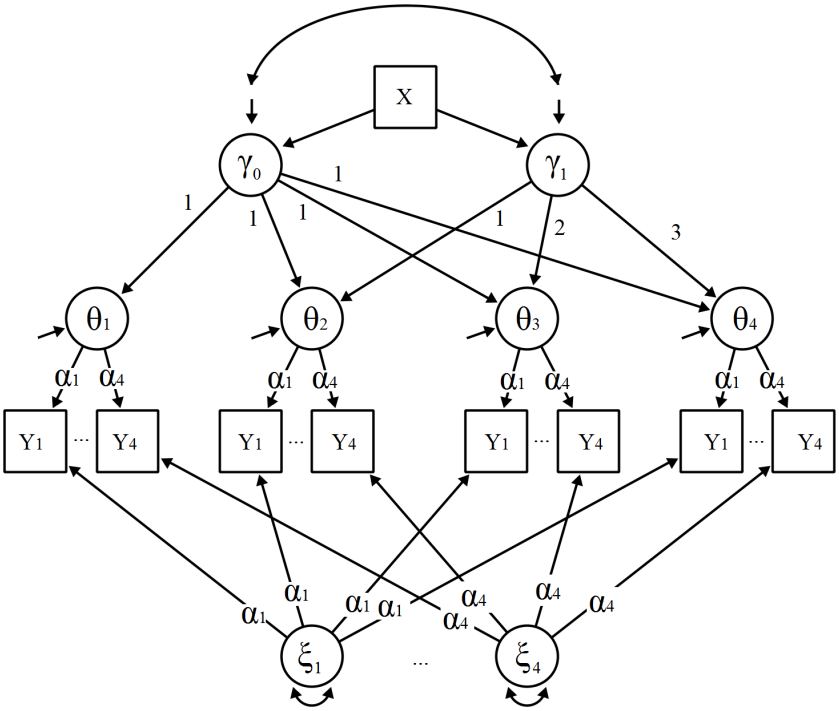

Figure 3.3. Second-order LGM-IRT model with one covariate for generating complete item response data in Simulation II. This example includes four time points and four repeated items. X is the time-invariance covariate for latent intercept γ_0 and latent slope γ_1. ϵ_n and ξ_i are the time-specific disturbance and common item effect respectively.

3.3.2 The Data Generation Model for MAR

The steps for generating the complete item response data were the same here as the situation under MAR-X (see Figure 3.3), while the method for generating missingness is different. Under the general MAR, the missing mechanism depends not only on the covariate, but also observed outcomes. After the complete item response data are created, the probability of examinee j dropping out at time point n ($n > 1$) was calculated as:

$$P(M_{nj} = 1) \propto \exp(-X_j - \sum_{i=1}^{I} Y_{i(n-1)j}). \tag{3.9}$$

The missing data were also generated from wave 1 to 4 sequentially. Besides the covariate, the probability of a person dropping out was also conditioned on his/her total score in previous wave. With this mechanism, people with higher covariate values and higher sum scores from previous test were less likely to drop out of the current assessment. The point-biserial correlations between the observed sum scores in previous test and the attrition indicator were approximately -0.75 and -0.6 under 20% and 10% attrition rats respectively, which represent a moderately strong correlation between the observe scores and the attrition indicator. The empirical point-biserial correlations between the missing indicators and the covariate were approximately -0.45 and -0.36 under conditions with 20% and 10% attrition per wave respectively. This indicates moderate correlation between the covariate and the attrition indicator. The correlations between the covariate and the missing

Table 3.7

Manipulated Factors in Simulation III

Manipulated factors	**Levels**
Number of repeated items	(4, 8)
Sample size	(500, 1000, 2000)
Permanent attrition rate per wave	(10%, 20%)
Missing mechanism	(MAR-X, MAR)
estimation method	(MCEM, MH-RM, DWLS)

indicator were stronger than the data generated under MAR-X. This was due to the fact that there was a positive correlation between the covariate and the sum scores, since the regression coefficients of the latent slope and intercept were set to be positive.

3.3.3 Manipulated Factors

There were two more manipulated factor in Simulation III than Simulation I, namely the attrition rate and the missing mechanism. The permanent attrition rate per wave was set to be 10% or 20%. For example, if the permanent attrition rate was set at 10% per wave, the second wave would have 90% of the initial sample size remaining. The third wave would have $90\% \times 90\%$ of the initial sample size remaining. The missing mechanisms were the MAR-X and the general MAR as discussed above. Under panel attrition, a sample of 200 hundred people would be too small. Therefore a sample size of 200 was not used as a condition here. A sample size of 1000 was added as a condition in Simulation III. All the manipulated factors are summarized in Table 3.7. In total, there are 72 conditions in this simulation.

3.3.4 Identification of Data Analysis Model in Simulation III

For Simulation III, the variance of latent intercept could not be fixed directly due to the inclusion of covariate in the model. Instead, the slope of the first item was set to the generating value for model identification. It is recognized that, in actual applied studies, the slopes of items might not be available. In that case, some reasonable arbitrary value (e.g., 1) can be used to fix the slope of one item. Post hoc transformation can be used to convert the variances, covariances and item parameters to any desired scale after the initial analysis as well.

3.3.5 Evaluation Criteria

The evaluation criteria used in Simulation III were the same with Simulation I, namely (1) estimation time, (2) convergence rate, (3) percent relative bias of structural and item parameters, (4) root mean square errors of point estimates, (5) Adequacy of the standard errors, and (6) coverage of parameter confidence intervals.

3.4 Empirical Data Analysis

The empirical illustration of this study used the "Language and Literacy" rating scale data from the "Multistate Study of Pre-Kindergarten 2001–2003" by the National Center for Early Development and Learning (Clifford, Bryant, Burchinal, & Barbarin, 2005). As part of this longitudinal study, the "Language and Literacy" skills of 1015 young children in the United States from pre-kindergarten to kindergarten were evaluated across four semesters from Fall 2001 to Spring 2003.

The teachers were asked to rate the students' skills every semester in approximately equal intervals, using the same 9 five-category Likert-scale items repeatedly.

The item response data are fitted with the LGM-IRT model with random item intercept effects using the three estimation methods of interests, using the graded response model (Samejima, 1969) as the measurement model. Additionally, the two reduced models will also be fitted using FIML-BAEM with three-dimensional integrations. In total, five analyses will be conducted. The results from the different models and estimation methods will be compared and contrasted.

Chapter 4: Results

The results of the three simulation studies as well as the empirical example are presented in this Chapter.

A pilot study with 50 replications was conducted to explore the configurations of the FIML-MH-RM algorithm before the simulation studies. It was found through the pilot trial that FIML-MH-RM algorithm is sensitive to starting values for the model being examined and can yield severely biased estimates; overestimated variance estimates; therefore, underestimated slopes due to the scale. To provide better starting values, DWLS point estimates were fed as the starting values for FIML-MH-RM estimation. To make fair comparison between FIML-MCEM and FIML-MH-RM, DWLS estimates were also used as starting values for FIML-MCEM estimation.

To verify that Schmid-Leiman tranformation was conducted properly, both the first- and second- order formulations of the LGM-IRT were tested using DWLS. The results confirmed that the two parameterization methods returned identical estimates.

The number of iterations in FIML-MH-RM estimation with post-convergence approximated standard errors was gradually increased to explore the number of

iterations needed to yield a positive definite information matrix. The results of the trial suggested that a large number of iterations (15,000) were needed to yield a convergence rate above 75% under complete data. The tuning constant was left at default (i.e., 1) and the proposal standard deviations were tuned (between 0.24 to 0.32 depending on the data condition) so that the acceptance rates of the Metropolis sampler were between 0.20 to 0.30.

In order to compare the estimation methods with the same number of replications, replacement datasets were added to conditions where the convergence rates were not 100%. Convergence rates were reported with original datasets, while the parameter estimates were compared using the original datasets that had converged along with the replacement datasets. For those situations where the convergence rates were below 50%, no results on point estimates or standard errors were reported.

The simulations were conducted on a computer with Intel(R) Xeon(R) dual CPU E5-1660 @ 3.30GHZ and 3.30GHZ with 36.0GB RAM. The computer was installed with Windows(R) 7 Professional 64-bit Operating System. One process (i.e., no parallel processing was utilized) was used for estimation of each replication.

4.1 Results of Simulation I

Simulation I examined the performance of the three estimation methods under complete data with a LGM-IRT. The results are summarize in this section.

4.1.1 Convergence and Estimation Time

The convergence rates of the estimation methods are presented in Table 4.1. DWLS estimations were able to convergence all the time under all the conditions considered in this simulation study. The convergence rates for FIML-MCEM were nearly 100% when the sample size was 500 or above, while the convergence rates dropped below 70% when the sample size was 200. FIML-MH-RM had almost perfect convergence rates across the conditions with recursively approximated standard errors, while the convergence rates ranged from 58% to 82% for FIML-MH-RM with post-convergence approximated standard errors using 15,000 iterations.

The estimation times for the estimation methods are reported in Table 4.2. As expected, DWLS was the fastest among all the methods. It took no more than 3 seconds for the model to be estimated. FIML-MCEM was the slowest estimation method. For a sample with 2000 people and eight items, FIML-MCEM took approximately one and a half hours for the model to converge. FIML-MH-RM with post-convergence approximated standard errors took 1/5 of the time required by FIML-MCEM. The recursive method was nearly 10 times faster than the post-convergence method.

Table 4.1

Convergence Rates (%) of Estimation Methods under Complete Data

		200 Examinees	500 Examinees	2,000 Examinees
4 Items	DWLS	100	100	100
	FIML-MCEM	61.6	88.0	98.4
	FIML-MH-RM (default)	95.6	99.2	100
	FIML_MHRM (Louis)	68.8	74.0	82.0
8 Items	DWLS	100	100	100
	FIML-MCEM	69.2	93.2	100.0
	FIML-MH-RM (default)	97.2	99.6	99.2
	FIML-MH-RM (Louis)	57.6	72.4	75.6

Table 4.2

Estimation Time (second) of Estimation Methods under Complete Data

		200 Examinees	500 Examinees	2,000 Examinees
4 Items	DWLS	1	1	1
	FIML-MCEM	363	997	2011
	FIML-MH-RM (default)	14	22	53
	FIML_MHRM (Louis)	153	244	575
8 Items	DWLS	1	2	3
	FIML-MCEM	1506	2538	5362
	FIML-MH-RM (default)	45	53	118
	FIML-MH-RM (Louis)	309	533	1040

4.1.2 Item Parameter Recovery

The relative bias of the item parameter estimates were calculated and then analyzed with an analysis of variance (ANOVA) with interactions. The independent variables were sample size, estimation method, test length, and all two-way and three-way interactions. It was found that there was a significant main effect for estimation method, $F(2, 528) = 6.50$, $\hat{\eta}^2 = 0.023$, $p < 0.01$, and a significant interaction between estimation method and sample size, $F(2, 528) = 4.05$, $\hat{\eta}^2 = 0.015$, $p = 0.02$.

To assess the relative bias in item parameter estimates, the true generating item parameters and their mean estimates across the replications are plotted in

Figure 4.1. The detailed relative bias for item parameter estimates is included in Table A.1 in Appendix A. In general, all the three estimation methods were able to return item parameter estimates that were less than 5% biased in either direction. The bias decreased when sample size increased. There were estimates, however, that were biased by more than 15% in either direction. The numbers of estimates that were more than 15% biased in either direction for DWLS, FIML-MCEM, and FIML-MH-RM across all six data conditions were 11, 15, and 3 respectively. Judging by the magnitudes of relative biases, FIML-MH-RM performed the best.

The RMSEs for item parameter estimates of these three estimation methods are plotted in Figure 4.2. In general, the RMSEs decreased as the sample size increased. RMSEs were larger for parameters with greater absolute values. The FIML-MH-RM algorithm yielded the smallest RMSEs, while DWLS produced the largest. The disadvantage of DWLS in terms of RMSE was more obvious under the smaller sample size condition. This indicates that DWLS estimates had the largest variability and disagreement among them especially when sample size was small.

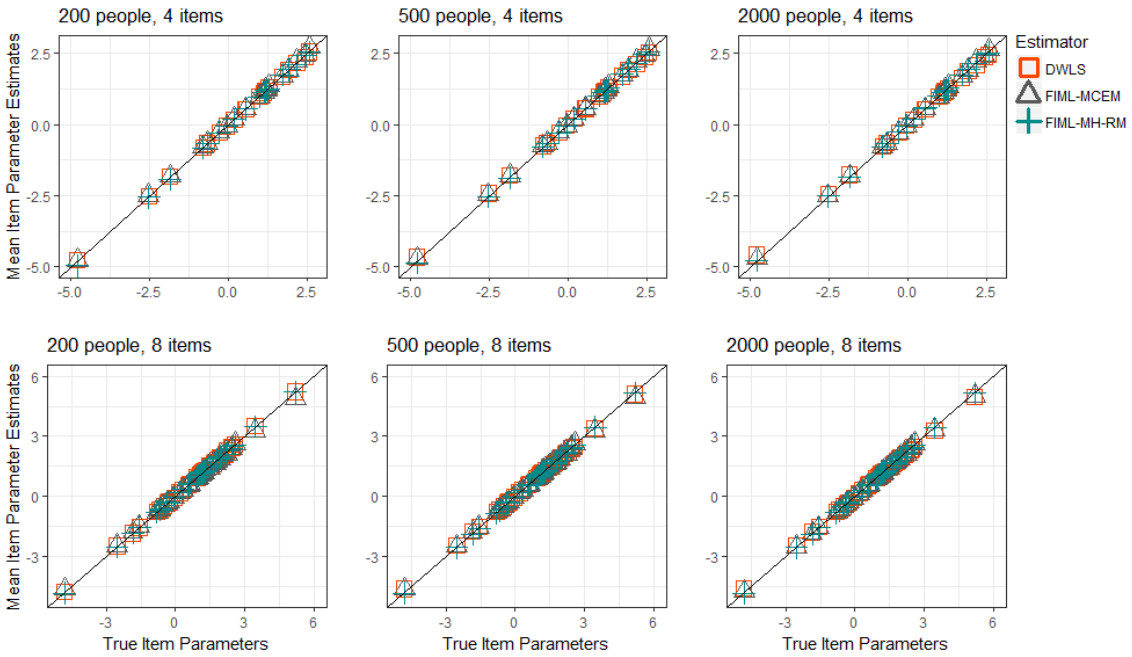

Figure 4.1. Recovery of item parameters across sample sizes and test lengths in Simulation I. All three estimation methods were able to yield almost unbiased item parameters as the points fell along the 45 degree line.

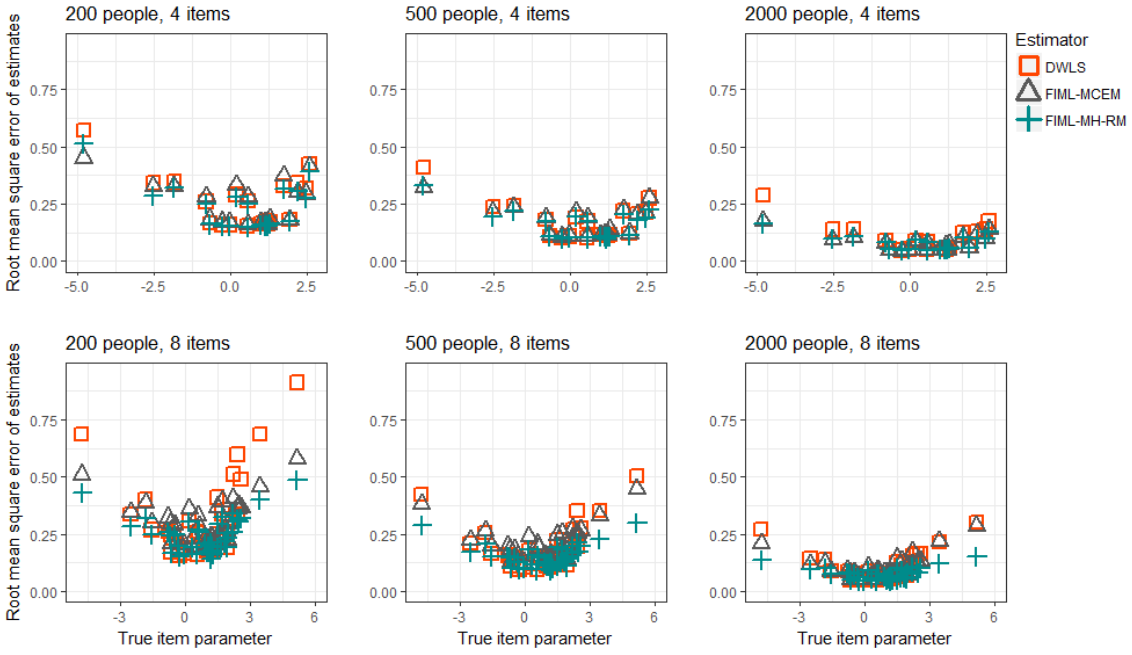

Figure 4.2. The RMSEs of the item parameter estimates for the estimation methods across sample sizes and test lengths in Simulation I. The FIML-MH-RM algorithm yielded the smallest RMSEs, while DWLS produced the the largest.

To examine the adequacy of the estimated standard errors for item parameters, the means of standard errors are plotted against the Monte Carlo standard deviations of the point estimates in Figure 4.3. Points above the 45 degree line indicate that the standard errors are overestimates, while points below the 45 degree line mean that the standard errors are underestimated. Among all the estimation methods, DWLS produced the most proper standard errors as the points mostly fell on the 45 degree line. DWLS produced larger standard errors than the FIML methods too. FIML-MCEM tended to overestimate the standard errors when sample size was small (200) and underestimate them with larger sample (2,000). This characteristic was more apparent with longer test as shown in the lower panels of Figure 4.2. The FIML-MH-RM underestimated the standard errors with both recursive approximation and the post-convergence approximation methods. The bias with the recursive approximation method was more severe than the post-convergence approximation method.

Taking both the point estimates and the standard errors into account, the coverage rates of the true item parameters in the 95% confidence intervals are plotted in Figure 4.4. With shorter tests, the coverage rates of the DWLS and FIML-MCEM were generally between 0.90 to 1 under 200 people and 500 people. When the sample size was increase to 2000, the coverage rates for item parameters with large absolute values dropped considerably for DWLS. With longer tests, the coverage rates dropped drastically for both DWLS and FIML-MCEM when sample size increased. For FIML-MH-RM, the coverage rates yielded by the two standard estimation methods were both underestimated and seemed unaffected by the sam-

ple size. This means that the estimated confidence intervals were all too liberal in terms of type-I error with FIML-MH-RM. The post-convergence approximation method was slightly better than the recursive approximation method due to the less underestimated standard errors it produced.

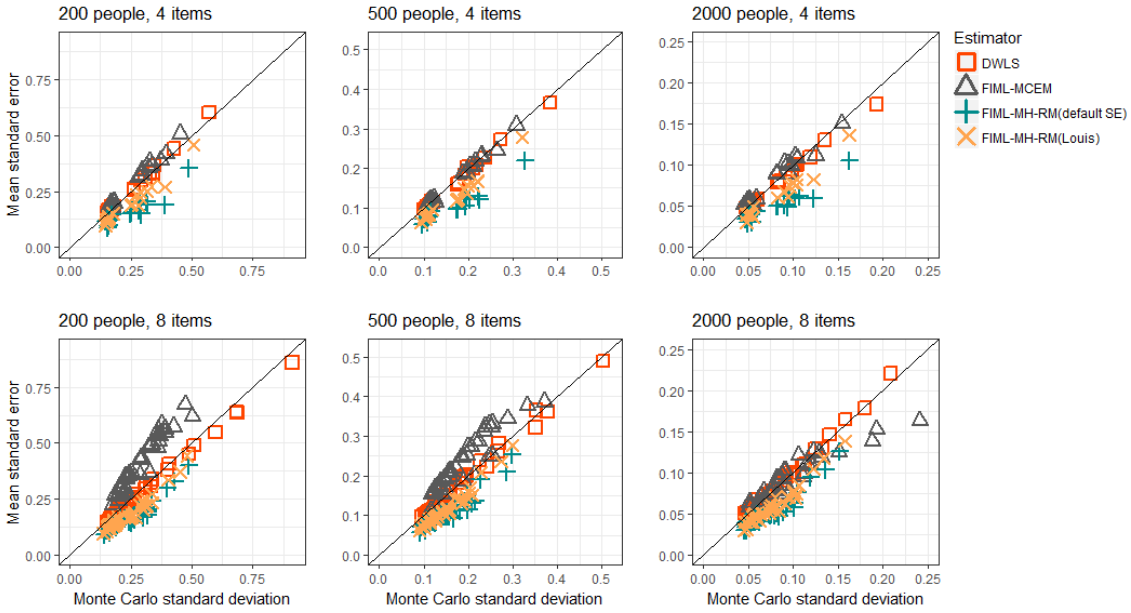

Figure 4.3. Monte Carlo standard deviations of item parameter estimates vs. mean item paramter standard error estimates across the replications across sample sizes and test lengths in Simulation I.

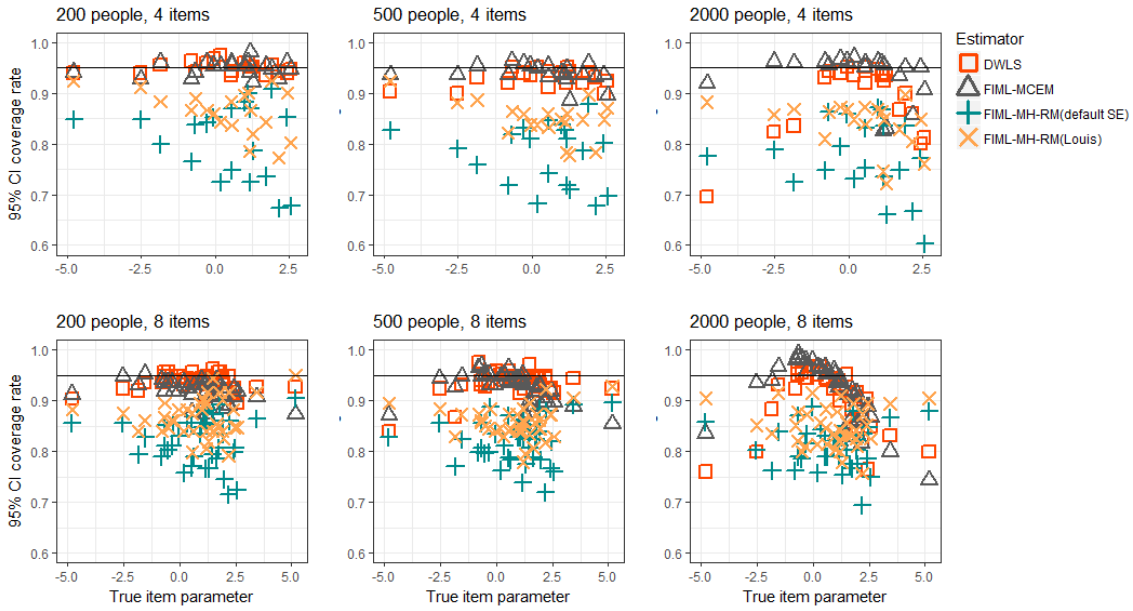

Figure 4.4. Coverage rates of the true item parameters in the 95% confidence intervals across sample sizes and test lengths in Simulation I. The horizontal line represents nominal level of 95%.

4.1.3 Structural Parameter Recovery

The relative bias of the latent slope mean, latent slope variance, and the covariance between latent slopes and intercepts is plotted in Figure 4.5. Overall, all the three estimation methods were able to recover mostly unbiased structural parameters. The relative bias was generally within 15% except when the sample size was 200 and the number of items was 4. DWLS produced the least biased structural parameter estimates (within ±2% in either direction). The bias for FIML-MCEM estimates was all negative, while the bias for FIML-MH-RM was all positive. In terms of the magnitude of relative bias, FIML-MH-RM slightly outperformed FIML-MCEM.

The RMSEs of the structural parameter estimates are plotted in Figure 4.6 to

compare the three estimation methods in terms of the variability of estimate across replication. When sample size was 200, FIML-MCEM outperformed DWLS and FIML-MH-RM. When the sample size was 500 or above, the FIML-MH-RM slightly outperformed the other two methods.

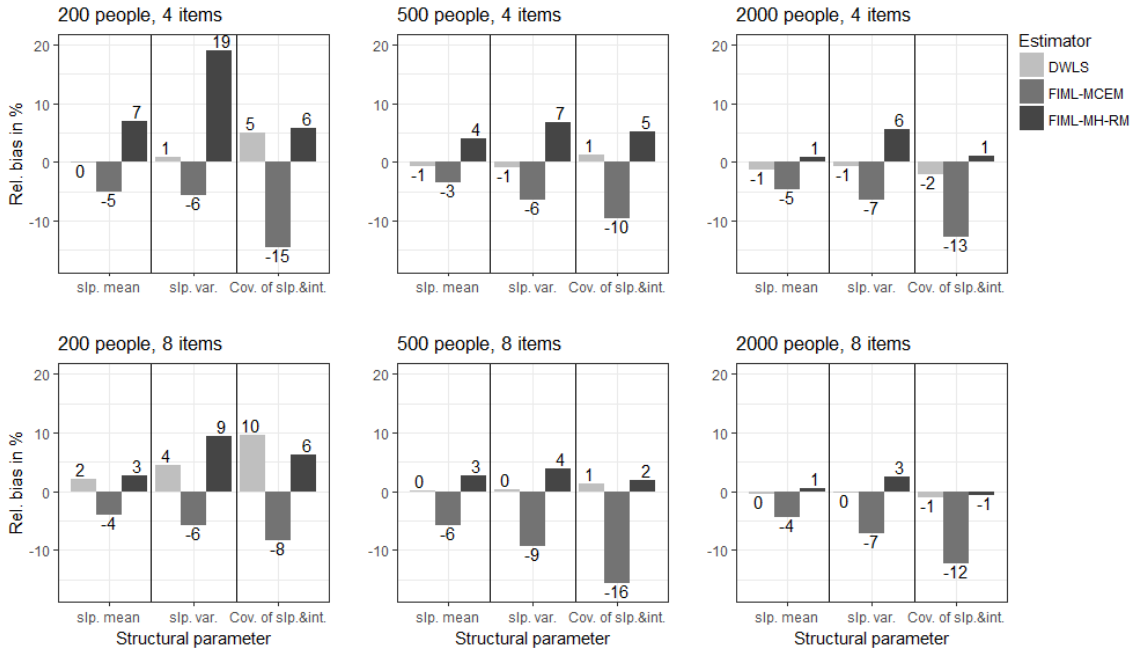

Figure 4.5. Relative bias of structural parameter estimates across sample size and test length in Simulation I. DWLS produced the least biased structural parameter estimates. FIML-MH-RM slightly outperformed FIML-MCEM

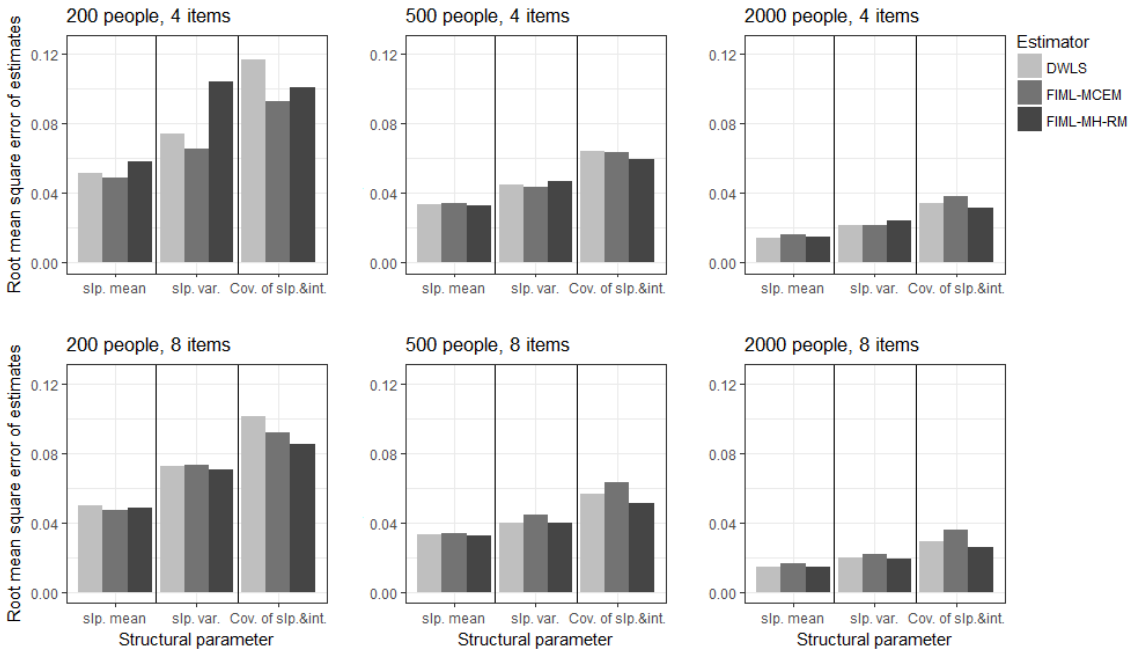

Figure 4.6. Root mean square errors of structural parameter estimates across sample size and test length in Simulation I.

In general, DWLS produced larger standard error estimates than the two FIML estimation methods. The ratios of mean estimated standard errors to Monte Carlo standard deviations of point estimates are plotted in Figure 4.7 to assess the adequacy of the estimated standard errors. A ratio larger than 1 means that the standard errors are overestimates, whereas a ratio smaller than 1 indicates underestimates standard errors. Most noticeably, the recursively approximated standard errors under FIML-MH-RM severely underestimated all the structural parameter standard errors. The other three methods were less biased. FIML-MCEM overestimated the standard errors when sample size was small, while it underestimated them with a large sample size. FIML-MH-RM with post-convergence approximated standard errors underestimated the standard error when the sample size was small (200) and the test length was four items.

Taking both the point estimates and standard errors into account, the coverage rates of the true structural parameters in 95% confidence intervals are plotted in Figure 4.8. The DWLS yielded the most accurate coverage rates as they were all very close to the nominal level. FIML-MCEM produced acceptable coverage rates when the number of items was four and the sample size was 500 or below, while it produced overly liberal confidence intervals in other situations. FIML-MH-RM with post-convergence approximated standard errors yielded acceptable coverage rates when sample size was 500 or below, while the coverage rates were too low with 2000 people. FIML-MH-RM with recursively approximated standard errors returned the poorest coverage rates, as they were all considerably under the nominal level.

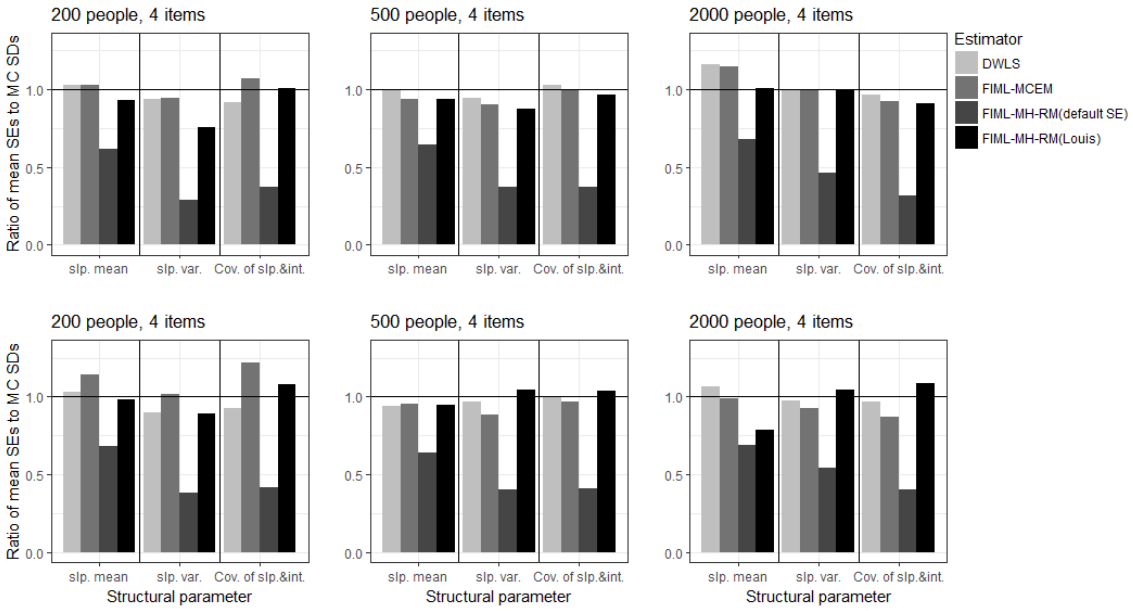

Figure 4.7. Ratio of mean structural parameter standard errors to Monte Carlo standard deviations of point estimates across sample sizes and test lengths in Simulation I. The horizontal line represents a ratio of 1. A ratio larger than 1 means that the standard errors are overestimates, whereas a ratio smaller than 1 indicates underestimates standard errors. FIML-MH-RM (default SE) is the recursively approximation method for standard errors. FIML-MH-RM (Louis) is the post-convergence approximation method for standard errors. FIML-MH-RM underestimated the item parameter standard errors with both methods.

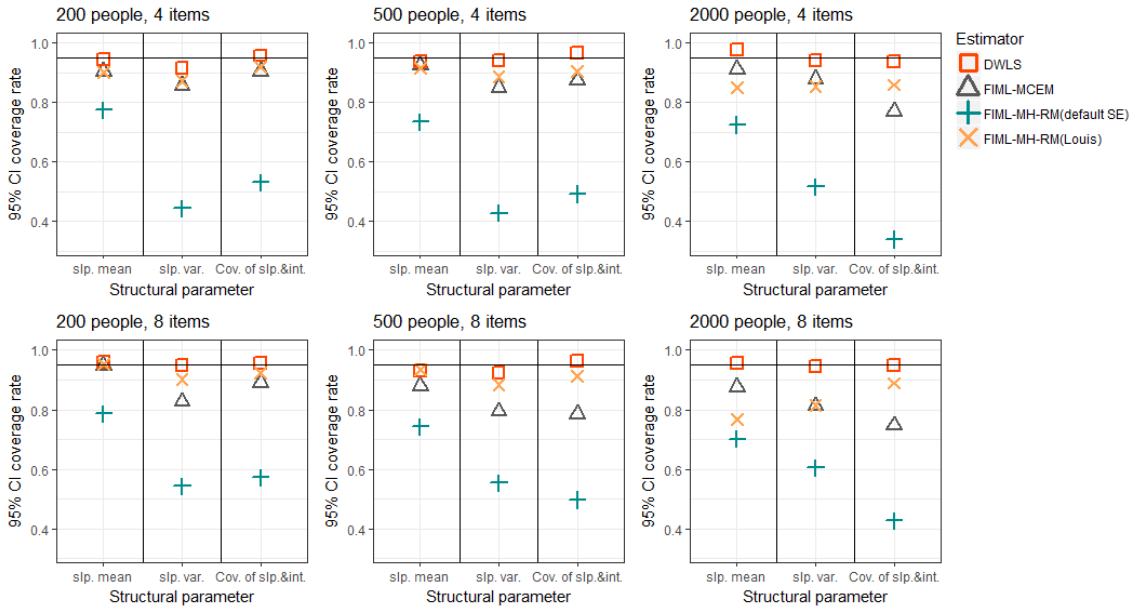

Figure 4.8. Coverage rates of the true structural parameters in the 95% confidence intervals across sample sizes and test lengths in Simulation I. The horizontal line was the nominal level of 95%. DWLS yielded the most appropriate confidence interval coverage.

4.2 Results of Simulation II

The results for Simulation II are summarized here, where either the common item effects or the time-specific disturbances were omitted to utilize the dimension-reduction technique for faster estimation under complete data.

4.2.1 Results of Omitting Common Item Effects

Convergence. The convergence rates of estimating the reduced model by omitting the common item effects are presented in Table 4.3. The reduced model ran into convergence issues when the time-specific disturbance variance was small (0.25). When the disturbance variance was 0.25 and the number of items was four,

Table 4.3

Convergence rates (%) of Estimating the Reduced Model by Omitting the Common Item Effects

			200 Examinees				**500 Examinees**				**2,000 Examinees**			
		C. var. =	0.25	0.50	0.75	1.00	0.25	0.50	0.75	1.00	0.25	0.50	0.75	1.00
4 Items	D. var.=0.25		0	0	0	0	69	0	0	0	96	0	0	0
	D. var.=0.50		95	81	62	0	99	96	86	65	100	100	100	97
	D. var.=1.00		97	98	92	94	100	100	100	100	100	100	100	100
8 Items	D. var.=0.25		100	100	87	0	75	50	0	0	91	78	0	0
	D. var.=0.50		100	100	100	100	97	95	95	94	99	100	99	96
	D. var.=1.00		100	100	100	100	99	98	99	97	100	100	100	99

Note. D. var=Time-specific disturbance variance; C. var=Common item effect variance.

the reduced model was not able to converge when the omitted common item effect variance was 0.50 or above. The convergence rates increased when the length of the test was increased to eight items. However, the reduced model was still not able to converge when the omitted common item effect variance was 0.75 or above with a small disturbance variance of 0.25 under 500 or 200 people.

Item parameter recovery. The relative biases of the item slope and intercept estimates were both negative when the common item effects were omitted. This means that the absolute values of these estimates were lower than those of the true values. The mean relative bias in item parameters for data conditions where the convergence rates were above 50% is plotted in Figure 4.9. The bias was larger in longer tests (from -12% to -32%) than in shorter tests (from -6% to -13%). The bias increased when the omitted common item effect variance was larger, while it was not sensitive to the magnitude of the disturbance variance.

Due to the magnitudes of bias, the coverage rates of the true item parameters (presented in Figure 4.10) were considerably under the nominal level especially under large sample sizes. In sum, the item parameter estimates produced by omitting the common item effects cannot be trusted.

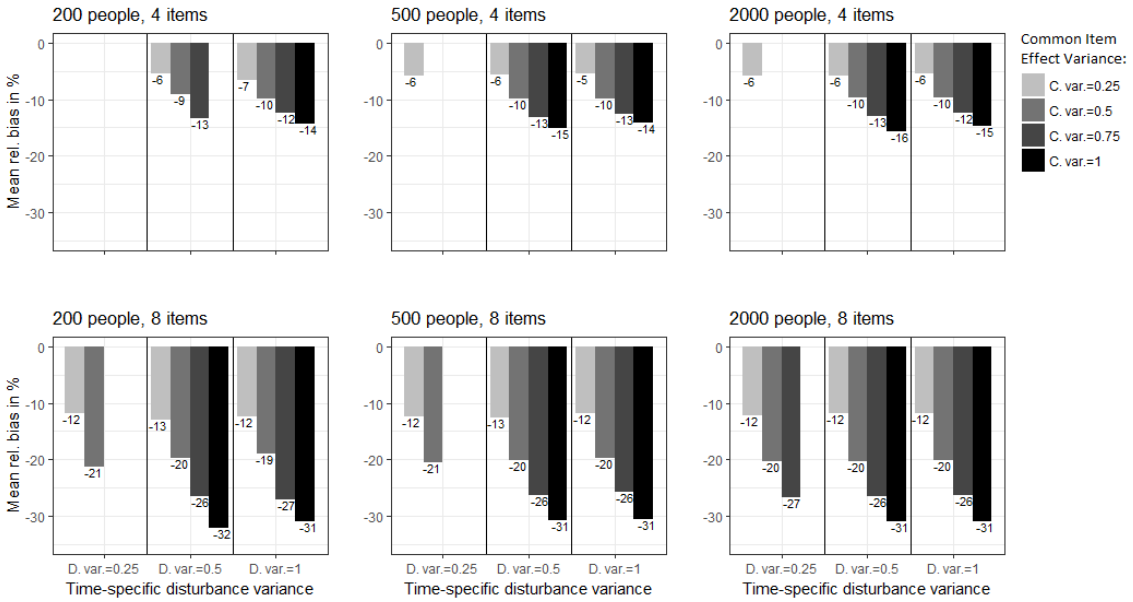

Figure 4.9. Relative bias in item parameter estimates across sample sizes and test lengths when common item effects were omitted in Simulation II. The estimates were all negatively biased. The bias increased along with the omitted common item effect variance. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

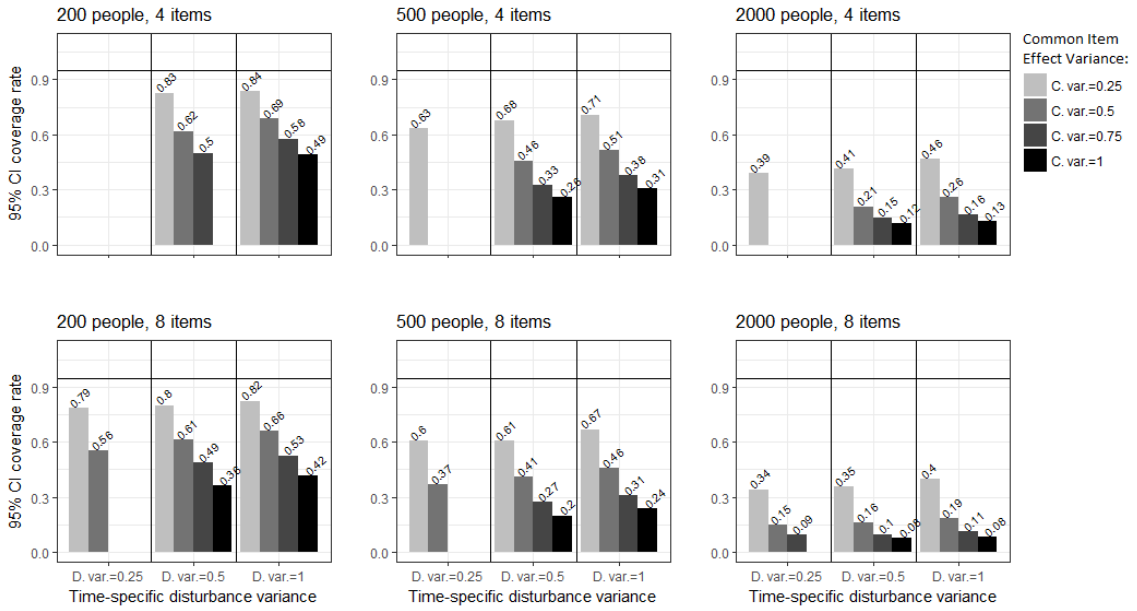

Figure 4.10. Coverage rates of true item parameters in 95% confidence intervals across sample sizes and test lengths when common item effects were omitted in Simulation II. The horizontal line represents the nominal level of 95%. All coverage rates were considerably below the nomoinal level. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

Structural parameter recovery. The relative biases in the estimates of latent slope mean, latent slope variance, and the covariance between latent slopes and intercepts are plotted in Figure 4.11 to Figure 4.13 for data conditions where convergence rates were above 50%. The estimated relative bias for the three structural parameter estimates was generally within 10% when sample size was 500 or above with longer tests. The coverage rates of the true structural parameters in the 95% confidence intervals are plotted in Figure 4.14 to Figure 4.16. The coverage rates were most appropriate when the sample size was 500 or below and the test length was eight items. The coverage suffered when the omitted common item effect variance increased and when the sample size was large (2,000).

In sum, when the common item effects were left out to achieve estimation efficiency, The item parameter estimates were severely downward biased. The structural parameter estimates were mostly unbiased when the test was longer and the sample size was sufficient (500 or above). However, when the sample size increased, the coverage of the structure parameters in the confidence interval was too liberal.

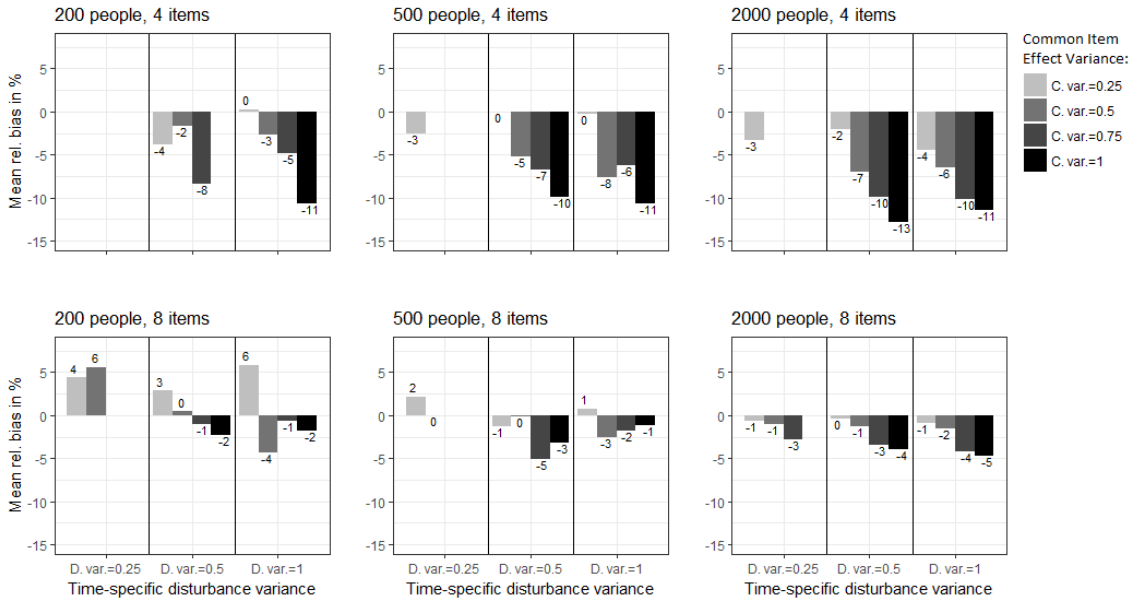

Figure 4.11. Relative bias in latent slope mean estimates across sample sizes and test lengths when common item effects were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

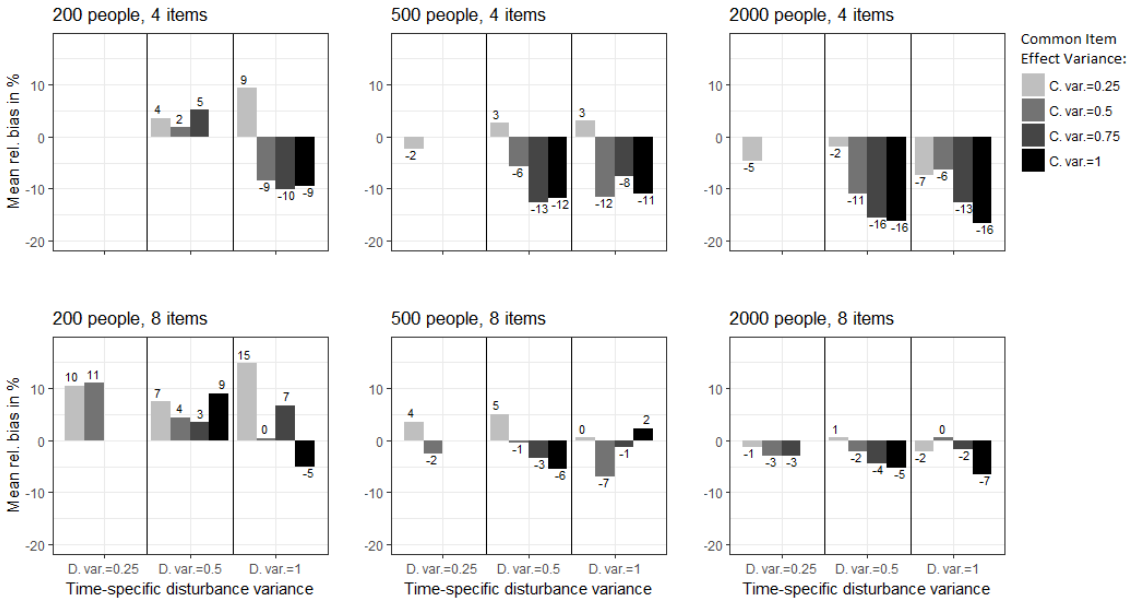

Figure 4.12. Mean relative bias in latent slope variance estimates across sample sizes and test lengths when common item effects were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

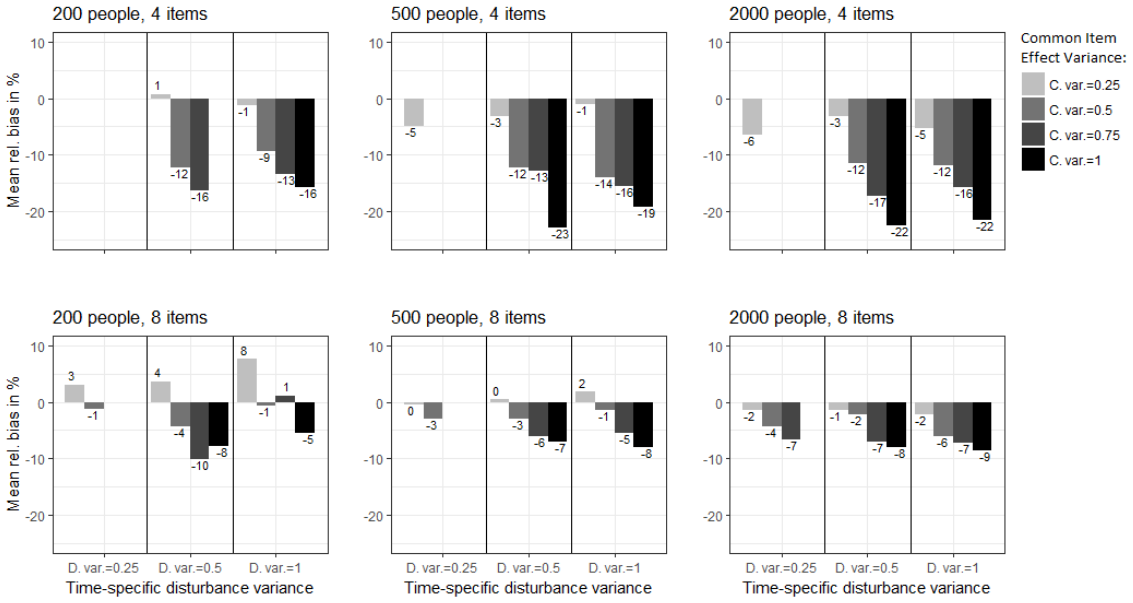

Figure 4.13. Relative bias in estimates of covariance between latent slopes and intercepts across sample sizes and test lengths when common item effects were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

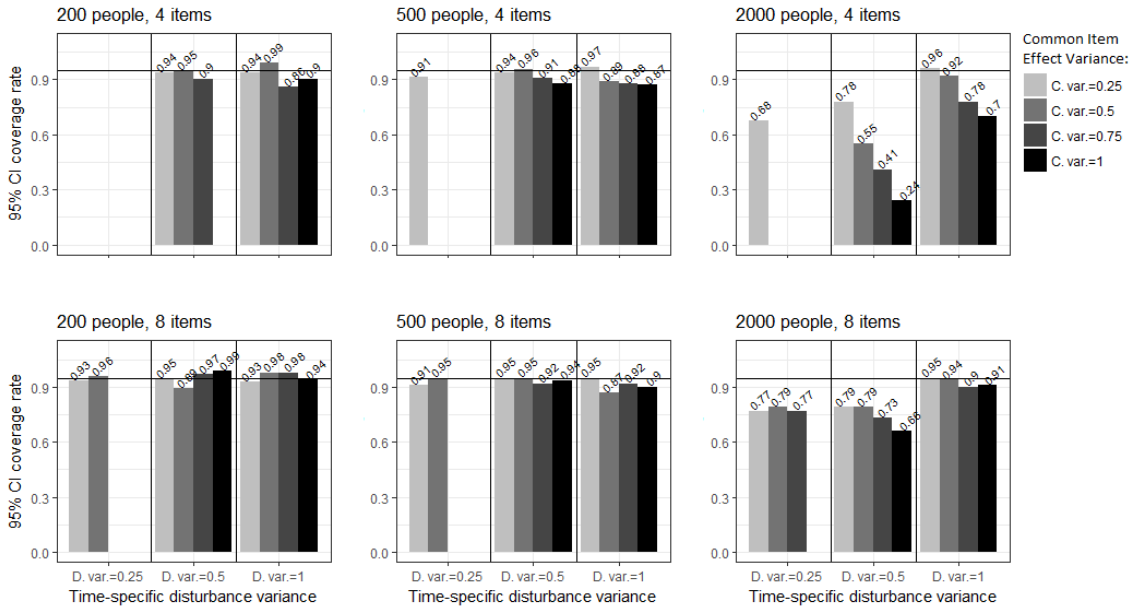

Figure 4.14. Coverage rates of latent slope means in 95% confidence intervals across sample sizes and test lengths when common item effects were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

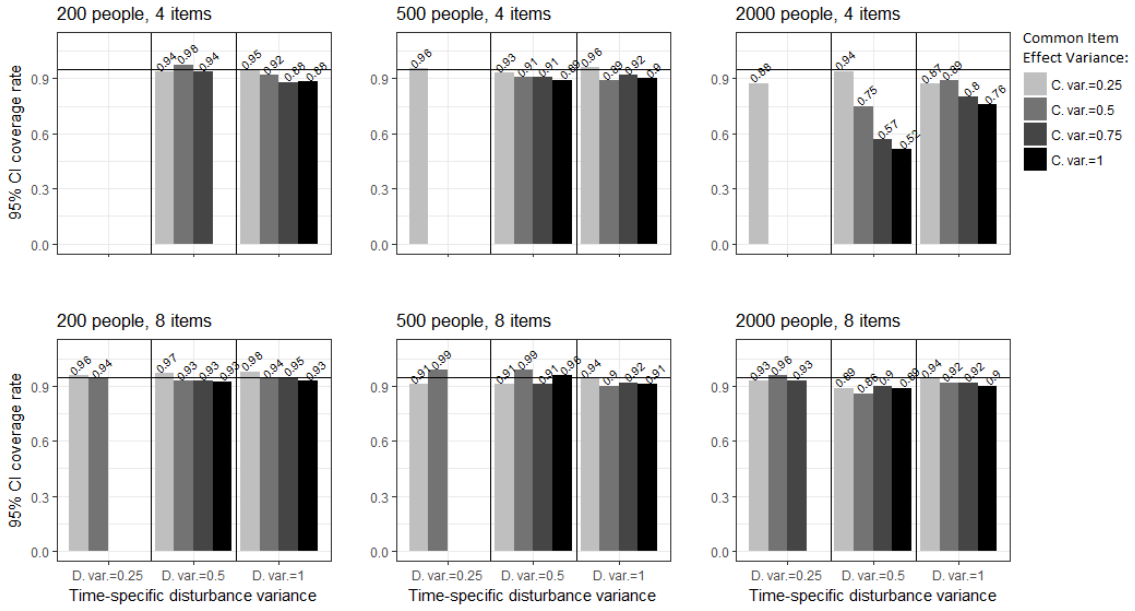

Figure 4.15. Coverage rates of latent slope variance in 95% confidence intervals across sample sizes and test lengths when common item effects were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

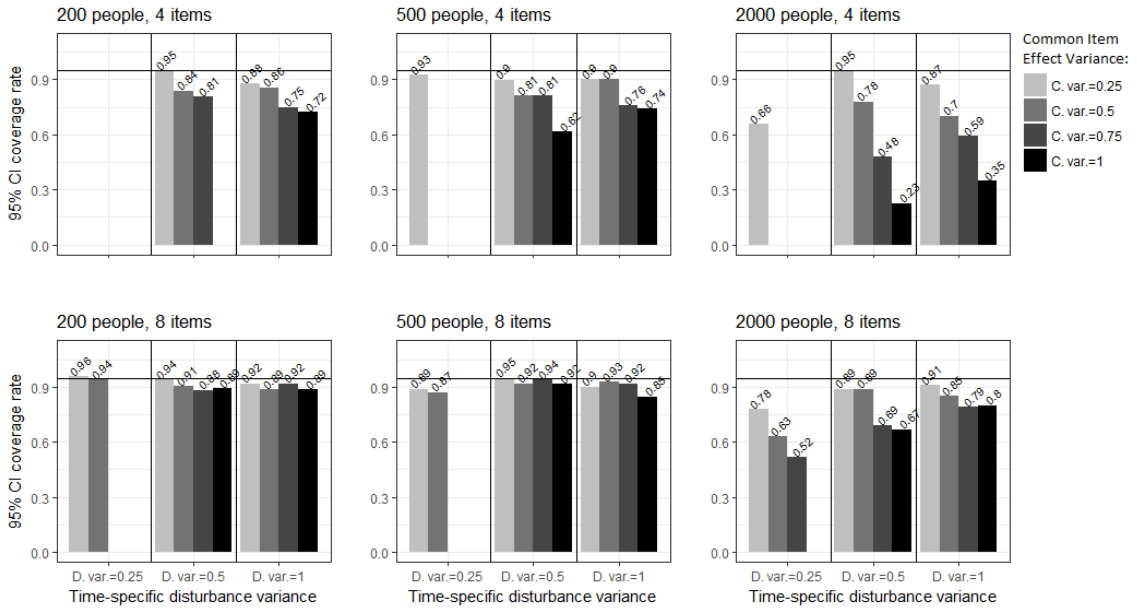

Figure 4.16. Coverage rates of the covariance between latent slopes and intercepts in 95% confidence intervals across sample sizes and test lengths when common item effects were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

4.2.2 Results of Omitting Time-Specific Disturbances

Convergence. The convergence rates of estimating the reduced model by omitting the time-specific disturbances are presented in Table 4.4. In general, the convergence rates decreased when the true common item effect variance was smaller. The model was not able to converge when the sample size was 200 with the longer tests.

It was worth mentioning that when the sample size was 500 and the number of items was 8, the model was not able to converge when the common item effect variance was 0.75 or above and the omitted disturbance variance was 0.50. However, under the same common item effect variances and larger disturbance variance (1.00),

the model had no convergence problem. The exact mechanisms of how sample size, test length, and the nuisance factor variance affect the convergence rates were not clear.

Table 4.4

Convergence rates (%) of Estimating the Reduced Model by Omitting the Time-Specific Disturbances

			200 examinees				**500 examinees**				**2,000 examineese**			
		C. var. =	0.25	0.50	0.75	1.00	0.25	0.50	0.75	1.00	0.25	0.50	0.75	1.00
4 Items	D. var.=0.25		0	88	100	100	0	99	100	100	77	100	100	100
	D. var.=0.50		0	73	98	100	0	97	100	100	0	100	100	100
	D. var.=1.00		0	0	83	98	0	0	97	100	0	88	100	100
8 Items	D. var.=0.25		0	0	0	0	0	100	100	100	88	100	100	100
	D. var.=0.50		0	0	0	0	0	98	0	0	0	0	100	100
	D. var.=1.00		0	0	0	0	0	0	99	100	0	95	100	100

Note. D. var=Time-specific disturbance variance; C. var=Common item effect variance.

Recovery of Item Parameters. When the time-specific disturbances were omitted, the item slope and intercept estimates were both negatively biased as well. The mean relative bias for item parameters when the time-specific disturbances were omitted is plotted in Figure 4.17 for data conditions where the convergence rates were over 50%. The bias increased when the omitted disturbances became bigger. For tests with four items, The bias was mostly within 10% across the sample sizes. When test length was eight items, the bias was within 15% when the omitted disturbance variances were 0.5 or lower. The bias was not sensitive to magnitudes of the common item effect variances, which were retained in the model.

The average coverage rates of the true item parameters in the 95% confidence intervals were plotted in Figure 4.18. The confidence intervals were all too liberal as the coverage rates were all below the nominal level. The coverage became worse when the sample size increased.

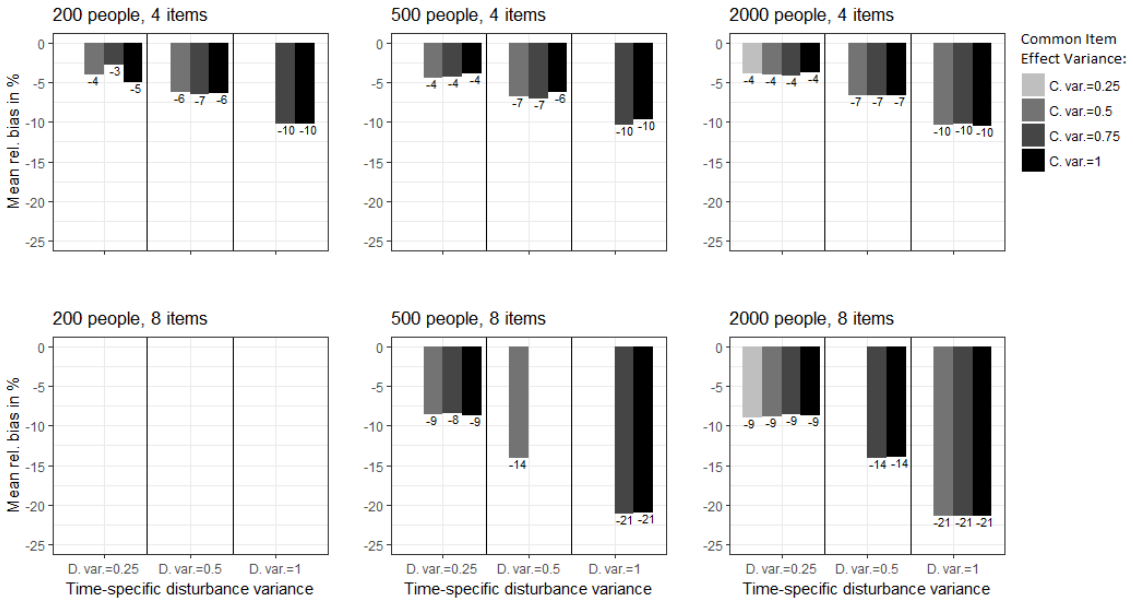

Figure 4.17. Average relative bias in item parameter estimates across sample sizes and test lengths when time-specific disturbances were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

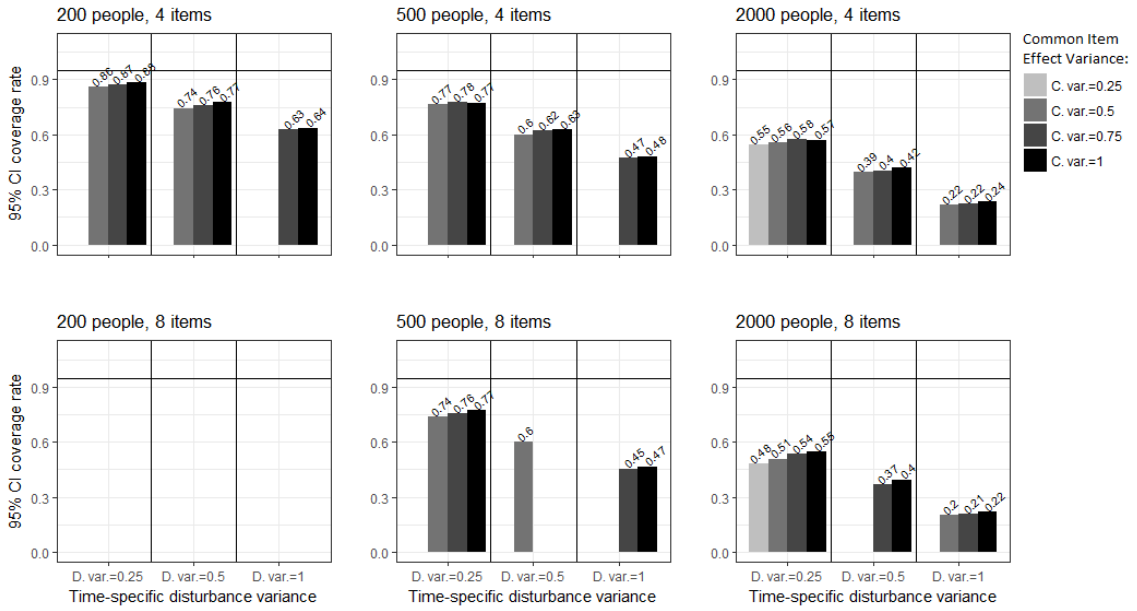

Figure 4.18. Coverage rates of true item parameters in 95% confidence intervals across sample sizes and test lengths when time-specific disturbances were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition. The horizontal line represents the nominal level of 0.95%. The coverage rates were all overly low.

Recovery of Structural Parameters. The relative bias in the estimates of latent slope mean, latent slope variance, and the covariance between latent slopes and intercepts is plotted in Figure 4.19 to Figure 4.21 for data conditions where convergence rates were above 50%. The estimates for the latent slope means were negatively biased (-1% to -21%). The estimates of the latent slope variances were severely downwardly biased (-35% to -125%). The covariance estimates were positively biased (8% to 31%). The bias was bigger under longer tests than under shorter tests and increased when larger disturbance variance was omitted.

The coverage rates of the true structural parameters in 95% confidence intervals are plotted in Figure 4.22 to Figure 4.24 for conditions where the convergence

rates were above 50%. The confidence intervals of the latent slope means and variance were all overly liberal except when the sample size was 500 or below with the shorter tests. The coverage rates for the covariance were all too low across all conditions.

In sum, when the disturbances were neglected in order to achieve faster estimation, the structural parameter estimates became severely biased, especially for the covariance components. The item parameter estimates were slightly biased when the test length was four items. However, the coverage of the true item parameters was all considerably lower than the nominal level. Generally speaking, omitting disturbances is not an ideal method for estimating the LGM-IRT.

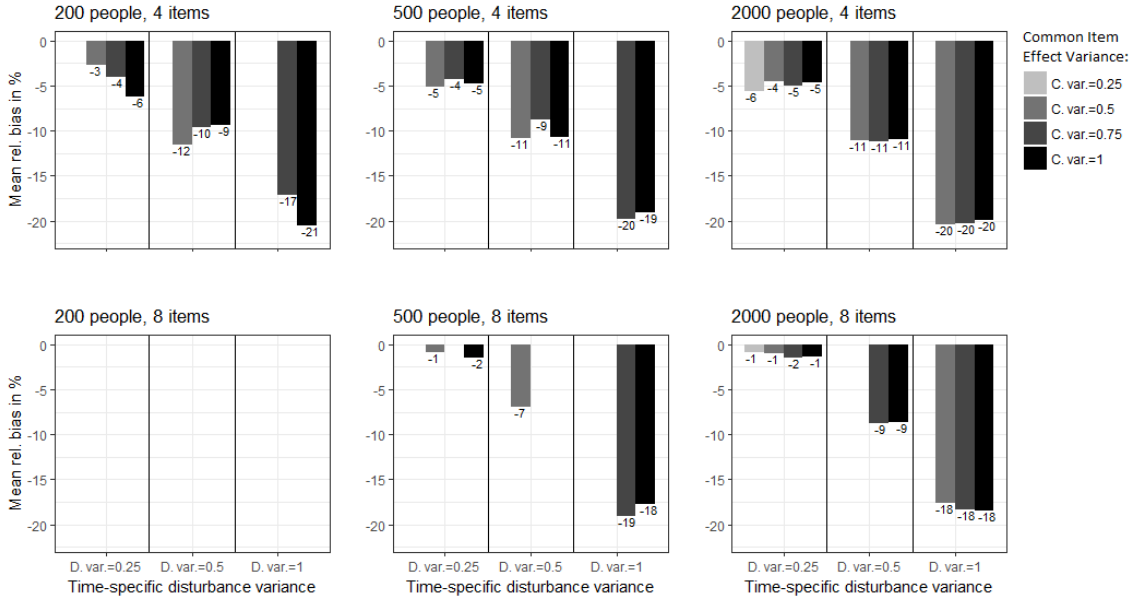

Figure 4.19. Relative bias in latent slope mean estimates across sample sizes and test lengths when time-specific disturbances were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

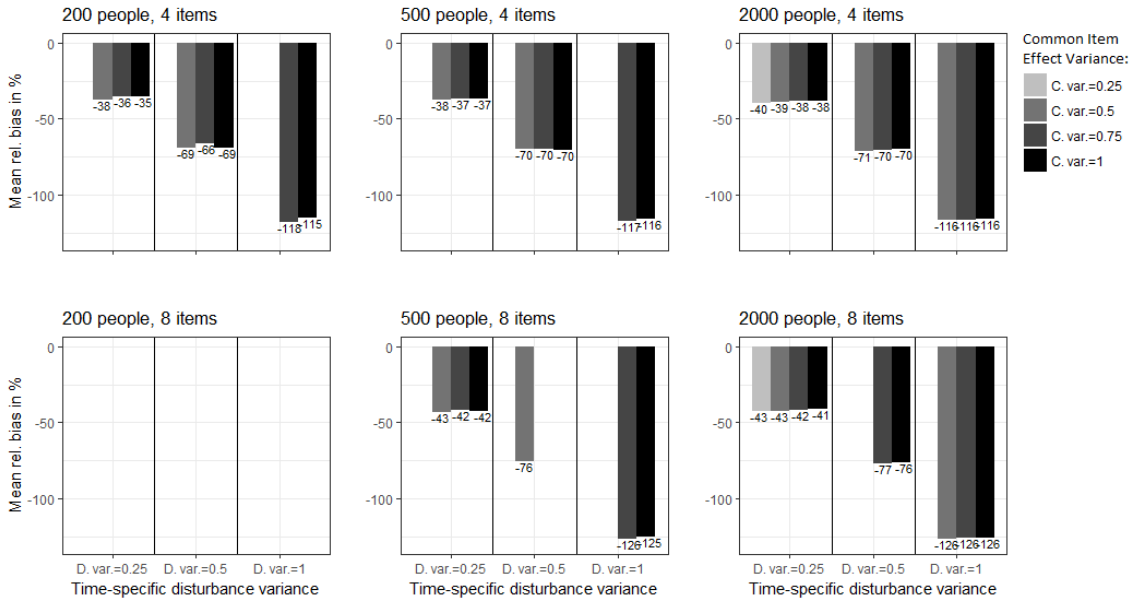

Figure 4.20. Relative bias in latent slope variance estimates across sample sizes and test lengths when time-specific disturbances were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

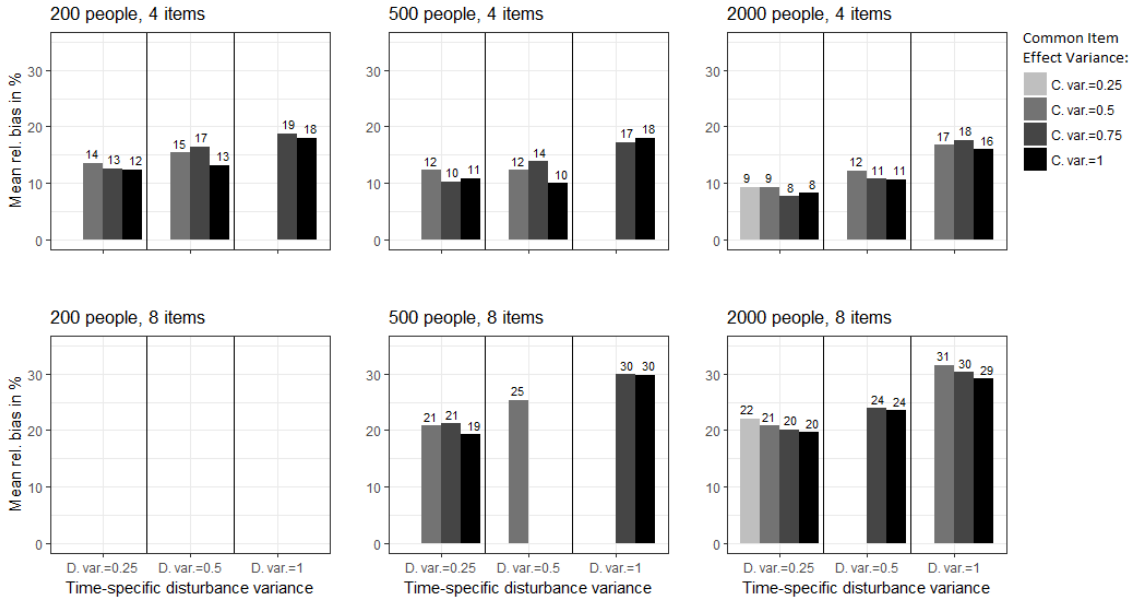

Figure 4.21. Relative bias in estimates of covariance between latent slopes and intercepts across sample sizes and test lengths when time-specific disturbances were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

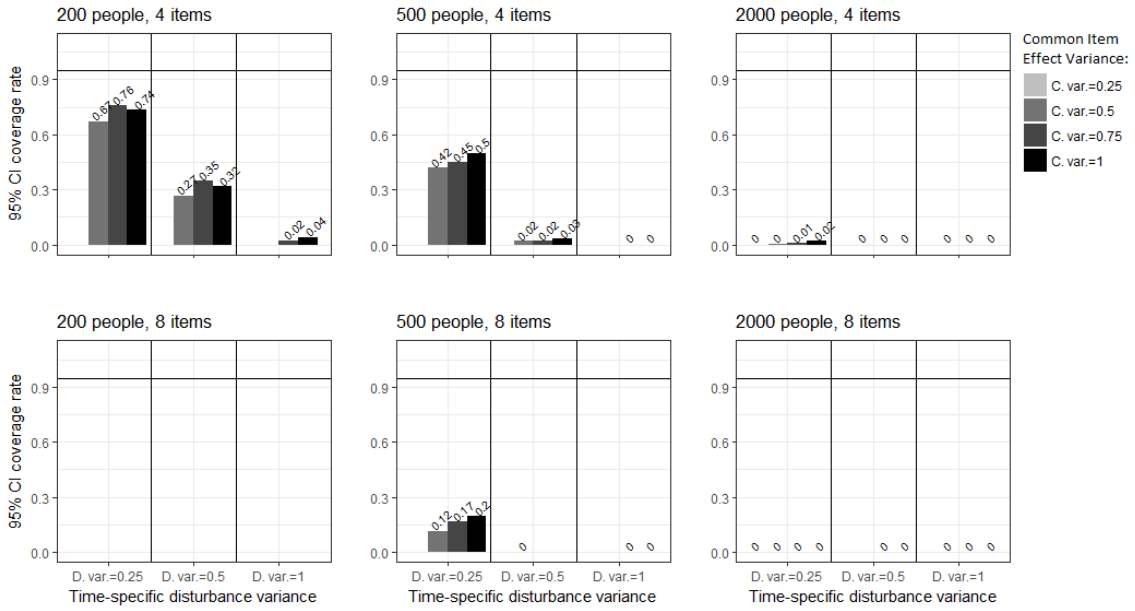

Figure 4.22. Coverage rates of latent slope means in 95% confidence intervals across sample sizes and test lengths when time-specific disturbances were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

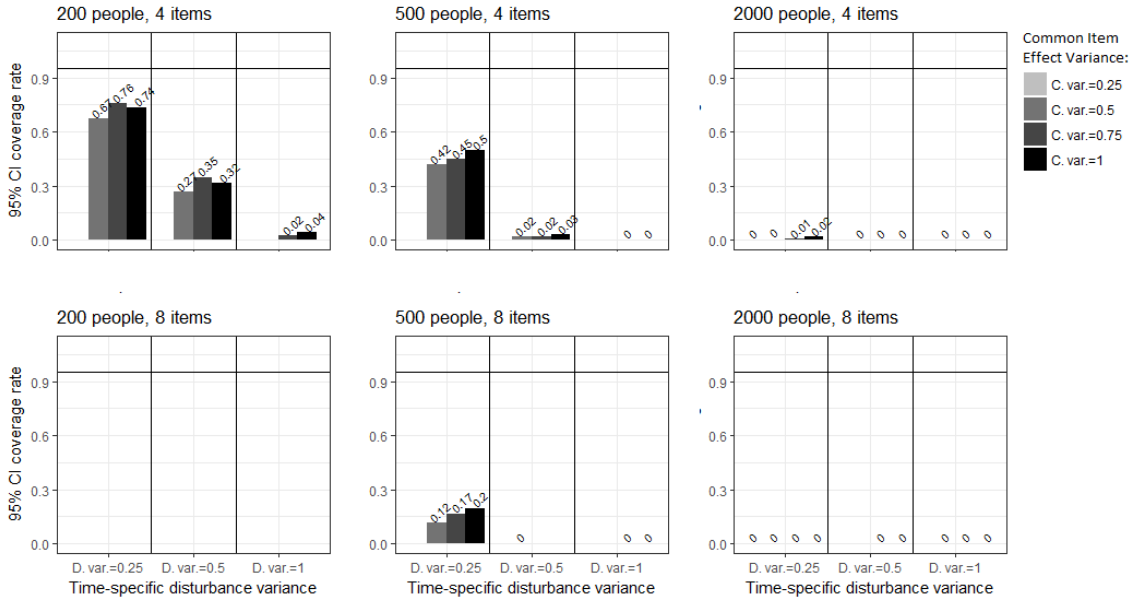

Figure 4.23. Coverage rates of latent slope variance in 95% confidence intervals across sample sizes and test lengths when time-specific disturbances were omitted were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

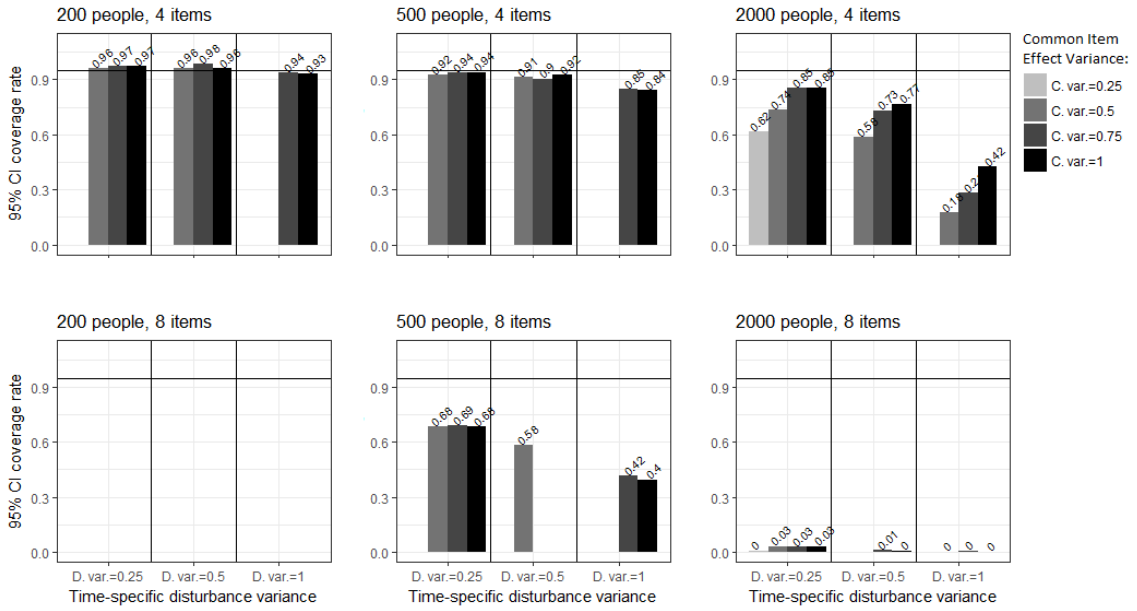

Figure 4.24. Coverage rates of the covariance between latent slopes and intercepts in 95% confidence intervals across sample sizes and test lengths when time-specific disturbances were omitted in Simulation II. D. var is the generating disturbance variance. C. var is the generating common item effect variance. A missing bar means that the convergence rate was below 50% under this data condition.

4.3 Results of Simulation III

This section summarizes the results of Simulation III, where the performance of the three estimation methods under MAR-X attrition and the general MAR attrition was examined in a conditional model with one time-invariant covariate.

4.3.1 Convergence and Estimation Time

The convergence rates for the two attrition rates (10% and 20%) crossed with the two attrition mechanisms (MAR-X and general MAR) are presented in Table 4.5 to Table 4.8. The convergence rates of all three estimation methods were above

90% across all conditions. Among them, DWLS estimations were able to converge all the time across all conditions. Since the FIML-MH-RM with post-convergence approximated standard errors was not able to produce positive definite information matrices, its results are not reported here.

Table 4.5

Convergence Rates (%) of Estimation Methods under 10% per Wave MAR-X Attrition

		500 Examinees	**1,000 Examinees**	**2,000 Examinees**
4 Items	DWLS	100	100	100
	FIML-MCEM	96.4	97.6	100
	FIML-MH-RM (default)	98.4	100	99.6
8 Items	DWLS	100	100	100
	FIML-MCEM	93.6	99.2	100
	FIML-MH-RM (default)	97.6	94.8	95.2

Table 4.6

Convergence Rates (%) of Estimation Methods under 20% per Wave MAR-X Attrition

		500 Examinees	**1,000 Examinees**	**2,000 Examinees**
4 Items	DWLS	100	100	100
	FIML-MCEM	97.2	99.6	100
	FIML-MH-RM (default)	96.0	98.8	99.6
8 Items	DWLS	100	100	100
	FIML-MCEM	95.6	99.6	99.6
	FIML-MH-RM (default)	97.6	97.6	94

Table 4.7

Convergence Rates (%) of Estimation Methods under 10% per Wave General MAR Attrition

		500 Examinees	**1,000 Examinees**	**2,000 Examinees**
4 Items	DWLS	100	100	100
	FIML-MCEM	94.0	99.6	100
	FIML-MH-RM (default)	99.6	100	98.8
8 Items	DWLS	100	100	100
	FIML-MCEM	92.4	100	100
	FIML-MH-RM (default)	100	98.4	96.4

Table 4.8

Convergence Rates (%) of Estimation Methods under 20% per Wave General MAR Attrition

		500 Examinees	1,000 Examinees	2,000 Examinees
4 Items	DWLS	100	100	100
	FIML-MCEM	96.0	99.6	99.6
	FIML-MH-RM (default)	98.0	99.6	98.4
8 Items	DWLS	100	100	100
	FIML-MCEM	95.2	99.2	100
	FIML-MH-RM (default)	98.0	93.6	91.2

The estimation times for the three estimation methods are presented in Table 4.9 to Table 4.12. On average, FIML-MCEM required the longest time to estimate the models, while DWLS took the least. For a sample of 2000 and a test of eight items, FIML-MCEM took over seven hours for the model to converge. In general, FIML-MCEM took more time under general MAR than under MAR-X. The speed of FIML-MH-RM with recursive approximated standard errors did not seem to be affected by the missing mechanism.

Table 4.9

Estimation Time (second) of Estimation Methods under 10% per Wave MAR-X Attrition

		500 Examinees	1,000 Examinees	2,000 Examinees
4 Items	DWLS	1	1	2
	FIML-MCEM	3053	3792	5592
	FIML-MH-RM (default)	36	49	106
8 Items	DWLS	1	2	4
	FIML-MCEM	10382	17425	39641
	FIML-MH-RM (default)	71	157	239

Table 4.10

Estimation Time (second) of Estimation Methods under 20% per Wave MAR-X Attrition

		500 Examinees	1,000 Examinees	2,000 Examinees
4 Items	DWLS	1	1	2
	FIML-MCEM	3884	7261	12363
	FIML-MH-RM (default)	68	51	139
8 Items	DWLS	2	4	4
	FIML-MCEM	8924	10634	27721
	FIML-MH-RM (default)	134	113	229

Table 4.11

Estimation Time (second) of Estimation Methods under 10% per Wave General MAR Attrition

		500 Examinees	1,000 Examinees	2,000 Examinees
4 Items	DWLS	1	1	2
	FIML-MCEM	3584	7826	14381
	FIML-MH-RM (default)	31	63	94
8 Items	DWLS	2	3	4
	FIML-MCEM	13417	15264	27576
	FIML-MH-RM (default)	66	134	203

Table 4.12

Estimation Time (second) of Estimation Methods under 20% per Wave General MAR Attrition

		500 Examinees	1,000 Examinees	2,000 Examinees
4 Items	DWLS	1	1	2
	FIML-MCEM	4055	7720	8162
	FIML-MH-RM (default)	61	94	147
8 Items	DWLS	2	4	5
	FIML-MCEM	9320	20881	26113
	FIML-MH-RM (default)	107	148	231

4.3.2 Item Parameter Recovery

An ANOVA identical to the one in Simulation I was conducted on the relative bias of the item parameter estimates. It was found that under MAR-X, the main effects for estimation method were significant for 10% per wave attrition, $F(2, 530) = 14.74$, $\hat{\eta}^2 = 0.005$, $p < 0.01$, and 20% per wave attrition, $F(2, 530) = 17.40$, $\hat{\eta}^2 =$

0.060, $p < 0.01$. All other factors were not significant. Under MAR attrition, none of the manipulated factors was significant.

The true generating item parameters and their mean estimates across the replications for the two attrition rates (10% and 20%) crossed with the two attrition mechanisms (MAR-X and general MAR) are plotted in Figure 4.25 to Figure 4.28. The detailed relative bias in item parameters under MAR-X is presented in Table B.1 to Table B.4 in Appendix B.

Under MAR-X attrition, all three estimation methods were able to return nearly unbiased item parameter estimates. The numbers of estimates demonstrated relative bias greater than 15% in either direction across the six data conditions under 10% per wave MAR-X attrition were 14, 25, and 3 respectively for DWLS, FIML-MCEM, and FIML-MH-RM. The same measures under 20% per wave MAR-X attrition were 13, 30, and 3 respectively for DWLS, FIML-MCEM, and FIML-MH-RM. Judging by the magnitude of relative bias, FIML-MH-RM performed the best, while FIML-MCEM estimates were the most baised.

Under general MAR attrition, DWLS returned more biased estimates than the two FIML algorithms. This can be observed in Figure 4.27 and Figure 4.28, where some DWLS estimates noticeably deviated from the 45 degree line. The numbers of estimates that demonstrated more than 15% in either direction across the six data conditions with 10% per wave MAR attrition were 54, 24, and 3 respectively for DWLS, FIML-MCEM, and FIML-MH-RM. Under 20% per wave MAR attrition, the same measures were 63, 15, and 18 respectively for DWLS, FIML-MCEM, and FIML-MH-RM. Judging by the magnitudes of relative bias, DWLS estimates were

the most biased. FIML-MH-RM outperformed FIML-MCEM with 10% per wave MAR attrition. The performance of the two FIML algorithm were comparable under 20% per wave MAR attrition.

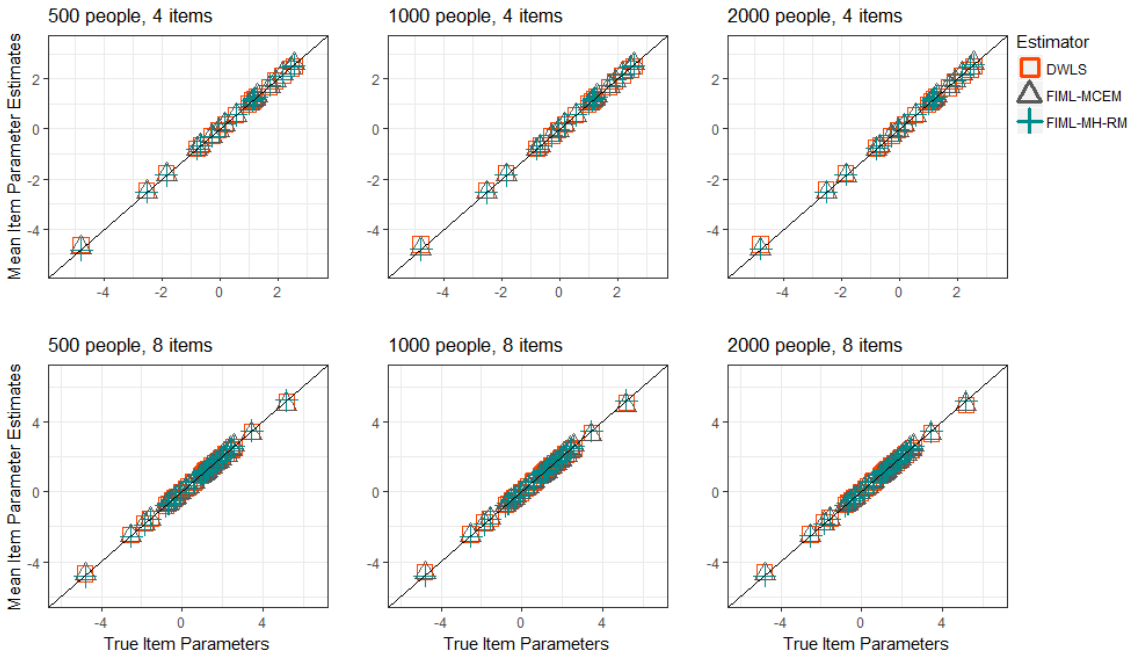

Figure 4.25. Recovery of item parameters across sample sizes and test lengths under 10% MAR-X attrition in Simulation III. All three estimation methods were able to yield almost unbiased item parameters as the points all fell on the 45 degree line.

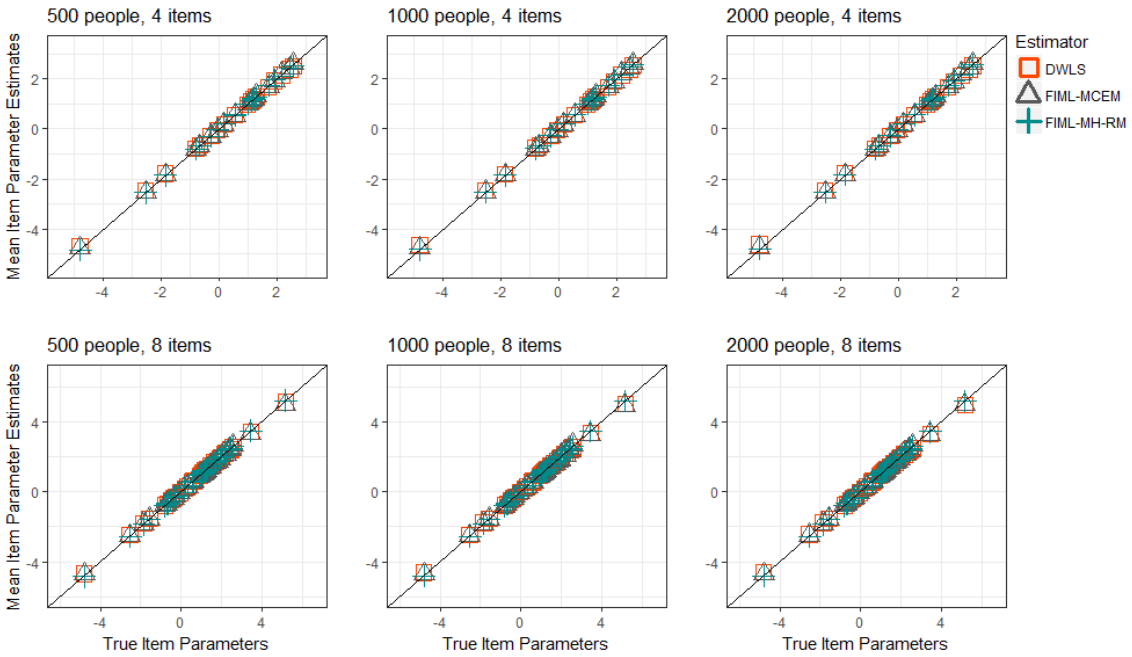

Figure 4.26. Recovery of item parameters across sample sizes and test lengths under 20% MAR-X attrition in Simulation III. All three estimation methods were able to yield almost unbiased item parameters as the points all fell on the 45 degree line.

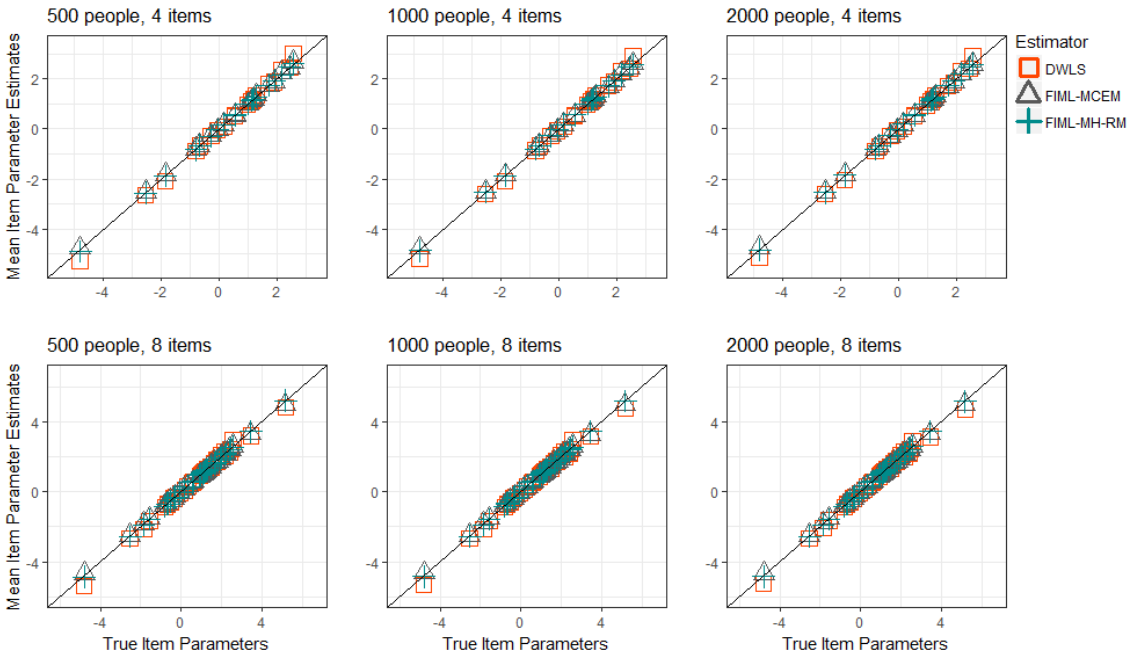

Figure 4.27. Recovery of item parameters across sample sizes and test lengths under 10% general MAR attrition in Simulation III. FIML methods were able to recover almost unbiased item parameters, while DWLS returned more biased item parameters.

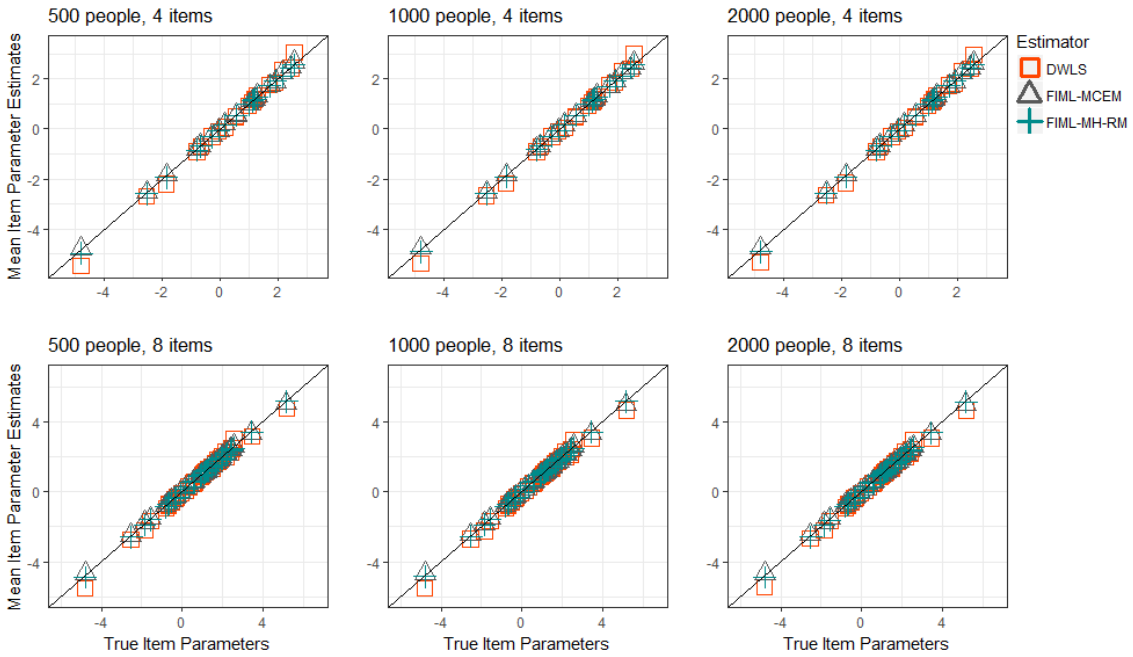

Figure 4.28. Recovery of item parameters across sample sizes and test lengths under 20% general MAR attrition in Simulation III. FIML estimation methods were able to recover almost unbiased item parameters, while DWLS returned more biased item parameters.

The RMSEs for item parameter estimates of these three estimation methods are plotted in Figure 4.30 to Figure 4.32 to examine the aggregated magnitudes of the errors for the item parameter estimates. In general, the FIML-MH-RM yielded the smallest RMSEs, while DWLS produced the the largest RMSEs especially under the smallest sample size condition (200).

Comparing the the RMSEs of item parameter estimates of the same attrition rate under the two attrition mechanisms (for example, Figure 4.30 vs. Figure 4.32), it can be observed that, the RMSEs produced by all three estimation methods were greater under general MAR than under MAR-X.

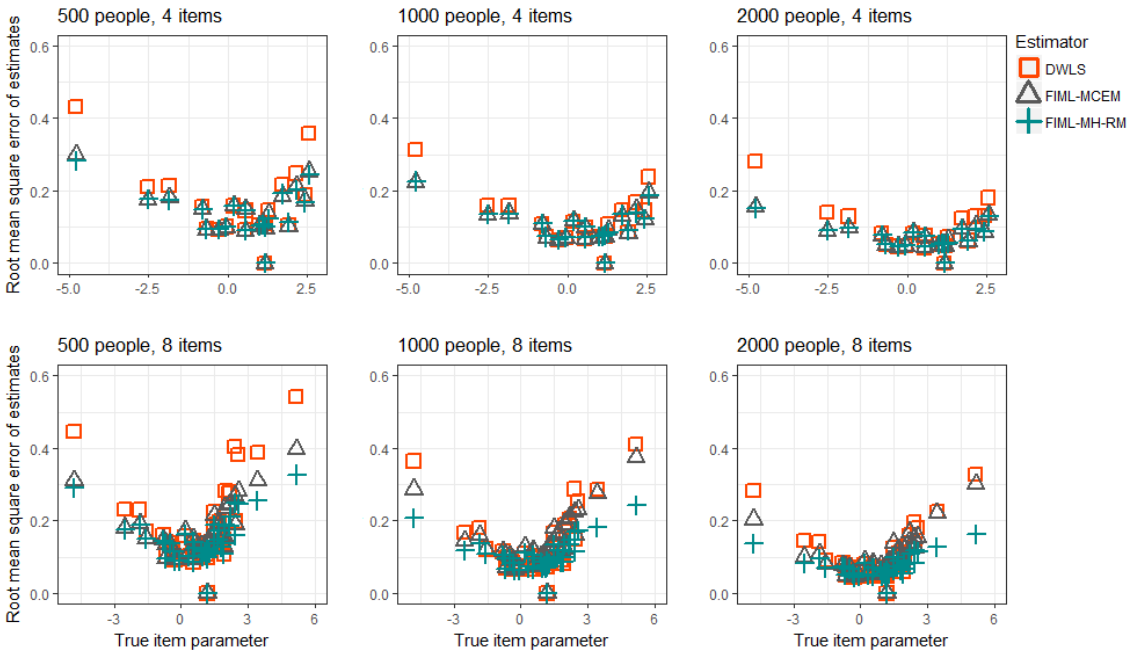

Figure 4.29. The RMSEs of the item parameter estimates for the estimation methods across sample sizes and test lengths under 10%/wave MAR-X attrition in Simulation III. The FIML-MH-RM algorithm yielded the smallest RMSEs, while DWLS produced the the largest RMSEs.

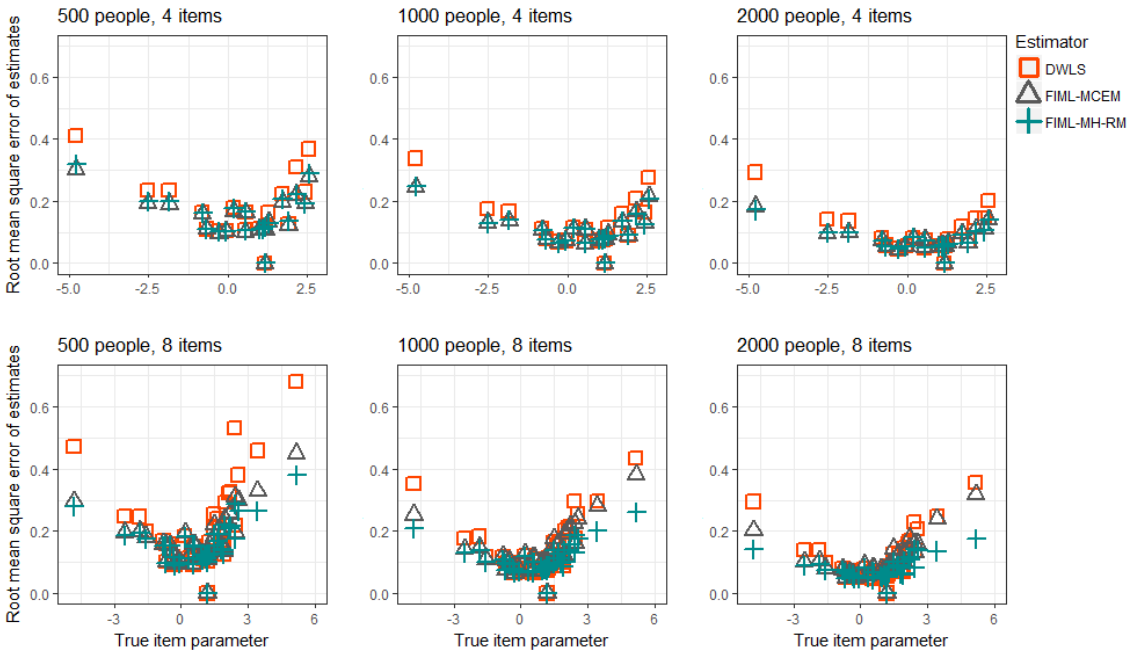

Figure 4.30. The RMSEs of the item parameter estimates for the estimation methods across sample sizes and test lengths under 20%/wave MAR-X attrition in Simulation III. The FIML-MH-RM algorithm yielded the smallest RMSEs, while DWLS produced the the largest RMSEs.

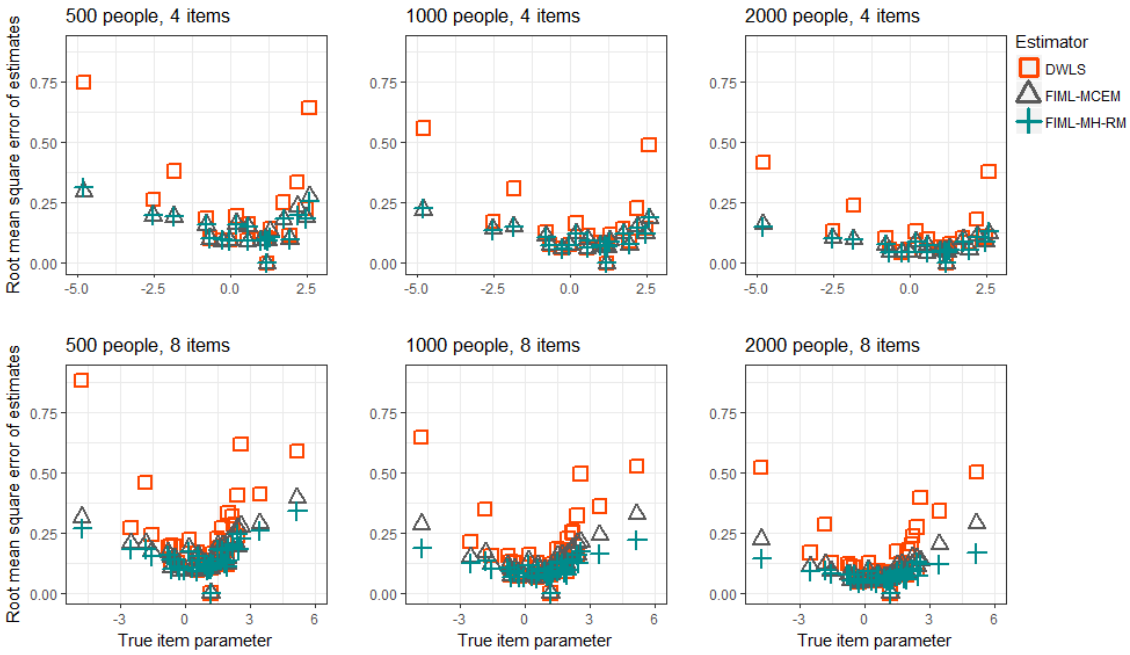

Figure 4.31. The RMSEs of the item parameter estimates for the estimation methods across sample sizes and test lengths under 10%/wave MAR-X attrition in Simulation III. The FIML-MH-RM algorithm yielded the smallest RMSEs, while DWLS produced the the largest RMSEs.

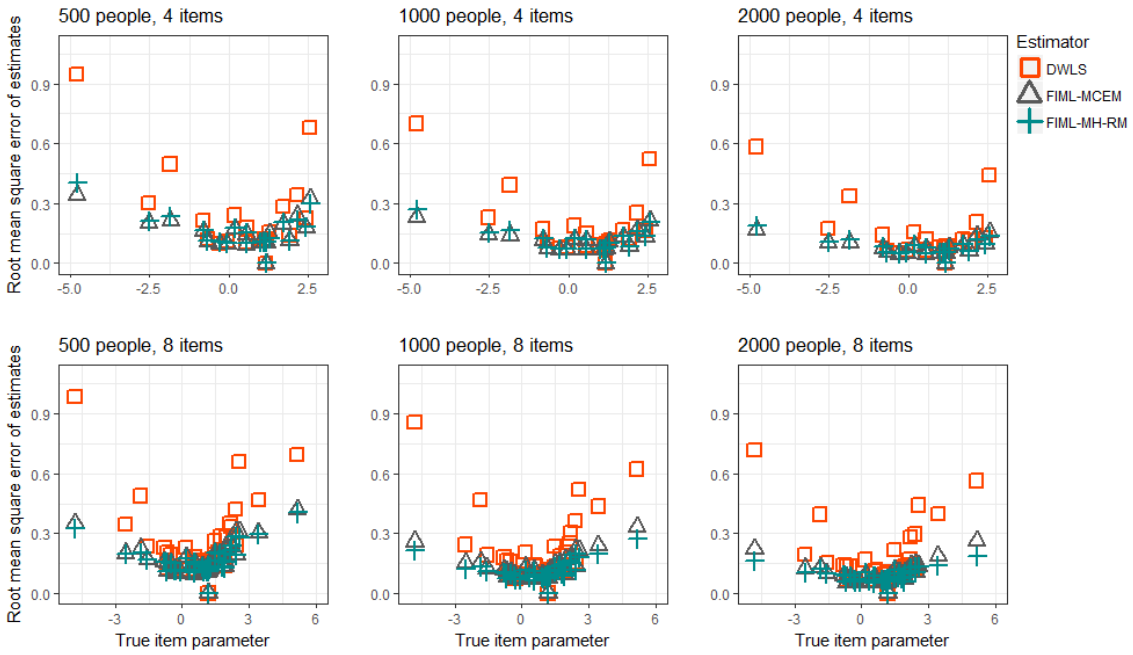

Figure 4.32. The RMSEs of the item parameter estimates for the estimation methods across sample sizes and test lengths in Simulation III. The FIML-MH-RM algorithm yielded the smallest RMSEs, while DWLS produced the the largest RMSEs.

The means of the estimated standard errors are plotted against the Monte Carlo standard deviations of item parameter estimates for the four conditions in Figure 4.33 to Figure 4.36 to assess the adequacy of the standard errors. Among all the estimation methods, DWLS produced the most proper standard errors as points mostly fell on the 45 degree line. FIML-MCEM underestimated the standard errors with larger sample (2000). The FIML-MH-RM underestimated the standard errors across all conditions with recursive approximated standard errors.

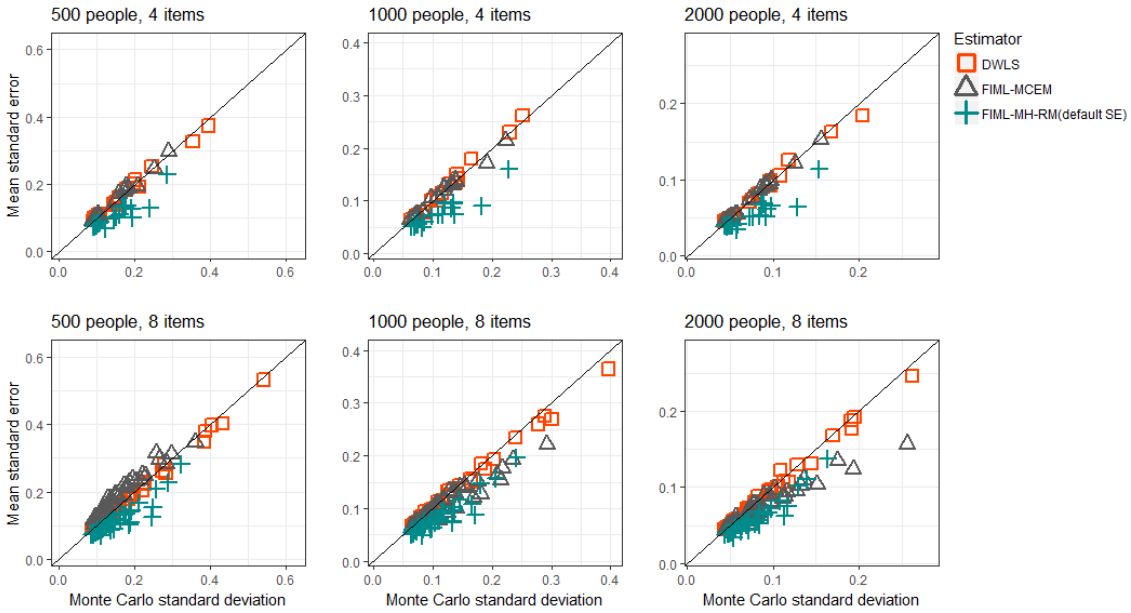

Figure 4.33. Monte Carlo standard deviations of item parameter estimates vs. mean item paramter standard error estimates across across sample sizes and test lengths under 10%/wave MAR-X attrition in Simulation III.

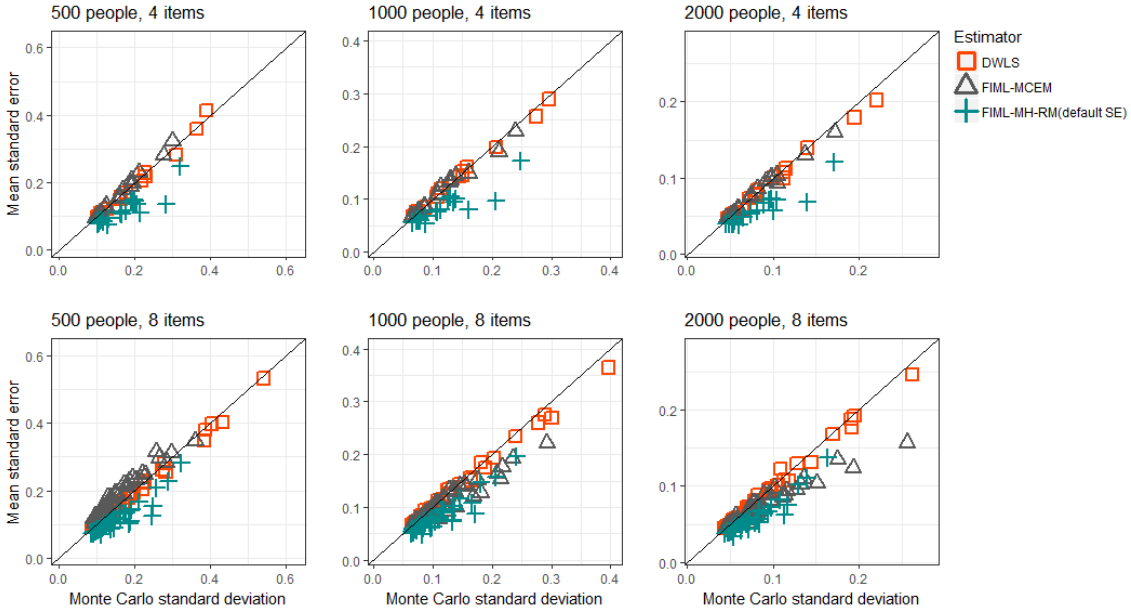

Figure 4.34. Monte Carlo standard deviations of item parameter estimates vs. mean item paramter standard error estimates across across sample sizes and test lengths under 20%/wave MAR-X attrition in Simulation III.

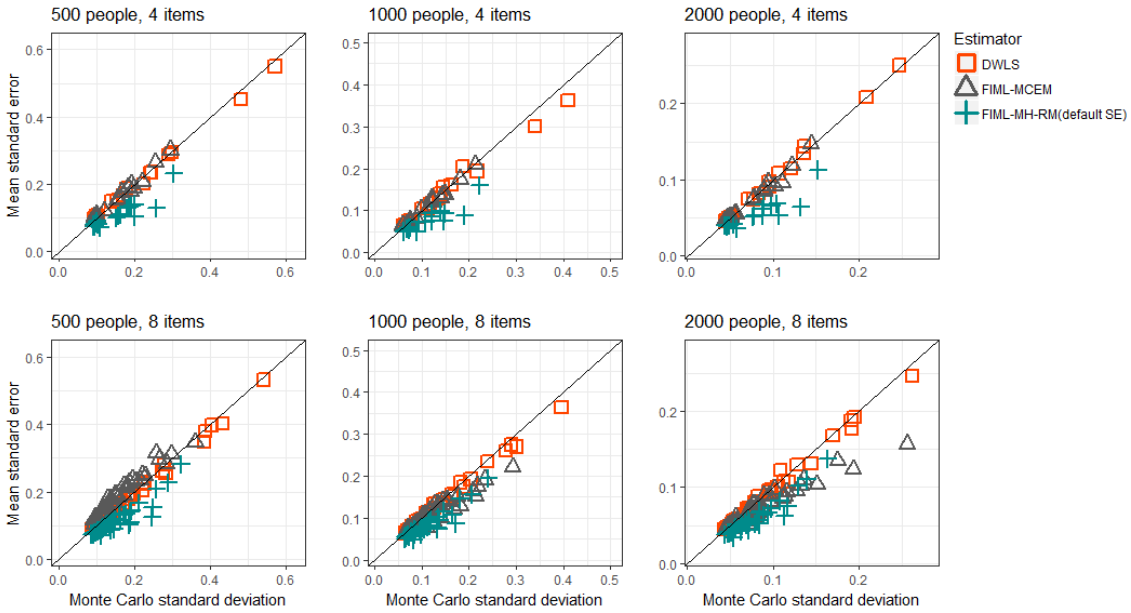

Figure 4.35. Monte Carlo standard deviations of item parameter estimates vs. mean item paramter standard error estimates across sample sizes and test lengths under 10%/wave general MAR attritionin Simulation III. The horizontal line represents nominal level of 95%.

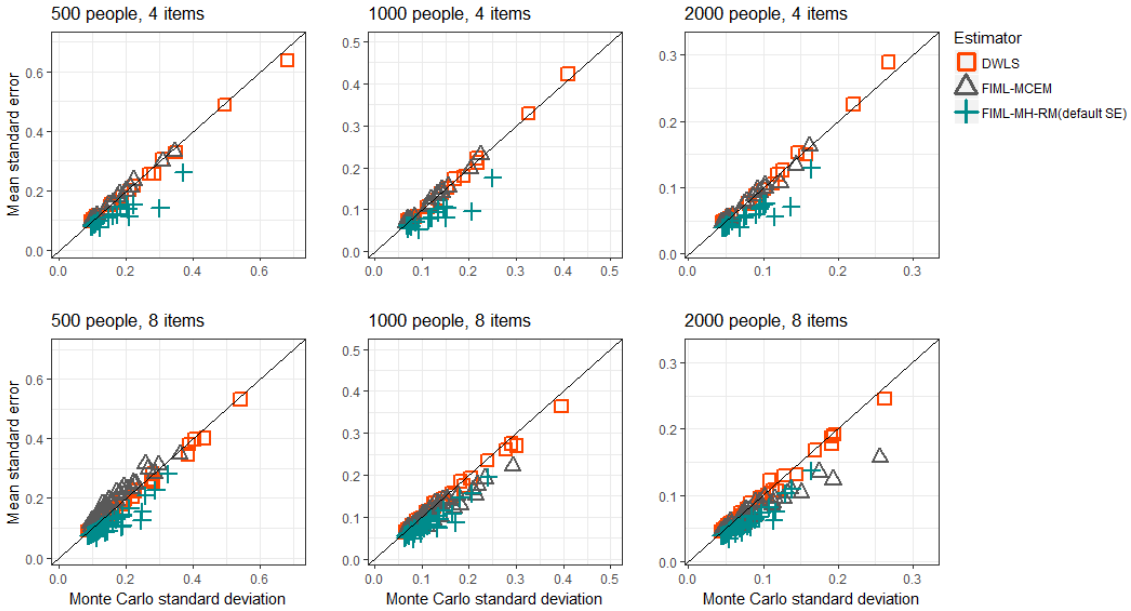

Figure 4.36. Monte Carlo standard deviations of item parameter estimates vs. mean item parameter standard error estimates across sample sizes and test lengths under 20%/wave general MAR attrition in Simulation III. The horizontal line represents nominal level of 95%

Taking both the point estimates and standard errors into account, the coverage rates of the true item parameters in the 95% confidence intervals are plotted in Figure 4.37 to Figure 4.40.

Under both MAR-X and general MAR attrition, FIML-MCEM yielded most accurate coverage rates when the sample size was 200. However, when the sample size increased, the coverage for those item intercepts with large absolute values began to drop noticeably below the nominal level for longer tests. Both DWLS and FIML-MH-RM produced confidence intervals that were too liberal. DWLS yielded generally better coverage rates than FIML-MH-RM under MAR-X. However, when the attrition mechanism was general MAR, the coverage rates produced by DWLS became poorer for some item parameters. The coverage rates of confidence intervals produced by FIML-MH-RM were consistently below the nominal level across all data conditions due to the underestimated standard errors.

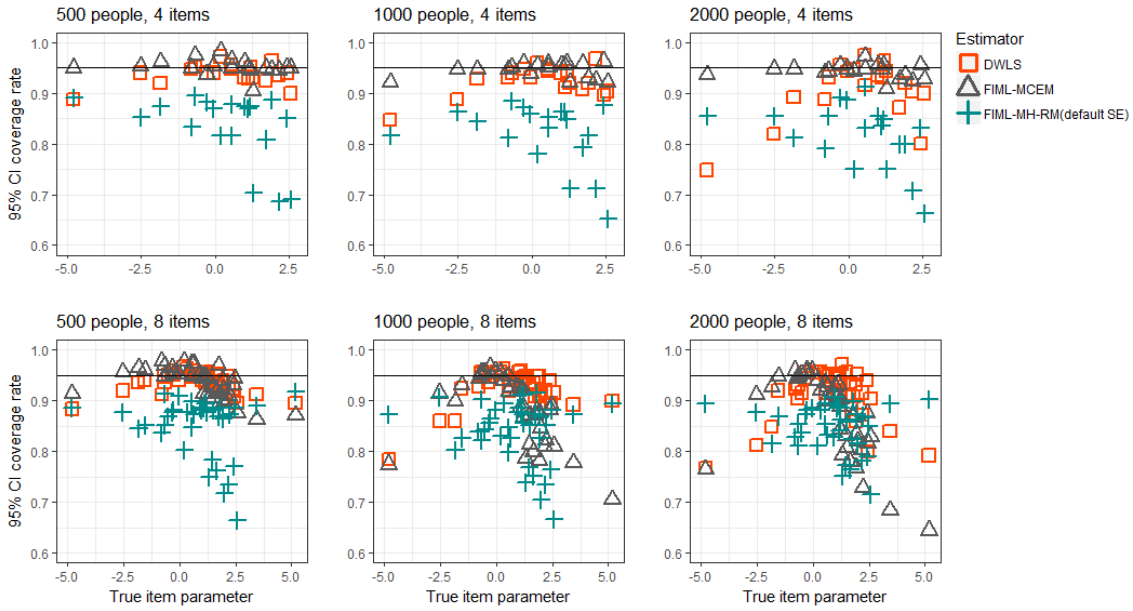

Figure 4.37. Coverage rates of the true item parameters in the 95% confidence intervals across sample sizes and test lengths under 10%/wave MAR-X attrition in Simulation III. The horizontal line represents nominal level of 95%.

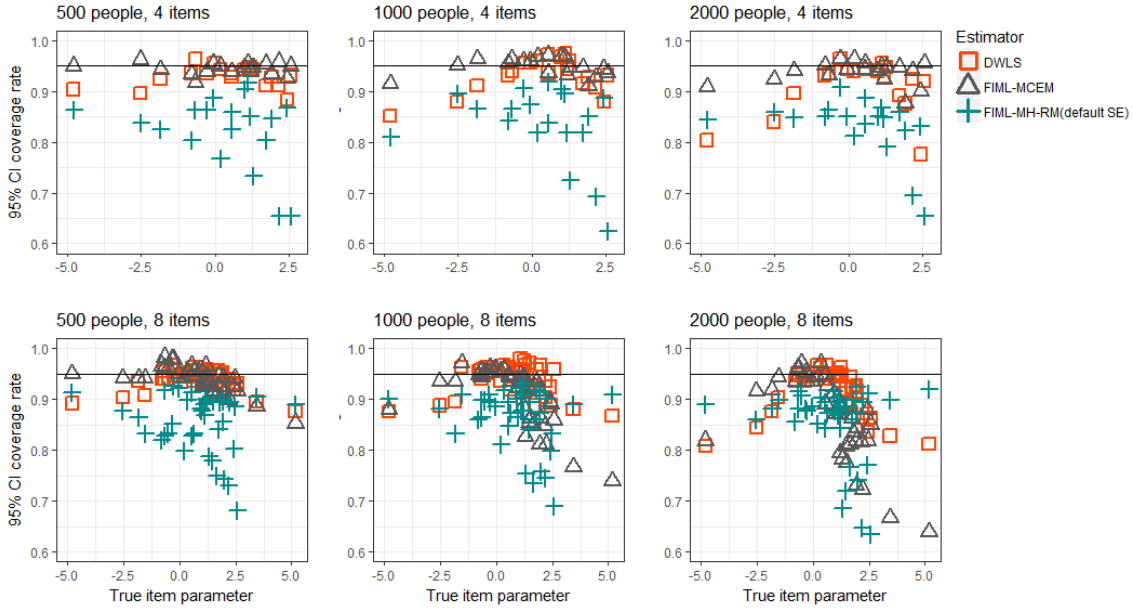

Figure 4.38. Coverage rates of the true item parameters in the 95% confidence intervals across sample sizes and test lengths under 20%/wave MAR-X attrition in Simulation III. The horizontal line represents nominal level of 95%.

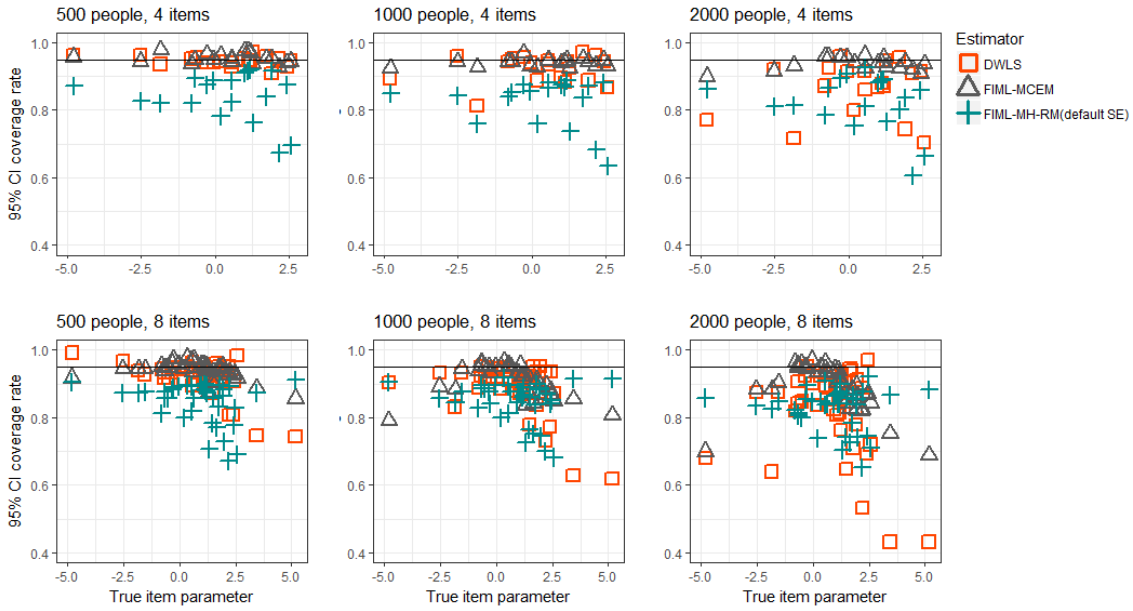

Figure 4.39. Coverage rates of the true item parameters in the 95% confidence intervals across sample sizes and test lengths under 10%/wave general MAR attrition in Simulation III. The horizontal line represents nominal level of 95%.

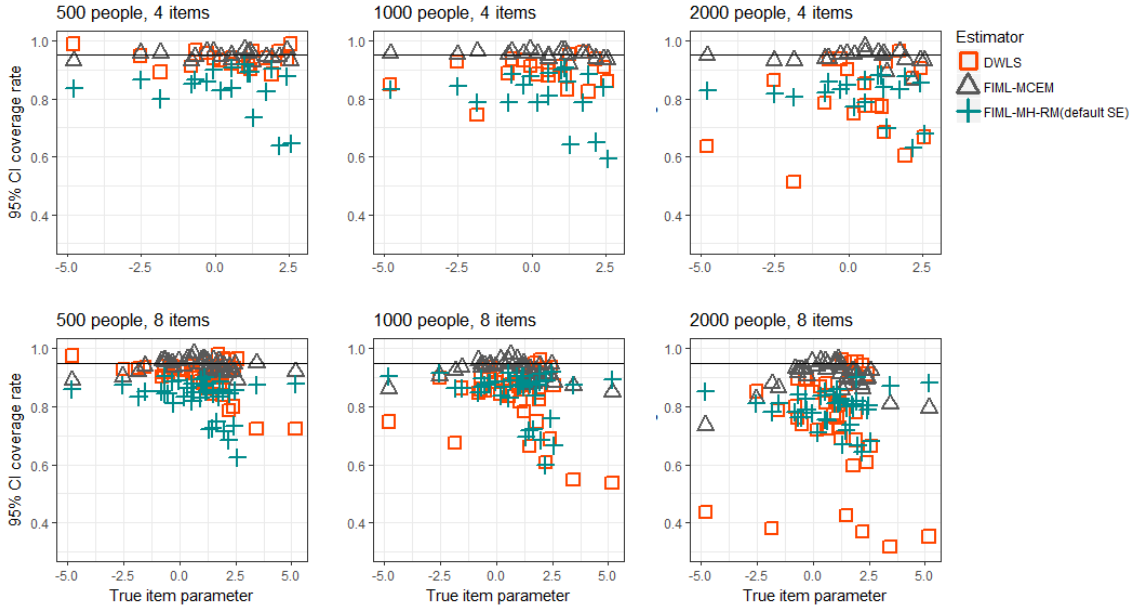

Figure 4.40. Coverage rates of the true item parameters in the 95% confidence intervals across sample sizes and test lengths under 20%/wave general MAR attrition in Simulation III. The horizontal line represents nominal level of 95%.

4.3.3 Structural Parameter Recovery

The mean relative bias for the structural parameter estimates for the two attrition rates (10% and 20%) crossed with the two attrition mechanisms (MAR-X and general MAR) are plotted in Figure 4.41 to Figure 4.44.

Under MAR-X attrition, all three estimation methods were able to yield mostly unbiased (from -7% to 5%) structural parameter estimates under all data conditions. FIML-MH-RM produced the least biased estimates except when the sample size was 1000 or below under shorter tests. DWLS and FIML-MCEM produced estimates that were consistently negatively biased. Overall, the bias of DWLS estimates was larger than that of FIML-MH-RM. FIML-MCEM produced the most biased estimates. The attrition rate under MAR-X did not seem to affect the relative bias produced by all the estimation methods.

Under general MAR attrition, DWLS produced severely biased estimates of latent slope means and the regression coefficients of the latent slope. The bias became more severe when the attrition increased from 10% to 20%. FIML-MCEM produced mostly unbiased structural parameter estimates under all data conditions. FIML-MH-RM yielded unbiased estimates when the MAR attrition rate was 10% per wave, while it produced positively biased (25% to 29%) latent slope mean estimates and negatively biased (-12% to -22%) regression coefficients of the latent slope. The directions of the bias by FIML-MH-RM and DWLS were the same across all conditions. One possible explanation of the pattern is that, when the biased DWLS estimates were used as starting values for FIML-MH-RM estimation, FIML-MH-RM

was not able to completely correct the bias. FIML-MCEM seemed to be able to correct such bias even when the severely biased starting values were used.

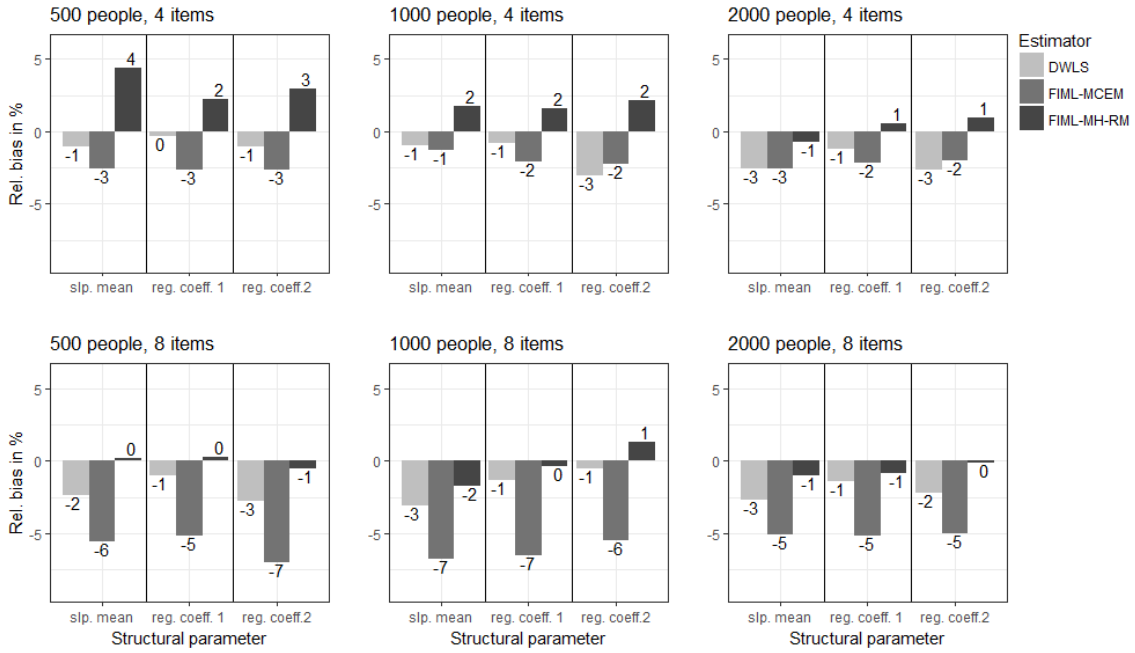

Figure 4.41. Relative bias of structural parameter estimates across sample size and test length under 10% per Wave MAR-X attrition in Simulation III.

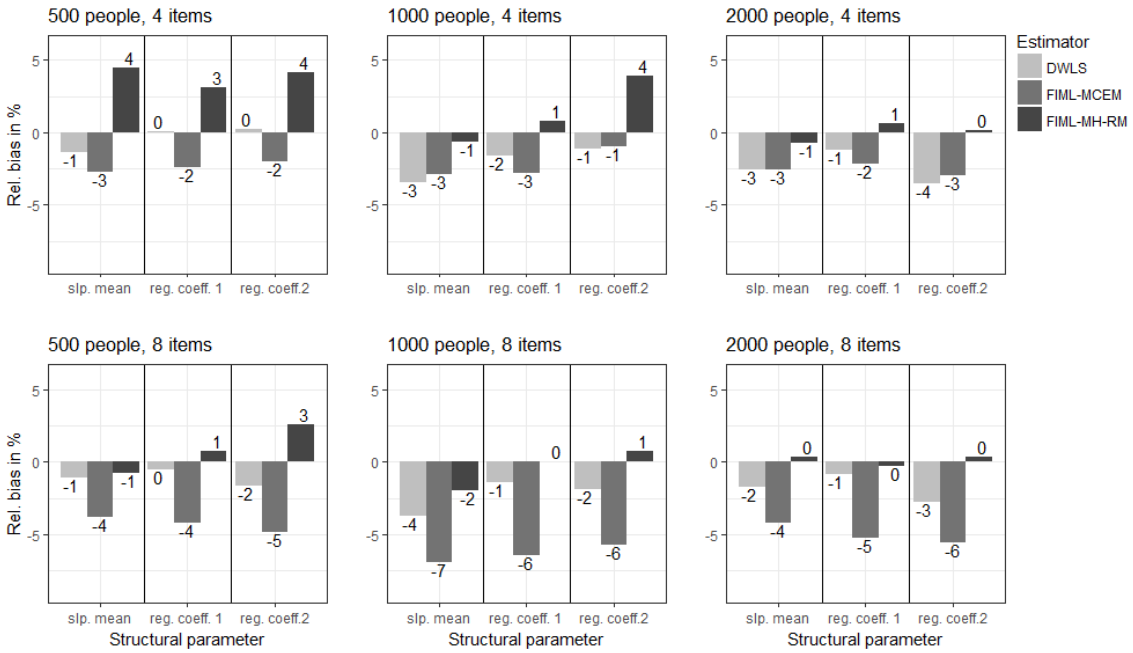

Figure 4.42. Relative bias of structural parameter estimates across sample size and test length under 20% per Wave MAR-X attrition in Simulation III.

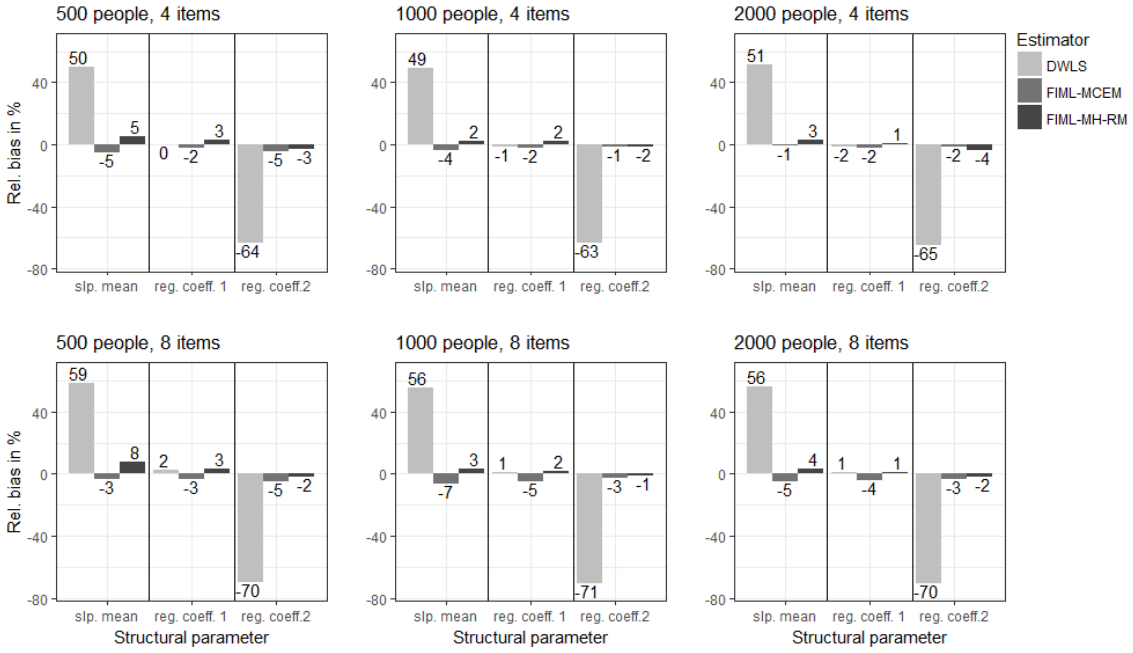

Figure 4.43. Relative bias of structural parameter estimates across sample size and test length under 10% per Wave general MAR attrition in Simulation III.

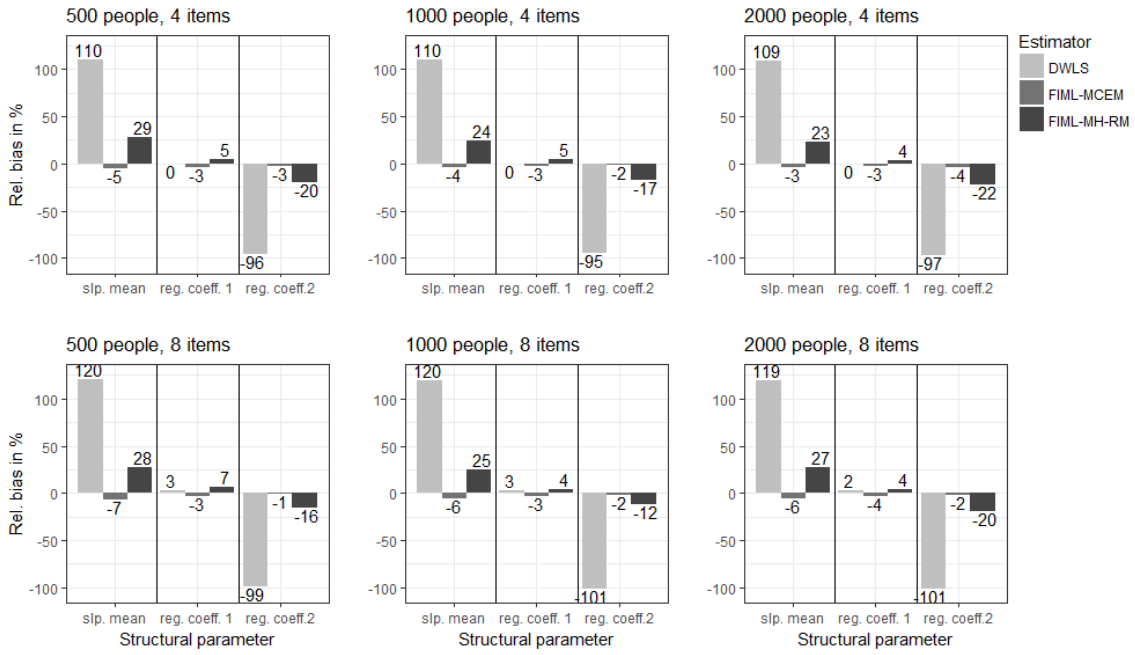

Figure 4.44. Relative bias of structural parameter estimates across sample size and test length under 20% per Wave general MAR attrition in Simulation III.

The root mean square errors of the structural parameter estimates for the two attrition rates (10% and 20%) crossed with the two attrition mechanisms (MAR-X and general MAR) are plotted in Figure 4.45 to Figure 4.48.

Under MAR-X attrition, all three estimation methods yielded mostly comparable RMSEs except that FIML-MCEM produced larger RMSEs for the regression coefficients of the latent intercept. Under general MAR attrition, the RMSEs for the DWLS estimates of the latent slope means and the regression coefficients of the latent slope were considerably larger than the other estimation methods due to the large relative bias of the two estimates.

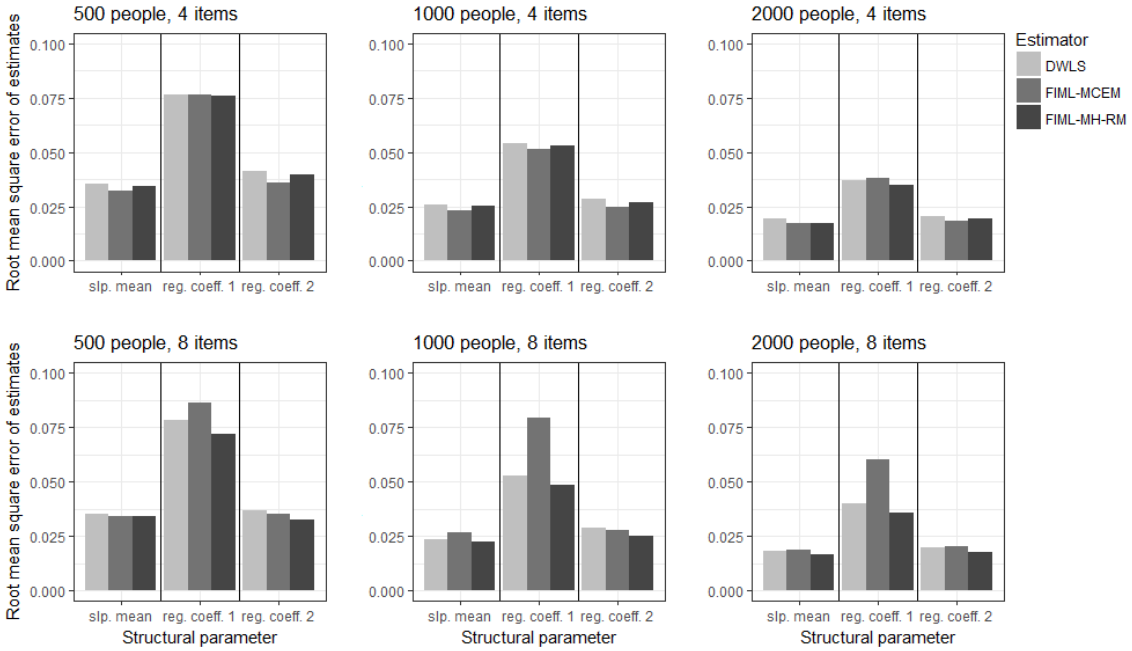

Figure 4.45. Root mean square errors of structural parameter estimates across sample size and test length under 10% per Wave MAR-X attrition in Simulation III.

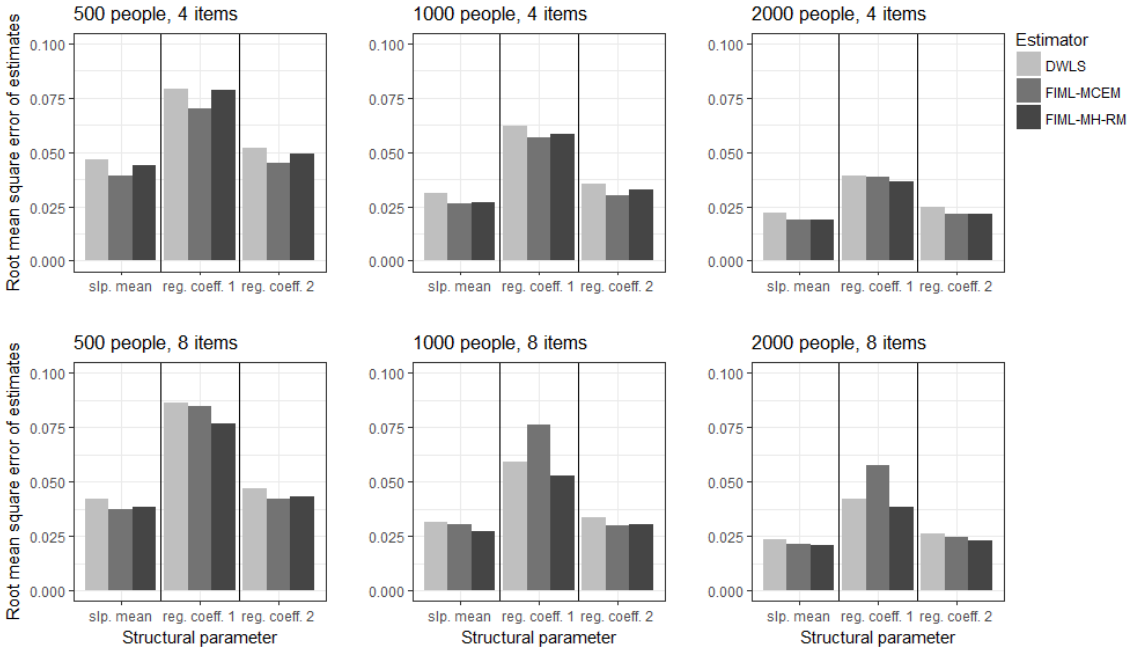

Figure 4.46. Root mean square errors of structural parameter estimates across sample size and test length under 20% per Wave MAR-X attrition in Simulation III.

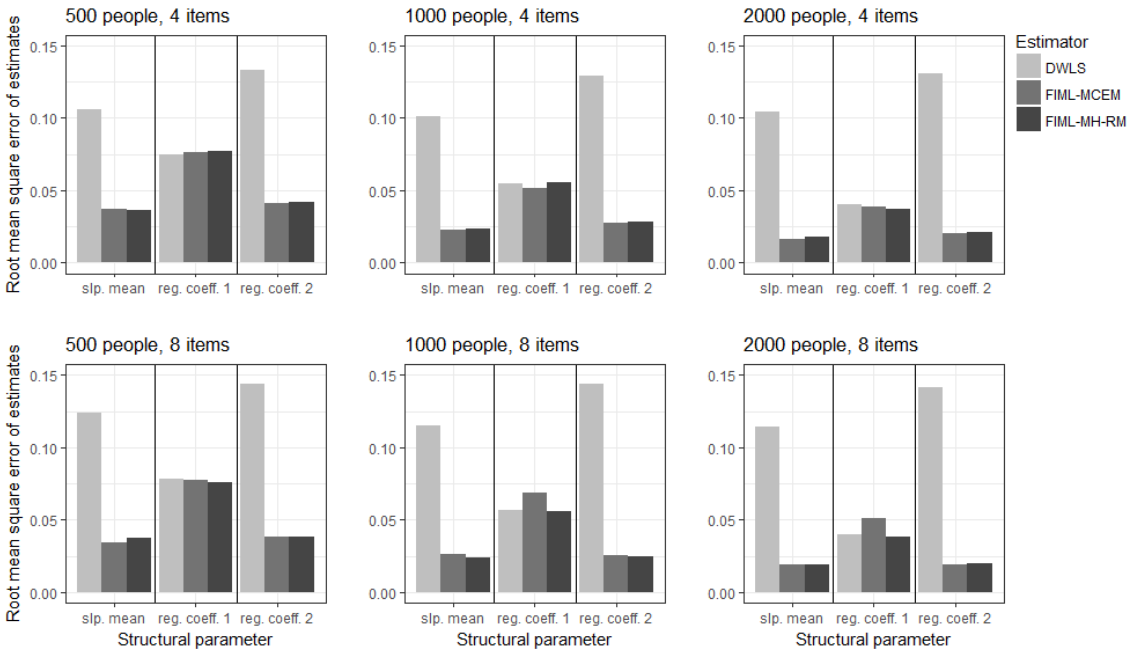

Figure 4.47. Root mean square errors of structural parameter estimates across sample size and test length under 10% per Wave general MAR attrition in Simulation III.

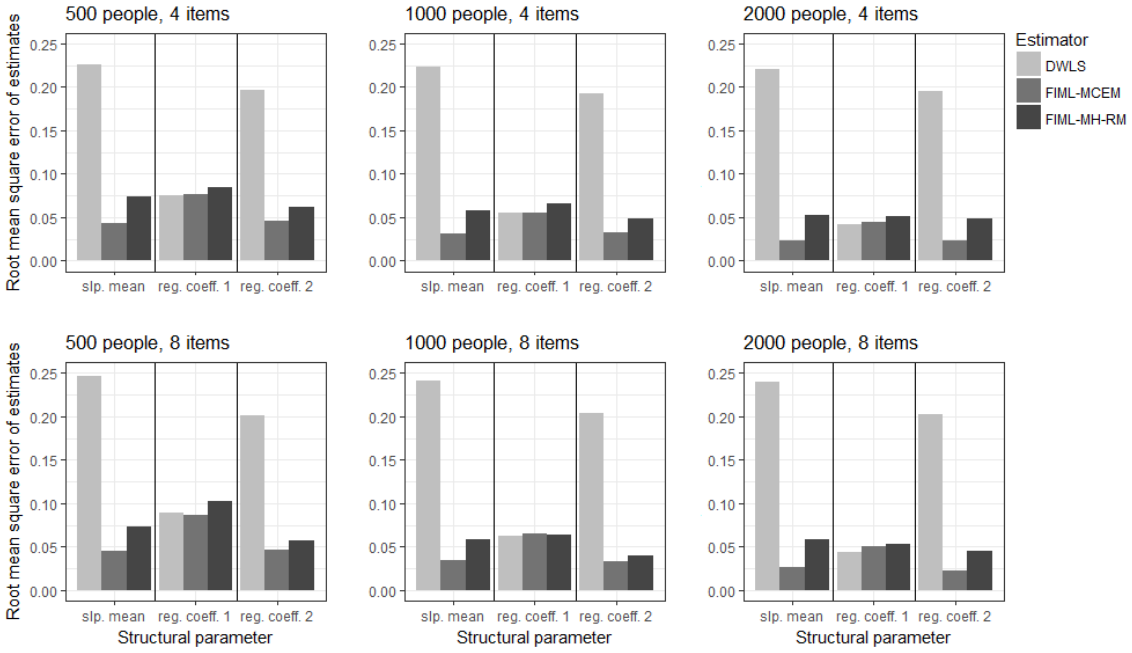

Figure 4.48. Root mean square errors of structural parameter estimates across sample size and test length under 20% per Wave general MAR attrition in Simulation III.

The ratios of the mean structural parameter standard errors to the Monte Carlo standard devaitions of the point estimates are plotted in Figure 4.49 to Figure 4.52. As can be observed from the Figures, the recursively approximated standard errors produced by FIML-MH-RM were severely underestimated which would lead to inflated type-I errors.

Taking both the point estimates and standard errors into account, the coverage rates of the true parameters in the 95% confidence intervals are plotted in Figure 4.53 to Figure 4.56. DWLS yielded appropriate coverage rates under MAR-X attrition, while it failed to cover the latent slope means and the regression coefficients of the latent slopes due to the large magnitude of bias under MAR. FIML-MCEM provided proper coverage for the latent slope means and the regression coefficients of the latent slopes under both MAR-X and general MAR, while it produced overly low coverage rates for the regression coefficients of the latent intercepts when the sample size was 500 or above under longer tests. FIML-MH-RM produced confidence interval coverage rates that were consistently below the nominal level for all structure parameters due to the underestimated standard errors.

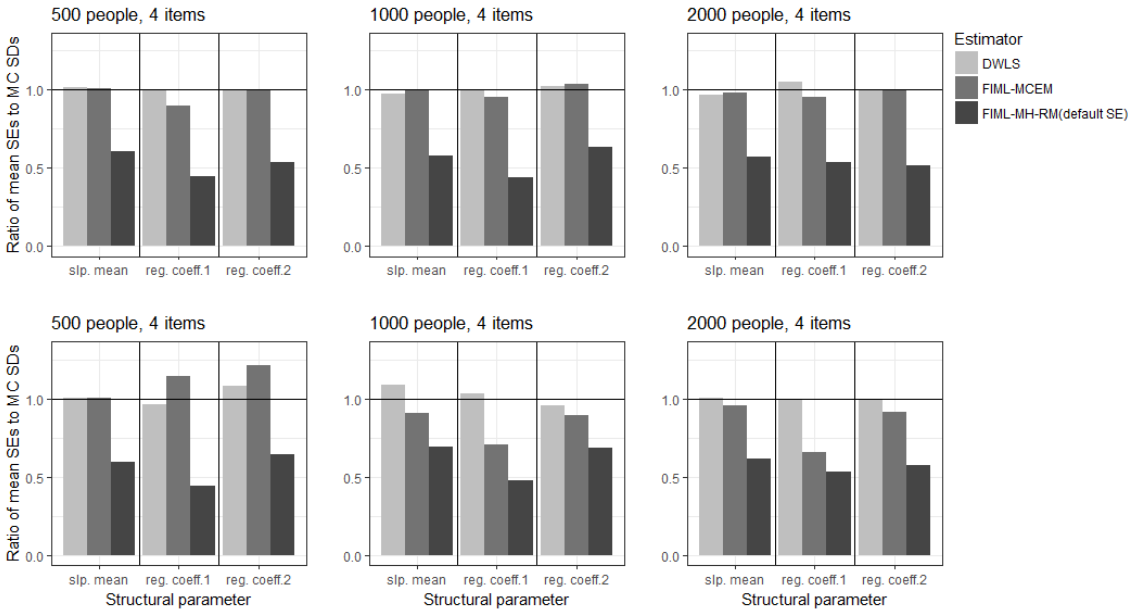

Figure 4.49. Ratio of mean structural parameter standard errors to Monte Carlo standard deviations of point estimates across sample size and test length under 10% per Wave MAR-X attrition in Simulation III.

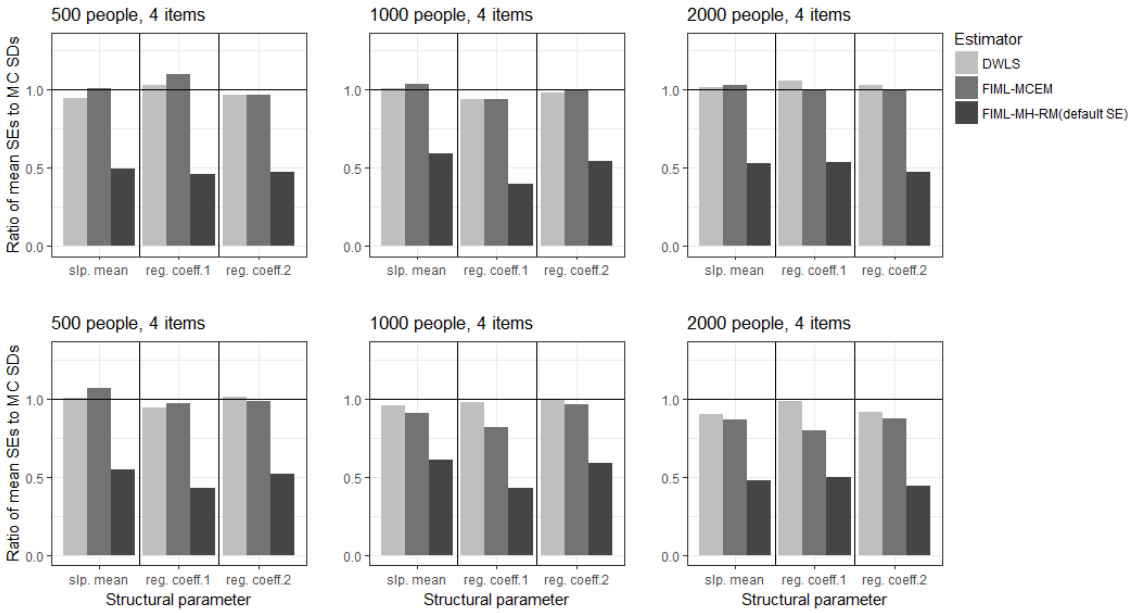

Figure 4.50. Ratio of mean structural parameter standard errors to Monte Carlo standard deviations of point estimates across sample size and test length under 20% per Wave MAR-X attrition in Simulation III.

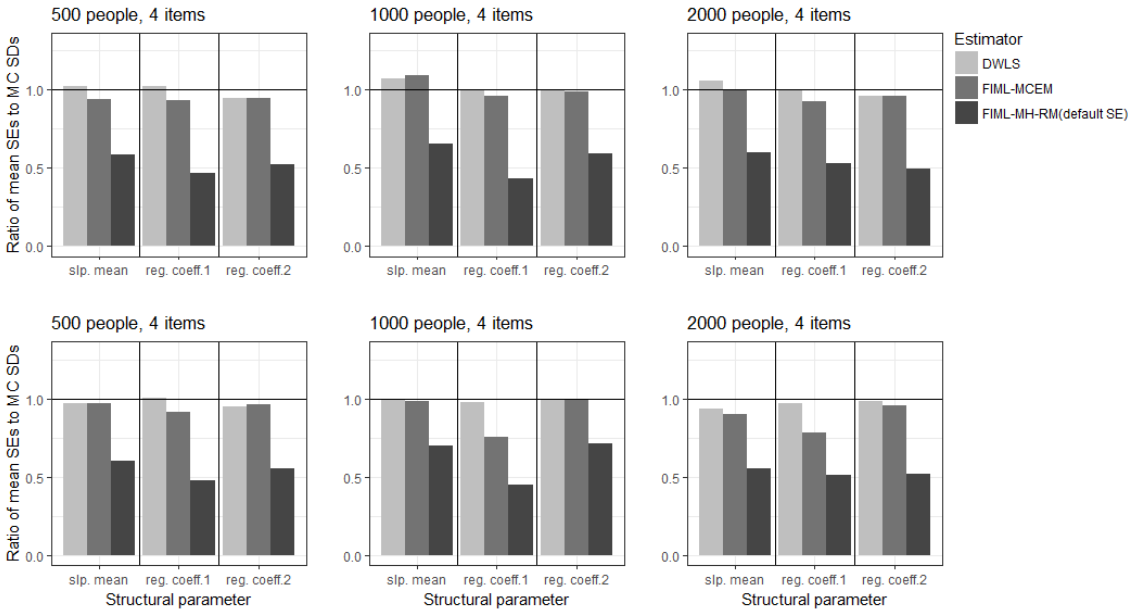

Figure 4.51. Ratio of mean structural parameter standard errors to Monte Carlo standard deviations of point estimates across sample size and test length under 10% per Wave general MAR attrition in Simulation III.

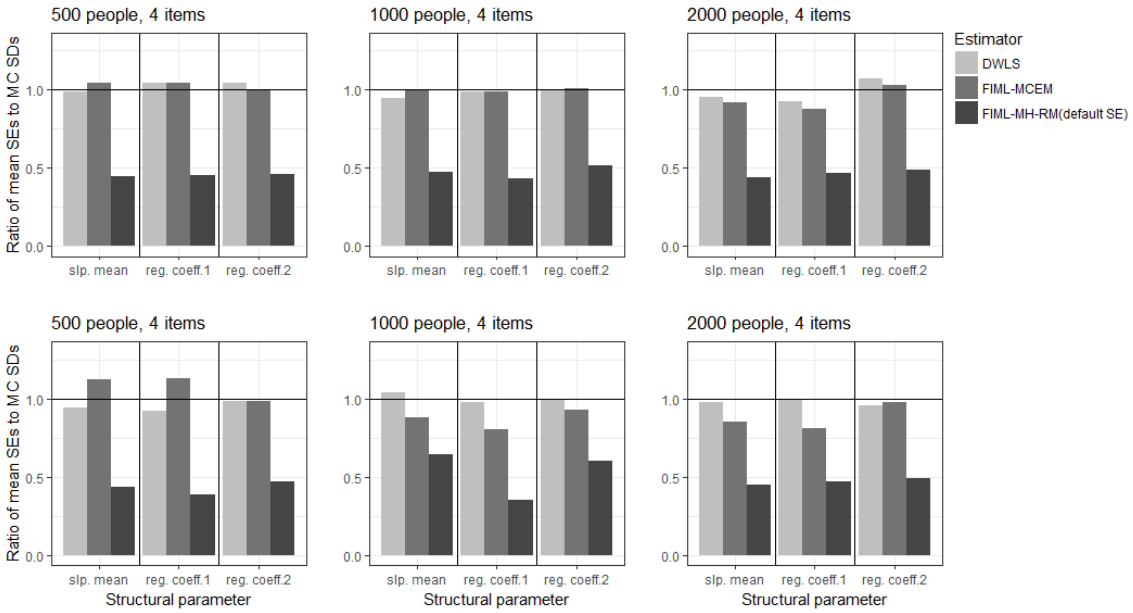

Figure 4.52. Ratio of mean structural parameter standard errors to Monte Carlo standard deviations of point estimates across sample size and test length under 20% per Wave general MAR attrition in Simulation III.

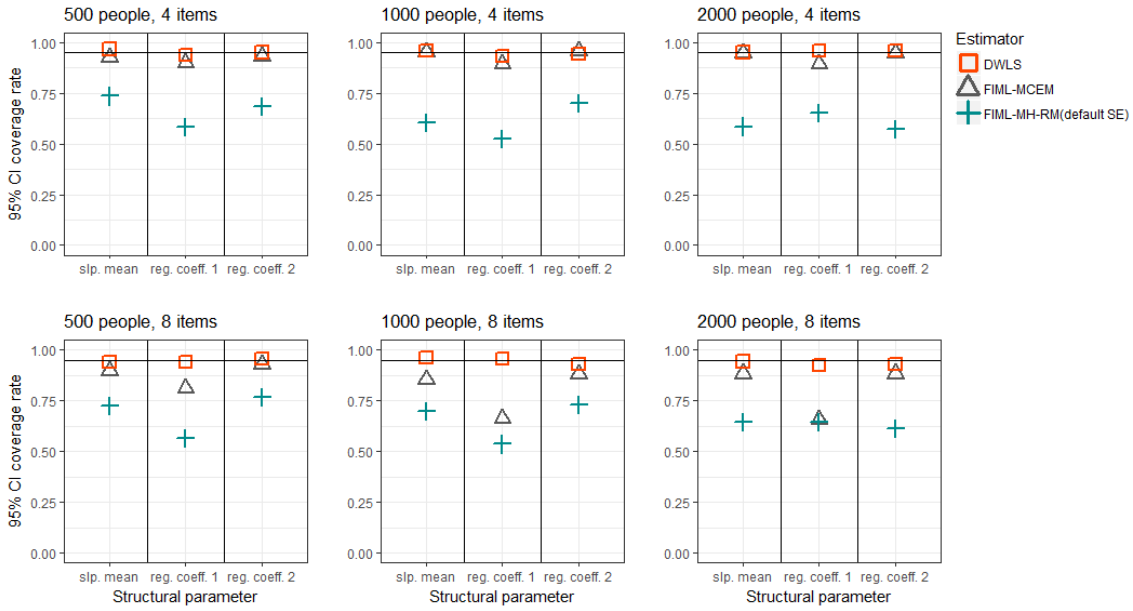

Figure 4.53. Coverage rates of the true structural parameters in the 95% confidence intervals across sample size and test length under 10% per Wave MAR-X attrition in Simulation III.

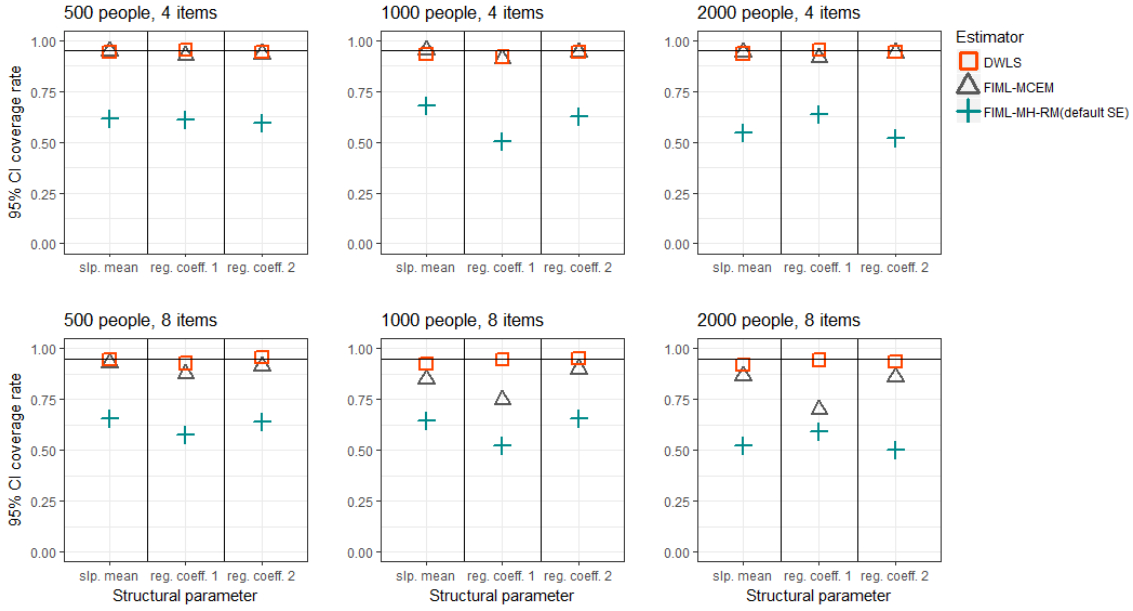

Figure 4.54. Coverage rates of the true structural parameters in the 95% confidence intervals errors to Monte Carlo standard deviations of point estimates across sample size and test length under 20% per Wave MAR-X attrition in Simulation III.

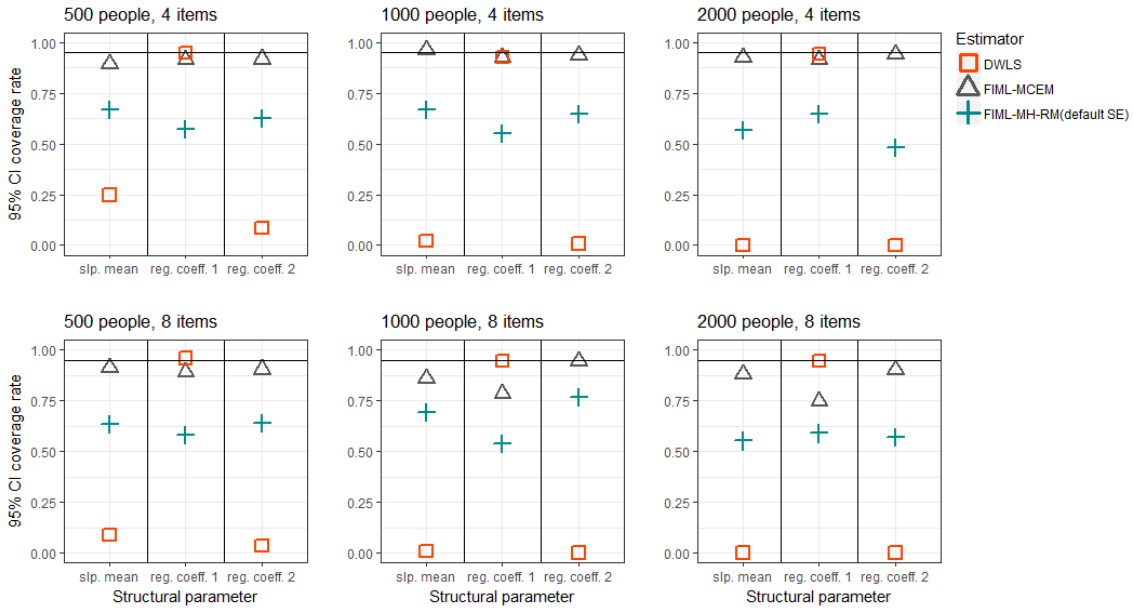

Figure 4.55. Coverage rates of the true structural parameters in the 95% confidence intervals across sample size and test length under 10% per Wave general MAR attrition in Simulation III.

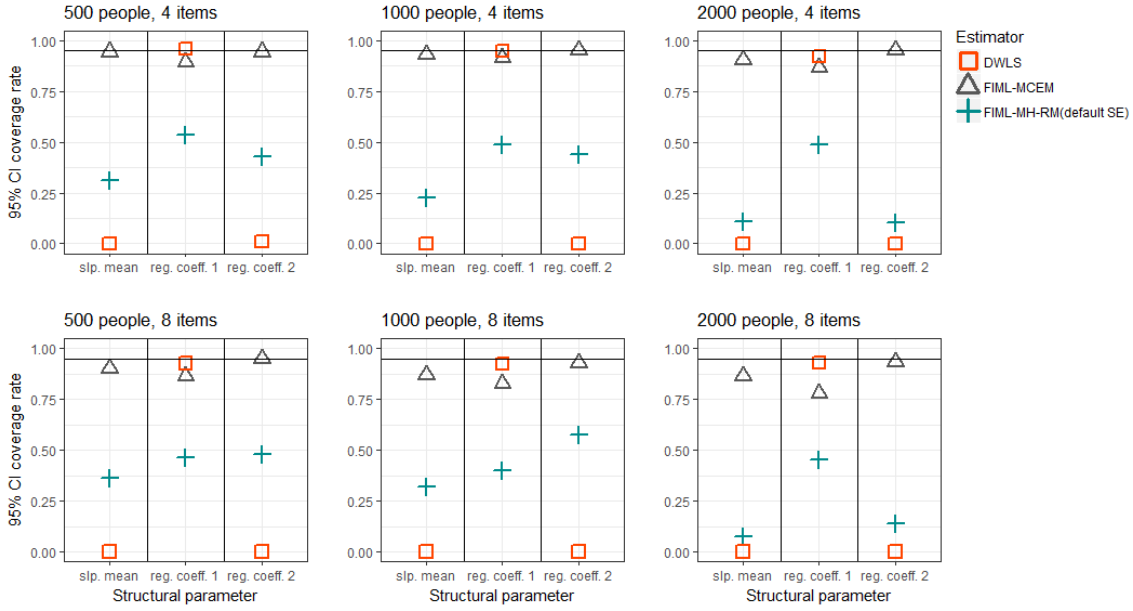

Figure 4.56. Coverage rates of the true structural parameters in the 95% confidence intervals across sample size and test length under 20% per Wave general MAR attrition in Simulation III.

4.4 Empirical Example

The "Language and Literacy" rating scale data from the "Multistate Study of Pre-Kindergarten 2001–2003" by National Center for Early Development and Learning (Clifford, Bryant, Burchinal, & Barbarin, 2005) were used to illustrate the application of different estimation strategies discussed before, namely using alternative estimation methods other than the FIML-BAEM algorithm and using the two reduced models.

4.4.1 Data

The "Language and Literacy" skills of 1015 young children in the United States from pre-kindergarten to kindergarten were evaluated across four semesters from Fall 2001 (Fall semester of pre-kindergarten) to Spring 2003 (Spring of kindergarten) in approximately equal intervals. The children were sampled using a stratified random sampling design to be representative of prekindergarten children in six participating states, including California, Illinois, New York, Ohio, Kentucky, and Georgia.

In each semester, teachers were asked to rate the children's language and literacy levels using the same nine Likert-scale items. The items are:

- Uses complex sentence structures.
- Understands and interprets a story or other text read to him/her.
- Easily and quickly names all upper– and lower-case letters of the alphabet.
- Produces rhyming words.

- Predicts what will happen next in stories by using the pictures and storyline for clues.
- Reads simple books independently.
- Demonstrates early writing behaviors.
- Demonstrates an understanding of some of the conventions of print.
- Uses the computer for a variety of purposes.

All items had five categories. The ratings of 1 to 5 corresponded to proficiency levels of "not yet", "beginning", "in progress","intermediate" and "proficient".

Both intermittent missing and permanent attrition were prevalent in this assessment. Of the 1015 children who participated in the assessment, only 226 had complete data across all four waves.

In the preliminary data exploration, the means and variances of the average scores were calculated for available participants at each time point. The means for the four time points were 2.09, 2.79, 2.40, 3.67, with variances of 6.50, 8.47, 8.38, and 9.12 respectively. The growth of language skills for the children did not follow a linear curve. There was a noticeable drop from the Spring of pre-kindergarten to the Fall of kindergarten. The phenomenon where a large portion of students lose what they have learned during academic semesters over the long summer break is referred to as "summer achievement loss" (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996; Entwisle & Alexander, 1992). In an empirical study using the same data, Paek et al. (2016) used a piece-wise solution to estimate and compare

the latent abilities at each time point. However, no general statement was made regarding growth in different periods. In this study, a linear growth LGM-IRT model was first fitted to the data. The data was then analyzed with the "summer effect" model developed by Raudenbush and Bryk (2002) to account for summer loss by estimating different growth rates for different periods. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) of the two models were compared to determine the final analytic model. In the "summer effect" model, the time values of the latent slope were set to (0, 1, 1, 2) instead of (0, 1, 2, 3) in a linear LGM. Additionally, a fixed effect (i.e., mean structure) was imposed on the four time points with loadings of (0, 0, 1, 1). With this setup, the random slope should be interpreted as a child's growth rate during the period from Fall semester to the Spring semester during pre-kindergarten or kindergarten academic year. The added mean structure should be interpreted as the average growth rate of all children from Spring semester of pre-kindergarten to Fall semester of kindergarten when the long summer break took place. This full LGM-IRT with the data using the second-order formation had 13 dimensions (four time-specific factors and nine common item effects), while the first-order formation had 15 dimensions (latent slope factor, latent intercept factor, four time-specific disturbances and nine common item effects). It should be noted that adding the mean structure did not add to the dimensionality of the model.

4.4.2 Results

The 95% confidence intervals of the AIC and BIC produced by FIML-MH-RM for the linear model were (65487.69, 65605.84) and (65787.97, 65906.12), respectively, while the same indices for the "summer effect" model were (63766.22, 63883.47) and (64071.42, 64188.67), respectively. The "summer effect" model yielded better fit than the linear model as indicated by the smaller AIC and BIC. Thus, the "summer effect" model was used as the final model.

The final model was estimated using the three estimation methods. Additionally, the reduced models without the common item effects or the time-specific disturbances were estimated using FIML-BAEM after dimension reduction techniques were implemented to reduce the dimensionality of the model to three. It was found that the post-convergence approximated standard errors of FIML-MH-RM estimates could not be produced with 30,000 iterations. Therefore, only the recursively approximated standard errors are reported. Additionally, the reduced model without time-specific disturbances could not converge, the results of which are not reported either.

The structural and item parameter estimates using the three estimation methods are presented in Table 4.13 and Table 4.14, respectively. Among the four methods, DWLS produced the largest growth rate as well as the largest summer setback of children's literacy and language skills. FIML-MH-RM yielded the second largest growth rate and summer effect. The two FIML algorithms and the misspecified model returned similar summer effect estimates, which were all smaller than the

DLWS estimate. All methods returned similar latent slope variance. As for covariance between latent slope and intercept, all methods returned negative estimates. However, DWLS and FIML-MH-RM produced smaller covariance than the other two methods. It is worth noting that, FIML-MCEM structual estimates were very similar to the reduced model, while DWLS estimates agreed with FIML-MH-RM more. The similarity between DWLS and FIML-MH-RM may be due to that FIML-MH-RM was sensitive to the starting values, which were DWLS estimates. The similarity between FIML-MH-RM and the reduced model may be due to the fact that the common item effect variances were small relative to the structural parameters. The differences in structural estimates create different interpretations of the growth rates and summer effects. According to the DWLS results, students grew their literacy skills very fast in the academic year but lost much of what they learned in the summer. Based on FIML-MH-RM results, the students' literacy skills grew very fast during academic years as well. But they lost less than indicated by the DWLS estimates. If the estimates of FIML-MCEM and the reduced model were to be trusted, the children grew slower during regular academic year but also lost less during the summer break.

Table 4.13

Structural Parameter Estimates of Empirical Example using Different Methods

	DWLS	FIML-MCEM	FIML-MH-RM	R1
	EST (SE)	EST (SE)	EST (Default SE)	EST (SE)
Latent intercept mean	0 (–)	0 (–)	0 (–)	0 (–)
Latent slope mean	1.49 (0.08)	0.98 (0.04)	1.41 (0.02)	0.99 (0.04)
Latent intercept variance	1 (–)	1 (–)	1 (–)	1 (–)
Latent slope variance	0.25 (0.05)	0.22 (0.02)	0.24 (0.01)	0.23 (0.02)
Covariance of intercept and slope	-0.11 (0.06)	-0.31 (0.02)	-0.15 (0.02)	-0.30(0.02)
Mean summer effect	-0.62 (0.07)	-0.35 (0.04)	-0.40 (0.02)	-0.37(0.05)

Note. Default SE=recursively approximated standard error; R1=reduced model by omitting common item effects.

In terms of item parameters, the four methods yielded different estimates. The rank order of item slopes from largest to smallest is FIML-MCEM, DWLS, the reduce model, and FIML-MH-RM. Considerable differences were observed in the item intercept estimates too. Noticeably, DWLS tended to return item intercept estimates with very large absolute values, while the reduced model returned the smallest intercept estimates. The item intercept estimates of FIML-MCEM and FIML-MH-RM were mostly comparable. The patterns of the item intercept estimates generally confirmed the patterns observed in Simulation III.

In sum, without knowing the true parameters, it is difficult to judge which set of estimates were closer to the true parameters. However, based on the results of Simulation III, FIML-MCEM yielded acceptable parameter recovery across all conditions. Additionally the reduced model, which was found to produce robust structural parameter estimates in Simulation II, agreed more with the FIML-MCEM structural estimates. Based on these observations, stronger argument can be made that the FIML-MCEM estimates were less biased.

Table 4.14

Item Parameter Estimates of Empirical Example using Different Methods

	DWLS	FIML-MCEM	FIML-MH-RM	R1
	EST (SE)	EST (SE)	EST (Default SE)	EST (SE)
α_1	1.99 (0.13)	1.67 (0.08)	1.13 (0.03)	1.50 (0.02)
α_2	3.18 (0.25)	2.28 (0.11)	1.55 (0.04)	2.06 (0.06)
α_3	2.85 (0.29)	3.27 (0.13)	2.28 (0.06)	2.63 (0.05)
α_4	3.01 (0.30)	3.52 (0.16)	2.41 (0.07)	3.09 (0.05)
α_5	2.71 (0.19)	3.05 (0.15)	2.13 (0.06)	2.82 (0.07)
α_6	1.67 (0.11)	2.43 (0.1)	1.69 (0.05)	2.11 (0.06)
α_7	2.32 (0.19)	3.17 (0.14)	2.25 (0.07)	2.79 (0.06)
α_8	1.76 (0.12)	2.88 (0.13)	1.99 (0.06)	2.50 (0.05)
α_9	1.13 (0.07)	1.89 (0.09)	1.32 (0.04)	1.66 (0.06)
λ_{11}	1.59 (0.12)	1.46 (0.11)	1.41 (0.06)	1.35 (0.04)
λ_{12}	-0.56 (0.12)	-0.07 (0.09)	-0.15 (0.06)	-0.07 (0.04)
λ_{13}	-2.88 (0.18)	-1.70 (0.10)	-1.81 (0.07)	-1.58 (0.03)
λ_{14}	-5.03 (0.25)	-3.18 (0.11)	-3.32 (0.09)	-2.96 (0.04)
λ_{21}	3.81 (0.27)	2.57 (0.14)	2.53 (0.08)	2.47 (0.05)
λ_{22}	-0.71 (0.18)	-0.07 (0.11)	-0.16 (0.06)	-0.04 (0.05)
λ_{23}	-4.51 (0.34)	-2.26 (0.12)	-2.40 (0.08)	-2.13 (0.07)
λ_{24}	-8.3 (0.56)	-4.43 (0.15)	-4.63 (0.11)	-4.22 (0.08)
λ_{31}	1.28 (0.18)	1.06 (0.15)	0.97 (0.09)	1.02 (0.05)
λ_{32}	-2.12 (0.24)	-1.43 (0.14)	-1.64 (0.09)	-1.17 (0.03)
λ_{33}	-4.84 (0.43)	-3.54 (0.17)	-3.85 (0.13)	-3.01 (0.06)
λ_{34}	-7.21 (0.60)	-5.32 (0.2)	-5.73 (0.16)	-4.58 (0.09)
λ_{41}	0.19 (0.16)	0.32 (0.16)	0.20 (0.07)	0.42 (0.06)
λ_{42}	-3.32 (0.30)	-2.34 (0.16)	-2.55 (0.10)	-2.09 (0.05)
λ_{43}	-6.57 (0.53)	-4.96 (0.21)	-5.26 (0.16)	-4.56 (0.04)
λ_{44}	-9.44 (0.73)	-7.19 (0.25)	-7.6 (0.21)	-6.68 (0.09)
λ_{51}	3.20 (0.20)	2.69 (0.17)	2.69 (0.09)	2.71 (0.07)
λ_{52}	-0.89 (0.15)	-0.51 (0.14)	-0.65 (0.07)	-0.43 (0.07)
λ_{53}	-4.05 (0.24)	-3.05 (0.15)	-3.31 (0.10)	-2.95 (0.09)
λ_{54}	-7.50 (0.38)	-5.80 (0.20)	-6.21 (0.15)	-5.68 (0.14)
λ_{61}	-0.67 (0.1)	-0.46 (0.12)	-0.58 (0.07)	-0.32 (0.05)
λ_{62}	-2.83 (0.15)	-2.47 (0.14)	-2.67 (0.09)	-2.22 (0.07)
λ_{63}	-4.58 (0.2)	-4.27 (0.18)	-4.55 (0.13)	-3.92 (0.10)
λ_{64}	-6.19 (0.26)	-5.92 (0.22)	-6.28 (0.16)	-5.49 (0.13)
λ_{71}	-1.26 (0.15)	-0.88 (0.15)	-1.09 (0.08)	-0.71 (0.07)
λ_{72}	-3.91 (0.25)	-3.25 (0.19)	-3.58 (0.11)	-2.98 (0.09)
λ_{73}	-6.08 (0.36)	-5.33 (0.24)	-5.8 (0.16)	-4.98 (0.11)
λ_{74}	-8.04 (0.45)	-7.26 (0.30)	-7.87 (0.21)	-6.83 (0.14)
λ_{81}	-0.79 (0.11)	-0.62 (0.14)	-0.79 (0.07)	-0.47 (0.06)
λ_{82}	-2.98 (0.16)	-2.90 (0.18)	-3.14 (0.11)	-2.63 (0.08)
λ_{83}	-4.81 (0.22)	-4.96 (0.24)	-5.28 (0.16)	-4.59 (0.11)
λ_{84}	-6.68 (0.29)	-7.04 (0.29)	-7.44 (0.21)	-6.57 (0.13)
λ_{91}	0.93 (0.09)	1.01 (0.11)	0.97 (0.07)	1.03 (0.06)
λ_{92}	-1.23 (0.09)	-1.18 (0.10)	-1.33 (0.07)	-1.04 (0.06)
λ_{93}	-3.18 (0.12)	-3.25 (0.13)	-3.50 (0.10)	-3.00 (0.09)
λ_{94}	-4.80 (0.16)	-5.05 (0.17)	-5.40 (0.14)	-4.73 (0.12)

Note. Default SE=recursively approximated standard error; R1=reduced model by omitting common item effects.

Chapter 5: Discussions

In this chapter, the research context, purposes, and the results are summarized and discussed. In addition, limitations of the current research are identified. Finally, potential future research areas are discussed.

5.1 Summary

Measuring change in a construct over time has been an active area in educational and psychological research. It is often achieved by administering the same (subset of) items to the same respondents repeatedly over time. When the response data are continuous, the second-order latent growth model (McArdle, 1988) has been adopted to investigate change on the latent level. When response data are categorical, using an item response theory model as the measurement model in a second-order LGM is a natural extension of the second-order LGM with continuous indicators. However, application of the LGM-IRT is limited due to several methodological issues. This study investigated three issues in using LGM-IRT, namely model parameterization, estimation of model parameters, and sample attrition.

The first goal of the research was to provide a review of selected longitudinal IRT models, with special attention paid to the parameterizations and interrelations

of selected longitudinal IRT models. A total of seven models was reviewed, including multilevel IRT model that decompose person parameters, within-item MIRT model, simple structure correlated factors model, two-tier item factor model, LGM-IRT with no local dependence consideration, LGM-IRT with order local dependence, and LGM-IRT with common item effects. It was revealed that the different types of LGM-IRT models were extensions of their respective MIRT models. The LGM-IRT models can be transformed from their second-order parameterization using the Schmid-Leiman (Schmid & Leiman, 1957) transformation so that they can be estimated using common multidimensional IRT software packages. It was further confirmed in the simulation studies that the transformation yielded equivalent models. However, it should be noted that the transformed model used in this study is an extremely constrained model, in which the item loadings were all constrained to be the same for time-specific latent variables, common item effects, and the disturbances. In a more general model, the loadings of the common item effects could be freely estimated when the model identification condition is imposed on variances as a function of factor loadings.

The second goal of the research was to compare the performance of three estimation methods (namely FIML-MCEM, FIML-MH-RM, and DWLS) in estimating the LGM-IRT with common item effects with no attrition. It was found that all three estimation methods were able to yield sufficiently unbiased item and structural parameter estimates. FIML-MH-RM outperformed the other two estimation methods in terms of the relative bias and RMSE of model estimates. DWLS estimation could produce the results with a fraction of the time required by the other

two estimation methods. In addition, DWLS yielded more varying estimates across replications. But the larger standard errors produced by DWLS adequately capture the variability. As a results, the confidence interval coverage rates produce by DWLS were also appropriate. FIML-MCEM performance was also acceptable in terms of bias, RMSE, and adequacy of standard errors. However, the time efficiency was much more problematic with FIML-MCEM. Given that applied researchers generally need to fit similar models multiple times to find the best-fitting model, this can be a serious limitation.

The standard error estimates of FIML-MH-RM were less ideal. The recursively approximated and post-convergence approximated standard errors were both consistently underestimated. Previous research (Cai, 2008) found that the recursively approximated standard errors were underestimated in a unidimensional model and a bifactor model. The results of this research regarding the recursively approximated standard errors were similar to those of Cai (2008). Additionally, this research found that the standard errors produced by the post-convergence approximation method were less underestimated than those produced by the recursive method. The results were different from previous research of the two methods (Yang & Cai, 2014). Yang and Cai (2014) examined the performance of the FIML-MH-RM algorithm in estimating the contextual effect in a nonlinear multilevel latent variable model and found that the recursively approximated standard errors were closer to the Monte Carlo standard deviations of the point estimates than the post-convergence approximated standard errors. The main drawback of the post-convergence method was that the standard errors for item intercepts were consistently underestimated. The

different conclusions regarding the comparison of the two standard error estimation methods might have been due to the different models being examined.

The third goal of the research was to assess the performance of the two reduced models under complete data. The first reduced model omitted the common item effects. It was found that when common item effects were omitted, the structural parameter estimates were slightly biased. The bias in the item parameter estimates increased when the omitted common item effect variance became larger. The item parameters could be underestimated by over 30% for longer tests. The model ran into convergence problem when the time-specific disturbance variances were small (0.25). One possible explanation is that when the time-specific disturbances were small, the model-implied variances of the latent factors at each time point were also small. As a result, the common item effect variances were large relative to the structural factors. Thus, when larger common item effects were omitted, the model-data misfit became larger and convergence issues occurred. If the model could converge, the structural point estimates might be of some use to researchers. However, it should be cautioned that the confidence interval coverage of the structural parameters might not be proper based on the simulation results, especially when sample size is large (2,000).

When time-specific disturbances were omitted, the model encountered convergence problem when the sample size was small and when the common item effect variances were small (0.25). This might be due to that the relative magnitudes of the common item effect variances (which were retained in the reduced model) were too small to be stably estimated. Similar phenomenon is found when the between-

level variance is small in multilevel model estimation. When the sample size was 500 for long test, the model was easier to converge when the disturbance variances were 1.00 than when the disturbance variances were 0.5. This might be because that some specific combinations of disturbance variances, omitted common item effects, and test length might make the structural part easier to estimate than other combinations. However, the exact mechanism was not yet clear and deserves further examination. It was also possible that non-convergence was due to the stabilization mechanisms in flexMIRT. For example, there might be no mechanism to prevent the variance or covariance of latent variable from going to unreasonable values. The exact mechanism that influenced the convergence required further investigation. As for recovery of model parameters, the item parameters were somewhat biased (-3% to -21%), while the structural parameter estimates were more biased (e.g., the bias in latent slope variance estimates ranged from -35% to -126%). Since the structural parameters are usually of primary interest to researchers, this method is not recommended to researchers who would like to make valid inference of the analysis.

The fourth goal of the research was to compare the performance of the three estimation methods in estimating a conditional LGM-IRT with common item effects when attrition occurs under MAR. Two types of MAR mechanisms were considered, namely, general MAR with respect to both covariates and observed outcomes, and MAR with respect to covariates only (MAR-X). The simulation results showed that all three estimation methods were consistent under MAR-X. FIML-MH-RM outperformed the other two in terms of relative bias and RMSE. However, under general MAR, the DWLS yielded biased means and regression coefficients of the latent slope.

FIML-MH-RM was sensitive to the biased DWLS estimates as starting values. Additionally, it was found that post-convergence approximated standard errors could not be produced in the conditional model in the current implementation of FIML-MH-RM in flexMIRT. Under MAR attrition, FIML-MCEM outperformed the other two in terms of parameter recovery.

The fifth goal of the research was to provide an empirical illustration of applying the LGM-IRT model to real-world data. The language and literacy assessment data in the "Multistate Study of Pre-Kindergarten 2001–2003" (Clifford et al., 2005) were used. A "summer effect" model (Raudenbush & Bryk, 2002) was fitted to the data using the three estimation methods. The estimates of the three estimates diverged considerably, leading to different inferences of students' growth and summer effect. In general, the pattern of the item parameter estimates confirmed with the simulation results. It is difficult to judge which estimation method should be more trusted without true parameters. Based on the results of Simulation III, FIML-MCEM yielded acceptable parameter recovery across all conditions under panel attrition. Additionally the reduced model, which was found to produce robust structural parameter estimates in Simulation II, agreed more with the FIML-MCEM structural estimates. Based on these observations, stronger argument can be made that the FIML-MCEM estimates were less biased.

Overall, the results provided implications for applied researchers in choosing the more appropriate models and estimation methods according to their research purposes, because each method exhibited its own strengths and weaknesses. If the dataset at hand has little missing, DWLS and FIML-MH-RM can be good choices

to yield fast results. However, if missing data is a concern of the study, caution is advised when applying DWLS estimation. Unless explicit assumption of MAR-X can be made about the missing mechanism, or some measures (i.e., multiple imputation or weight adjustment) have been taken to address the missing, DWLS is not recommended to estimate the LGM-IRT. If missing data is a concern of the study, FIML-MCEM can be utilized. As shown in simulation II, FIML-MCEM was able to yield generally unbiased estimates under either MAR or MAR-X attrition. However, the large amount of time FIML-MCEM requires to estimate a model is a serious limitation of the method. For example, the lengthy estimation of FIML-MCEM can be extremely cumbersome when applied researchers need to fit similar models multiple times to find the best-fitting model. FIML-MHRM could also be a more efficient alternative to FIML-MCEM. However, caution is advised regarding the interpretation of the standard errors, as both standard error estimation methods of FIML-MH-RM yielded underestimated standard errors. If the interest of the researcher lies only in the structural parameters, the misspecified model by omitting common item effects could be utilized as it was able to yield almost unbiased structural parameters. However, if item calibration is a focus of the study, the model may not be suitable due to the biased item parameter estimates. For example, if a researcher wants to use the calibrated item parameters for other studies or use pre-calibrated items as anchor items in an LGM-IRT, the misspecified model is not recommended. The misspecified model by omitting time-specific disturbances is generally not recommended due to the biased structural parameter estimates, which are usually one of the key focuses of researchers. As shown in Simulation II, the model was shown to

produce slightly biased item parameter estimates. This characteristic of the model may be useful for obtaining starting values for FIML estimation algorithms.

5.2 Limitations

There are several limitations associated with the current research. First, the research used simple linear growth curve, fixed number of time points, and uncorrelated disturbances for the structural model. Extensions to models with more complex structural setup are not investigated. If a researcher would like to use a quadratic curve, more time points, or correlated disturbances, additional structural factors need to be included and the same Schmid-Leiman transformation can be applied. While such extensions seem straightforward, the performance of the three estimation methods is unclear due to the increase in dimensionality of the model. The maximum number of dimensions examined in this study was 14 with four time points and eight items, which was already a highly complex model. There might be a diminishing return on investment on concurrently estimating item and structural parameters using the three estimation methods. One potential solution is to separate the estimation of item parameters and estimation of structural parameters. However, the multiple-stage approach would impose another issue in estimating appropriate standard errors and not necessarily avoid the computational challenge. Another solution is to use dimension reduction techniques to analytically reduce the model to four dimension. The idea is further discussed in "Future Studies" section of the chapter.

If more elaborate structural models are to be used, the Schmid-Leiman transformation becomes complex. Take the autoregressive latent trajectory (ALT; Bollen & Curran, 2004) for an example, a latent variable from previous test can determine the current value of the same construct. The directional arrow from one time-specific latent variable to the next makes the proportional constraints more difficult to implement. Whether an ALT model can be used as the structural model was not examined.

Second, the study only considered structural parameters of the LGM-IRT. The latent scores for individual examinees were not examined. The estimation of individual scores is usually conducted in two stages, namely item calibration and scoring. The bias incurred in the calibration stage can be carried over to the subsequent scoring stage and causes biased latent score estimates and misleading inference (Cheng & Yuan, 2010; Liu & Yang, 2017; Patton et al., 2013, 2014; Thissen & Wainer, 1990; Yang et al., 2012). It is conceivable that more accurate item parameter estimates (such as those produced by FIML-MH-RM in complete data or MAR-X attrition condition) would result in less biased latent score estimates. Additionally, if the standard errors are used in the second stage to adjust for the carried-over sampling errors (e.g., Thissen & Wainer, 1990; Yang et al., 2012), the large standard errors for the item parameters produced by DWLS would yield wider confidence intervals for latent score estimates, which would be less efficient. However, without a simulation study, it is not yet clear how the three estimation methods would compare with each other when the two-stage IRT scoring is conducted.

Third, the study used 15,000 iterations to estimate post-convergence standard

errors when FIML-MH-RM algorithm was used. It is not clear whether further increasing the number of iterations could improve the convergence rate and/or the adequacy of the standard errors.

Fourth, this research did not consider methods to handle MAR other than FIML estimation. Approaches such as multiple imputations and variance weight adjustment techniques were not implemented when the DWLS estimator was used. The potential improvement of the DWLS estimator combined with these approaches was not examined.

Fifth, the two reduced models were examined under complete data. The performance of the two misspecified models was not examined with panel attrition. Thus, the conclusions about the reduced models are limited to complete-data situations. Furthermore, the performance of available model fit indices was not examined because each estimation method yielded not necessarily comparable model fit indices.

Sixth, the LGM-IRT considered in this research was highly constrained in that all the item loadings of the common item effects were set to be equal to those of the main factors. In a more general parameterization, the variances of the specific dimensions can be fixed and the loadings of the specific dimensions can all be freely estimated. The performance of the estimation methods with this more general model was not examined in this the current study. It is possible that different conclusions regarding the three estimation methods could be made when the parameterization of the LGM-IRT model changes.

5.3 Future Studies

Based on the findings in this research, several future research areas are identified. First, standard error estimation methods in FIML-MH-RM require further exploration. Based on the results of this research, the recursively approximated and the post-convergence approximated standard errors were both underestimated. Additionally, the post-convergence approximated standard errors could not be produced in a conditional model with covariates. It is not clear whether this was due to missing data, convergence issues or the current implementation of the algorithm in flexMIRT. For example, flexMIRT recently started incorporating covariates and implementing post-convergence approximation of standard errors. It is possible that the post-convergence approximation method might have not been modified for the conditional model with covariates. More efforts should be devoted to exploring standard error estimation using FIML-MH-RM algorithm.

Second, the issue of starting values in FIML-MH-RM should be further examined. This research indicates that the algorithm could be sensitive to biased DWLS estimates as starting values in a high-dimensional model under general MAR. Other starting values can be explored. One possibility is to use estimates from the reduced models as starting values for FIML-MH-RM. Simulation II results showed that the structural parameter estimates were robust in the reduced model without common item effects, while the item parameter estimates were somewhat robust in the reduced model without disturbances (if the model could converge at all). However, before doing this, simulations need to be conducted to see if the two reduced models

can still produce robust estimates under MAR panel attrition.

Third, methods to reduce the bias of DWLS estimates under MAR attrition can be explored. It was found in the study that DWLS estimates of structural parameters were severely biased when the attrition mechanism was general MAR. One potential remedy for such bias is to multiply impute the missing data based on observed outcomes and auxiliary variables. It is currently not clear how much bias could be reduced by using multiple imputation or what imputation methods perform best in reducing bias. Efforts can be devoted to examining the performance of different multiple imputation methods combined with limited-information estimation. Candidate imputation methods include polytomous regression, predictive mean matching, classification and regression tree, and so on.

Fourth, this research used a linear function as a first step in the investigation of the LGM-IRT. Schmid-Leiman transformation and estimation of LGM-IRT with other structural models (e.g, ALT model) can be explored. The performance of the estimators in a LGM-IRT with more time points and other structural formulations can be examined.

Last but not least, as discussed in Chapter 2, the full LGM-IRT model with common item effect could be analytically reduced to a four-dimensional problem regardless of the number of repeated items and time points. Even though four-dimensional integration with FIML-MH-RM could be time consuming, it might still be more efficient than FIML-MCEM, especially when there are a lot of time points and repeated items. The dimension reduction method of the LGM-IRT (formulated as a first-order model) is currently not supported in flexMIRT. Efforts could be

devoted to developing the codes to implement the method.

Appendix A: Relative Bias in Item Parameter Estimates of the Three Estimation Methods in Simulation I

Table A.1

Percent Relative Bias in Item Parameter Estimates by the Three Estimators under Complete Data in Simulation I

		4 items, 200 examinees			**4 items, 500 examinees**			**4 items, 2,000 examinees**			**8 items, 200 examinees**			**8 items, 500 examinees**			**8 items, 2,000 examinees**		
	True value	DWLS	MCEM	MH-RM	DWLS	MCEM	MH-RM	DWLS	MCEM	MH-RM	DWLS	MCEM	MH-RM	DWLS	MCEM	MH-RM	DWLS	MCEM	MH-RM
α_1	1.18	0.5	2.5	-1.8	0.1	1.8	-0.9	0.6	1.9	-0.6	-0.2	2.0	-0.8	0.0	2.8	0.0	-0.6	1.8	-0.2
α_2	1.29	-1.5	-8.4	5.8	-1.2	-7.7	1.9	-1.6	-6.2	2.4	-1.1	-6.7	1.7	-0.8	-9.0	0.2	1.6	-5.9	0.6
α_3	2.17	1.1	4.6	-4.9	3.5	3.8	-2.4	4.5	4.2	-1.6	-3.9	2.9	-3.0	-3.1	5.7	-0.3	-5.0	3.9	-0.3
α_4	2.57	-1.0	-7.0	1.5	-2.4	-4.5	0.5	-6.0	-3.3	2.2	-0.2	-3.4	0.9	4.3	-4.7	0.1	5.3	-4.0	-0.3
α_5	1.64										-0.3	2.7	-1.4	-1.2	4.3	-0.1	-1.9	2.8	-0.3
α_6	1.97										-1.6	6.5	-3.2	-4.4	7.9	-1.8	-4.7	6.9	-0.5
α_7	2.41										7.1	-0.9	-7.7	2.2	11.1	0.5	-14.3	8.7	-1.0
α_8	1.47										-2.8	-12.1	0.9	2.0	-9.6	1.4	3.2	-7.6	0.4
λ_{11}	1.92	6.5	2.0	4.4	7.6	-0.4	3.4	16.1	-5.9	4.3	0.0	-0.6	0.7	-0.1	-0.7	0.2	-0.6	-0.6	0.0
λ_{12}	1.13	2.1	3.1	1.3	1.5	2.4	2.5	8.6	-4.3	4.1	0.2	-0.7	1.2	0.2	-0.9	0.4	-0.8	-1.0	0.0
λ_{13}	2.45	1.4	1.5	-1.2	0.0	1.2	-0.3	1.6	-1.2	1.4	0.5	-0.1	1.4	0.5	-0.5	0.2	-0.4	-0.8	-0.1
λ_{14}	1.72	2.0	1.0	-2.9	1.3	2.5	-0.6	0.7	0.0	1.8	1.1	1.4	2.2	1.0	0.5	-0.1	0.2	-0.5	-0.3
λ_{21}	1.92	2.8	-1.3	-1.5	0.7	1.7	1.4	3.3	-1.6	2.5	2.8	-7.3	14.2	-0.5	-13.1	2.2	-8.7	-11.9	-2.3
λ_{22}	0.62	3.7	-2.5	-3.0	0.3	2.0	1.4	2.9	-1.5	2.4	1.2	-8.1	12.0	0.3	-11.1	1.9	-7.4	-10.8	-2.3
λ_{23}	5.17	0.9	-0.1	-1.1	0.7	1.3	0.4	0.1	-0.1	0.9	0.1	0.1	3.1	1.3	-0.4	-0.4	-0.3	-1.9	-1.3
λ_{24}	1.80	0.1	0.3	-0.2	0.4	0.6	0.0	0.3	0.1	0.4	0.5	3.0	3.2	2.2	2.3	-0.8	1.2	-0.4	-1.4
λ_{31}	1.20	5.9	-2.6	-1.5	6.4	-2.0	0.3	8.0	-2.2	1.7	-8.4	-15.6	9.7	-13.3	-15.0	-0.7	-20.0	-10.7	-0.9
λ_{32}	0.99	-8.4	1.5	5.1	0.0	-6.8	-3.1	-6.8	1.7	-6.8	-1.1	-2.8	8.6	-0.8	-3.9	-1.4	-4.6	-4.2	-1.4
λ_{33}	0.56	0.3	1.2	-1.7	4.9	4.3	0.7	3.1	0.8	1.5	4.6	6.7	7.8	5.5	2.5	-4.0	6.3	0.2	-1.5
λ_{34}	0.18	-2.3	-6.5	4.3	-16.0	-10.3	-0.5	-14.2	-3.4	-1.3	13.0	17.9	6.7	17.9	14.2	-3.3	20.9	6.2	-1.7
λ_{41}	1.18	7.1	0.1	4.1	4.9	0.3	6.7	12.9	-4.6	4.8	0.0	-5.0	4.1	-3.7	-5.3	-0.5	-5.2	-3.0	0.4
λ_{42}	-0.51	-0.3	-1.1	0.3	-0.5	-1.0	-0.6	-0.2	0.0	-1.0	0.4	-0.1	1.5	0.6	0.0	-0.2	-0.3	-0.5	-0.2
λ_{43}	3.44	8.0	16.2	-37.3	32.2	28.2	-11.9	43.3	17.3	10.8	1.1	4.3	1.2	4.5	4.2	-0.6	3.9	1.4	-0.7
λ_{44}	1.23	-1.0	-0.6	3.3	-2.9	-2.1	1.4	-4.5	-1.8	-0.4	2.5	13.3	-0.8	13.1	12.6	-2.0	13.7	5.5	-1.9
λ_{51}	0.55										0.5	-3.7	4.0	-2.8	-4.4	0.3	-3.9	-3.0	0.1
λ_{52}	-0.05										-3.4	8.2	-15.2	4.7	11.6	-1.2	9.1	9.0	-0.8
λ_{53}	-0.80										1.3	-1.7	4.3	-0.9	-2.7	0.1	-2.1	-2.3	0.1
λ_{54}	-1.85										-2.3	-3.0	-5.4	-3.8	-2.1	-0.5	-1.8	0.4	-0.3
λ_{61}	0.93										4.0	2.7	9.6	-2.3	-5.5	-3.2	-4.3	-4.8	-1.6
λ_{62}	-0.68										-0.9	-1.9	-1.7	-0.6	-0.3	0.7	-0.3	0.3	0.4
λ_{63}	2.21										10.3	25.8	19.3	11.0	6.9	-9.7	7.2	-2.7	-5.2
λ_{64}	1.07										-0.5	-1.4	-0.6	-0.9	-0.9	0.3	-0.8	-0.2	0.2
λ_{71}	-0.28										5.2	-24.4	3.5	-4.9	-21.3	0.1	-18.3	-13.0	-0.2
λ_{72}	-0.67										-20.6	48.8	-14.6	5.1	46.2	-1.7	41.6	32.4	2.6
λ_{73}	-2.53										-11.3	21.8	-9.8	0.5	20.9	0.6	18.7	16.2	2.3
λ_{74}	-4.80										-3.0	4.7	-2.9	-0.2	5.0	0.6	4.1	3.9	0.6
λ_{81}	-0.35										0.3	-0.9	1.5	-0.9	-1.9	-0.4	-1.2	-1.2	-0.1
λ_{82}	-1.56										0.9	-1.3	3.4	-1.3	-3.3	-1.0	-1.8	-2.1	-0.4
λ_{83}	1.48										1.1	-1.3	3.9	-1.1	-3.3	-0.9	-1.9	-2.4	-0.5
λ_{84}	0.33										3.1	-2.3	11.5	-1.6	-6.5	-4.8	-2.3	-4.9	-1.5

Appendix B: Relative Bias in Item Parameter Estimates of the Three Estimation Methods in Simulation III

Table B.1

Percent Relative Bias in Item Parameter Estimates by the Three Estimators under 10% per Wave MAR-X Attrition in Simulation III

		4 items, 200 examinees			**4 items, 500 examinees**			**4 items, 2,000 examinees**			**8 items, 200 examinees**			**8 items, 500 examinees**			**8 items, 2,000 examinees**		
	True value	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM
α_1	1.18	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
α_2	1.29	0.4	22.9	-13.5	2.7	18.3	-11.1	5.3	19.4	-2.5	10.4	37.0	1.3	7.5	41.3	4.6	3.1	32.4	5.3
α_3	2.17	-0.7	-1.0	1.2	-0.6	-1.2	0.4	-1.1	-0.9	0.4	0.5	-2.1	-0.2	0.8	-2.1	-0.3	1.1	-1.8	-0.3
α_4	2.57	8.1	5.2	-5.2	8.0	5.6	-4.8	8.4	6.1	-1.7	-0.9	8.8	2.1	-7.4	7.2	0.7	-5.5	7.0	1.6
α_5	1.64										-1.1	-22.7	-2.9	3.8	-22.4	-0.3	3.0	-20.5	-3.8
α_6	1.97										-4.4	-19.1	-2.9	2.9	-18.0	-1.2	5.4	-14.1	-1.4
α_7	2.41										-1.6	-5.5	-2.3	0.0	-5.3	-2.3	3.0	-4.4	-1.2
α_8	1.47										0.4	2.2	0.2	-0.1	2.0	0.1	-0.1	1.8	0.3
λ_{11}	1.92	0.9	-3.6	9.2	5.4	-4.9	8.5	7.4	-5.4	4.9	-0.9	-3.8	1.0	-1.2	-5.0	1.0	-1.9	-3.8	0.2
λ_{12}	1.13	7.9	-22.2	48.8	25.0	-37.6	45.1	18.5	-25.2	42.6	0.0	-3.0	0.8	-0.6	-4.2	0.5	-0.9	-2.9	0.4
λ_{13}	2.45	1.1	-0.6	1.7	0.0	-1.2	2.6	0.5	-0.6	3.0	-1.0	-7.6	-0.7	0.4	-7.5	1.7	-0.5	-5.1	1.3
λ_{14}	1.72	0.5	0.8	-1.6	2.1	1.6	1.8	2.3	1.3	2.5	0.3	0.4	-0.5	0.8	1.0	0.3	0.6	0.7	0.2
λ_{21}	1.92	0.6	0.4	2.8	1.5	-0.7	3.1	2.0	-0.6	3.1	1.3	-5.3	5.2	-2.3	-9.9	0.5	-3.4	-5.7	0.6
λ_{22}	0.62	0.5	0.2	2.4	1.0	-0.6	3.2	1.7	-0.7	2.8	10.8	-46.3	41.0	-14.4	-79.6	5.8	-23.7	-44.9	6.2
λ_{23}	5.17	1.0	0.5	0.9	0.4	-0.2	0.9	0.2	-0.4	0.8	2.1	-0.3	2.3	1.1	-1.4	0.5	0.4	-1.3	-0.6
λ_{24}	1.80	0.6	0.3	0.2	0.4	-0.1	0.4	0.2	-0.3	0.3	2.3	1.4	1.7	1.6	1.0	0.0	1.8	0.3	-0.8
λ_{31}	1.20	4.8	-2.0	-0.1	5.8	-1.2	2.1	7.5	-1.1	1.4	-12.2	-12.3	2.5	-15.8	-17.3	1.9	-18.6	-11.4	-0.4
λ_{32}	0.99	-4.7	4.0	1.9	-5.2	3.4	-5.3	-7.2	3.9	-4.7	-2.0	-5.4	0.2	-1.7	-4.8	2.1	-3.2	-3.8	0.1
λ_{33}	0.56	2.3	0.4	-0.2	2.0	-1.0	0.3	3.4	-0.2	1.7	5.5	0.5	-0.4	7.5	3.5	1.6	6.4	0.6	-0.7
λ_{34}	0.18	-10.4	-3.6	-0.1	-10.7	0.8	1.1	-15.1	-0.9	-2.8	14.3	6.8	-2.1	17.4	11.3	-1.4	20.3	7.3	-0.5
λ_{41}	1.18	8.3	-5.1	2.9	8.9	-1.9	4.9	13.9	-4.6	2.5	-3.0	-4.6	0.4	-5.8	-6.7	-0.1	-4.5	-3.3	0.7
λ_{42}	-0.51	-0.2	0.1	-0.3	0.0	0.3	-0.6	-0.6	0.7	-0.5	-0.4	-1.2	-0.5	-0.4	-0.9	-0.1	0.1	-0.3	0.3
λ_{43}	3.44	40.6	19.3	-0.7	42.1	6.8	6.2	40.4	-1.3	7.9	2.0	1.2	-1.2	3.7	2.6	-0.4	4.3	2.1	0.3
λ_{44}	1.23	-3.7	-1.7	0.4	-3.9	-0.7	0.0	-4.0	-0.1	-0.1	7.9	6.4	-3.1	14.1	11.3	-1.0	13.9	7.1	-0.2
λ_{51}	0.55										-2.8	-5.3	0.5	-3.0	-5.1	1.0	-3.9	-3.5	0.0
λ_{52}	-0.05										4.6	14.7	-3.1	6.1	14.4	-4.1	9.2	10.8	0.0
λ_{53}	-0.80										-1.1	-3.8	0.5	-1.2	-3.4	1.0	-1.9	-2.6	0.1
λ_{54}	-1.85										-1.3	1.1	-0.2	-2.5	-0.5	-1.7	-1.0	1.3	0.4
λ_{61}	0.93										-3.3	-6.6	-2.8	-1.8	-4.6	1.3	-1.6	-2.5	1.0
λ_{62}	-0.68										0.2	0.6	0.8	-0.6	-0.3	-0.1	-0.8	-0.2	-0.1
λ_{63}	2.21										-1.2	-6.3	-12.5	11.1	6.6	1.8	13.9	4.0	1.5
λ_{64}	1.07										-0.3	-0.2	0.4	-0.8	-0.7	0.1	-1.0	-0.4	0.0
λ_{71}	-0.28										-5.0	-14.6	4.8	-9.5	-19.8	3.7	-16.7	-13.7	0.6
λ_{72}	-0.67										9.2	36.9	-8.5	20.0	48.1	-7.6	34.0	31.2	-2.7
λ_{73}	-2.53										4.6	19.7	-1.9	7.9	22.9	-3.4	14.1	14.9	-1.2
λ_{74}	-4.80										0.5	4.6	-0.5	1.8	5.7	-0.5	3.0	3.8	-0.1
λ_{81}	-0.35										-0.5	-1.5	0.0	-0.8	-1.9	0.1	-1.0	-1.1	0.1
λ_{82}	-1.56										-1.0	-3.0	-0.5	-1.5	-3.6	-0.1	-1.3	-1.8	0.2
λ_{83}	1.48										-0.9	-3.2	-0.5	-1.5	-3.9	-0.1	-1.4	-2.0	0.1
λ_{84}	0.33										-0.5	-7.0	-0.9	-2.2	-8.7	-0.9	-1.0	-4.2	0.1

Table B.2

Percent Relative Bias in Item Parameter Estimates by the Three Estimation Methods under 20% per Wave MAR-X Attrition in Simulation III

		4 items, 200 examinees			**4 items, 500 examinees**			**4 items, 2,000 examinees**			**8 items, 200 examinees**			**8 items, 500 examinees**			**8 items, 2,000 examinees**		
	True value	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM
α_1	1.18	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
α_2	1.29	6.5	25.4	-14.7	9.9	20.6	-6.7	4.1	15.8	-4.1	8.8	33.7	1.7	13.9	45.6	7.2	4.1	32.3	2.4
α_3	2.17	-0.5	-1.3	1.1	-0.6	-1.0	0.3	-0.8	-0.8	0.2	0.0	-2.1	-0.3	0.5	-2.5	-0.3	0.9	-1.8	-0.2
α_4	2.57	7.7	5.0	-8.7	3.9	5.7	-2.3	6.5	3.7	-1.6	-1.0	7.6	1.8	-3.3	10.6	2.0	-6.7	6.5	0.8
α_5	1.64										-7.1	-22.7	-4.9	0.5	-26.2	-2.9	3.9	-19.7	-1.8
α_6	1.97										-2.5	-18.0	-5.1	-2.7	-22.3	-3.2	3.1	-16.2	-2.1
α_7	2.41										-2.2	-3.6	-1.3	1.8	-5.5	-0.9	5.3	-3.8	-0.8
α_8	1.47										0.7	1.8	0.1	0.1	2.3	0.2	-0.2	1.8	0.1
λ_{11}	1.92	2.5	-5.1	9.6	6.9	-7.5	6.8	9.4	-8.5	2.7	-1.1	-4.1	0.8	-1.9	-5.3	0.3	-2.2	-4.9	0.0
λ_{12}	1.13	7.1	-37.1	39.2	6.0	-31.3	58.0	45.7	-59.1	13.4	0.2	-2.8	1.3	-0.8	-4.2	0.3	-1.3	-4.0	-0.1
λ_{13}	2.45	0.9	-2.3	-0.2	1.0	-1.0	3.0	1.6	-2.9	0.7	2.1	-5.2	3.6	0.5	-7.3	1.5	-2.2	-8.1	-0.8
λ_{14}	1.72	0.1	0.7	-2.3	1.8	0.8	0.9	0.5	-0.3	0.6	1.9	1.4	1.4	1.4	1.3	0.6	0.0	0.2	-0.5
λ_{21}	1.92	1.8	-2.3	0.5	2.3	-2.3	2.0	2.9	-2.4	2.0	-2.9	-8.9	1.3	-0.9	-8.0	2.0	-2.9	-7.5	0.0
λ_{22}	0.62	1.9	-2.7	-0.2	1.9	-2.3	1.9	2.6	-2.5	1.8	-21.7	-74.3	9.2	-4.0	-66.2	16.2	-22.5	-63.4	-2.9
λ_{23}	5.17	0.5	-0.1	0.3	0.4	-0.2	0.8	0.2	-0.4	0.8	0.4	-2.1	0.6	1.7	-1.3	1.0	0.3	-1.8	-0.8
λ_{24}	1.80	0.2	-0.2	-0.1	0.3	-0.1	0.2	0.3	0.0	0.4	0.7	-0.6	0.2	1.4	-0.1	-0.3	1.3	0.2	-1.0
λ_{31}	1.20	5.9	-3.2	-0.4	6.5	-2.6	1.4	7.8	-3.3	0.7	-9.2	-10.2	4.5	-15.0	-15.6	1.0	-18.0	-15.7	-1.8
λ_{32}	0.99	-6.6	7.0	2.0	-5.7	4.5	-4.3	-7.5	6.0	-4.1	-0.2	-3.2	2.9	-2.2	-6.4	0.8	-3.5	-5.9	-1.0
λ_{33}	0.56	1.9	-0.6	-0.6	3.3	0.6	1.5	2.3	-0.7	0.6	4.0	0.3	0.3	4.2	-2.0	-1.7	4.9	-0.4	-1.8
λ_{34}	0.18	-8.8	-0.8	2.1	-13.3	-3.2	-1.1	-12.7	-1.4	-0.4	10.8	5.0	-1.8	13.6	4.1	-4.3	17.6	8.0	-1.6
λ_{41}	1.18	8.3	-6.3	3.3	8.9	-4.8	4.2	10.7	-5.4	4.5	-1.9	-3.4	2.3	-4.7	-5.7	0.5	-5.8	-5.4	-0.5
λ_{42}	-0.51	-0.1	0.5	-0.4	0.0	0.3	-0.7	-0.1	0.3	-0.9	0.3	-0.3	0.7	0.0	-0.8	0.4	-0.3	-0.9	-0.3
λ_{43}	3.44	34.0	12.0	4.7	32.2	8.6	2.4	40.4	11.8	11.3	2.2	1.6	-0.4	3.9	2.1	0.3	4.0	2.0	-0.3
λ_{44}	1.23	-2.8	-1.1	0.5	-3.4	-1.2	0.1	-4.0	-1.4	-0.4	7.9	6.5	-2.1	12.3	8.2	-0.4	13.6	7.9	-1.3
λ_{51}	0.55										-1.4	-3.6	1.9	-3.2	-4.6	1.1	-3.7	-4.3	0.1
λ_{52}	-0.05										4.4	13.5	-3.4	9.1	16.1	-1.8	9.0	13.6	0.0
λ_{53}	-0.80										-0.9	-3.3	0.9	-1.8	-3.8	0.6	-1.9	-3.3	0.0
λ_{54}	-1.85										-0.8	1.1	-0.5	-1.4	1.2	-0.5	-1.4	1.1	0.3
λ_{61}	0.93										-1.2	-4.0	1.2	-3.0	-6.8	-0.2	-2.7	-4.8	0.2
λ_{62}	-0.68										-0.4	0.2	0.0	-0.2	0.4	0.1	-0.6	0.0	-0.1
λ_{63}	2.21										5.6	-2.1	-3.1	5.3	-2.9	-1.3	10.4	1.7	1.1
λ_{64}	1.07										-0.6	-0.3	0.2	-0.5	-0.2	0.2	-0.8	-0.4	0.0
λ_{71}	-0.28										-5.3	-16.7	2.7	-14.0	-19.0	1.3	-20.0	-16.7	1.0
λ_{72}	-0.67										8.2	38.6	-8.1	26.6	44.3	-4.0	42.4	39.9	-2.6
λ_{73}	-2.53										2.3	18.3	-4.0	12.0	23.0	-0.3	17.3	18.5	-1.6
λ_{74}	-4.80										0.1	4.2	-1.2	2.6	5.8	0.0	3.5	4.3	-0.5
λ_{81}	-0.35										-0.2	-1.2	0.3	-0.6	-1.4	0.4	-1.0	-1.4	0.0
λ_{82}	-1.56										-0.1	-2.0	0.7	-0.6	-2.5	0.8	-1.3	-2.4	0.1
λ_{83}	1.48										-0.1	-2.3	0.7	-0.5	-2.8	0.9	-1.4	-2.7	0.1
λ_{84}	0.33										0.9	-4.3	1.5	1.4	-5.6	2.7	-1.5	-6.2	-0.8

Table B.3

Percent Relative Bias in Item Parameter Estimates by the Three Estimation Methods under 10% per Wave MAR Attrition in Simulation III

		4 items, 200 examinees			**4 items, 500 examinees**			**4 items, 2,000 examinees**			**8 items, 200 examinees**			**8 items, 500 examinees**			**8 items, 2,000 examinees**		
	True value	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM
α_1	1.18	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
α_2	1.29	19.4	31.2	-6.2	20.2	17.0	-8.9	19.1	11.4	-1.7	20.6	28.7	-9.5	16.1	24.9	-9.2	16.8	21.8	-1.7
α_3	2.17	-3.3	-1.8	0.5	-2.7	-1.0	0.2	-2.5	-0.7	-0.2	-2.2	-1.3	0.8	-2.9	-1.3	0.5	-2.8	-1.1	0.1
α_4	2.57	53.3	13.3	2.4	43.9	5.1	-0.2	39.7	0.8	1.2	33.2	5.5	-2.4	31.6	4.1	-2.1	27.6	3.6	-0.3
α_5	1.64										-0.2	-15.6	7.1	2.3	-15.0	6.5	0.7	-13.7	0.3
α_6	1.97										-28.8	-10.7	5.5	-24.2	-12.2	3.9	-20.8	-9.3	1.1
α_7	2.41										11.6	-2.5	0.9	13.9	-2.0	1.3	13.8	-2.1	-0.4
α_8	1.47										-0.9	1.5	-0.6	-1.3	1.4	-0.4	-1.2	1.1	-0.1
λ_{11}	1.92	15.6	-7.1	1.9	15.5	-4.8	4.9	18.7	-6.2	0.4	-3.1	-3.4	0.3	-3.1	-3.4	0.5	-3.5	-2.9	0.0
λ_{12}	1.13	95.8	-42.6	-6.6	81.3	-22.1	30.6	104.3	-35.1	3.8	-2.5	-2.5	-0.2	-2.5	-2.4	0.1	-2.8	-2.1	-0.2
λ_{13}	2.45	5.2	-1.4	-2.0	2.6	0.9	2.4	5.1	-1.4	0.2	-7.1	-3.8	-3.1	-5.4	-1.9	-0.4	-7.0	-2.8	-1.9
λ_{14}	1.72	2.3	1.3	-3.5	0.5	3.1	1.0	1.3	0.8	0.3	-2.5	1.3	-2.7	-1.4	2.3	-1.2	-1.6	1.5	-1.1
λ_{21}	1.92	5.7	1.0	0.9	7.9	-1.4	0.8	9.0	-2.2	0.7	-7.2	-4.6	-0.9	-9.3	-6.1	-2.4	-9.3	-4.7	-3.1
λ_{22}	0.62	5.7	0.7	0.4	7.1	-0.9	1.2	8.4	-1.9	0.9	-71.6	-42.2	-17.6	-81.0	-48.7	-23.1	-81.3	-37.2	-28.2
λ_{23}	5.17	0.8	1.3	0.5	1.6	0.2	0.5	2.1	-0.2	0.3	-6.7	-0.5	-2.8	-4.7	1.3	-1.3	-5.3	0.4	-2.3
λ_{24}	1.80	0.3	0.6	0.0	0.4	0.4	0.4	0.8	0.0	0.1	-6.5	2.0	-2.7	-3.9	4.3	-1.2	-4.9	2.2	-2.7
λ_{31}	1.20	1.3	-1.0	-0.1	3.4	-2.0	0.7	4.4	-2.3	0.8	-1.2	-12.5	-4.3	0.7	-13.0	-2.1	-2.7	-12.4	-5.0
λ_{32}	0.99	-15.0	-2.5	0.1	-19.5	2.6	-1.1	-22.2	4.7	-0.2	-10.5	-2.6	-4.8	-11.9	-3.7	-3.8	-11.7	-3.6	-4.8
λ_{33}	0.56	7.6	0.2	-2.0	5.6	1.2	0.4	6.7	-0.2	-0.5	-17.9	3.1	-6.4	-19.3	4.0	-4.2	-18.2	2.0	-5.3
λ_{34}	0.18	-14.4	-2.0	4.1	-8.1	-4.4	0.0	-8.0	-3.2	0.4	-23.2	12.4	-6.7	-25.3	13.7	-4.7	-23.0	10.1	-5.2
λ_{41}	1.18	13.0	1.2	6.0	1.8	-5.4	1.0	0.8	-6.4	1.7	6.0	-3.9	-1.8	5.1	-4.2	-1.0	3.0	-3.9	-1.7
λ_{42}	-0.51	-3.1	-0.3	0.1	-3.7	0.4	0.3	-3.6	0.3	-0.1	-3.6	-0.4	-1.3	-3.3	0.0	-0.7	-3.3	-0.1	-0.9
λ_{43}	3.44	138.7	16.4	-16.9	122.2	11.7	-10.8	110.1	16.0	-2.4	-15.7	2.3	-2.5	-13.9	3.5	-1.3	-12.1	2.8	-1.3
λ_{44}	1.23	-10.1	-1.0	1.8	-8.0	-1.3	0.9	-7.0	-1.6	0.5	-39.7	8.7	-4.5	-34.1	12.2	-2.0	-29.2	9.4	-2.2
λ_{51}	0.55										-4.7	-3.7	-0.6	-5.7	-3.9	-0.4	-6.5	-3.5	-1.2
λ_{52}	-0.05										16.1	9.0	2.9	20.6	11.9	3.7	22.0	10.3	5.9
λ_{53}	-0.80										-3.9	-1.7	-0.6	-5.1	-2.5	-0.8	-5.6	-2.4	-1.6
λ_{54}	-1.85										4.6	-0.3	2.5	3.3	-1.5	1.1	4.0	-0.2	2.4
λ_{61}	0.93										-7.9	-1.4	-2.3	-9.7	-2.7	-1.8	-10.7	-3.0	-3.2
λ_{62}	-0.68										3.5	-0.8	0.8	3.0	-0.7	0.4	2.9	-0.4	0.7
λ_{63}	2.21										-49.9	16.5	-8.3	-45.2	11.2	-7.4	-42.4	7.4	-9.3
λ_{64}	1.07										2.4	-1.2	0.3	1.9	-1.0	0.2	1.8	-0.7	0.3
λ_{71}	-0.28										-29.9	-15.3	2.3	-35.1	-17.0	1.2	-38.0	-14.6	-0.4
λ_{72}	-0.67										68.6	35.7	-2.1	83.2	41.4	2.1	87.1	33.5	2.1
λ_{73}	-2.53										32.2	16.4	1.1	39.1	19.0	2.6	41.2	16.0	2.9
λ_{74}	-4.80										7.7	3.8	0.8	8.7	3.9	0.7	9.7	3.7	1.1
λ_{81}	-0.35										-1.8	-1.1	-0.3	-2.3	-1.3	-0.1	-2.4	-1.1	-0.4
λ_{82}	-1.56										-3.2	-1.7	-0.8	-4.0	-2.1	-0.5	-4.3	-2.0	-1.0
λ_{83}	1.48										-3.5	-1.8	-1.1	-4.5	-2.3	-0.8	-4.6	-2.1	-1.2
λ_{84}	0.33										-8.2	-2.0	-5.2	-8.5	-2.1	-2.1	-9.5	-2.8	-4.4

Table B.4

Percent Relative Bias in Item Parameter Estimates by the Three Estimation Methods under 20% per Wave MAR Attrition in Simulation III

		4 items, 200 examinees			**4 items, 500 examinees**			**4 items, 2,000 examinees**			**8 items, 200 examinees**			**8 items, 500 examinees**			**8 items, 2,000 examinees**		
	True value	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM	DWLS	MCEM	MHRM
α_1	1.18	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
α_2	1.29	18.8	29.5	-19.6	8.6	19.8	-20.4	11.5	24.0	-7.8	15.2	27.4	-23.2	5.1	20.0	-21.4	6.7	17.3	-16.3
α_3	2.17	-3.3	-1.9	1.3	-2.7	-1.3	1.0	-2.8	-1.4	0.5	-2.9	-1.4	1.6	-2.2	-0.7	1.7	-2.0	-0.8	1.1
α_4	2.57	58.9	13.9	-1.2	51.2	8.8	-3.4	47.9	7.8	-0.9	35.5	4.7	-7.5	33.9	3.6	-7.1	32.4	2.4	-4.2
α_5	1.64										9.0	-10.9	20.0	10.1	-10.5	14.7	10.5	-9.7	10.3
α_6	1.97										-14.1	-11.3	13.9	-14.0	-7.5	13.4	-12.2	-7.4	9.1
α_7	2.41										12.4	-2.3	4.5	17.1	0.0	4.9	15.9	-0.9	2.8
α_8	1.47										-1.3	1.1	-1.7	-1.7	0.8	-1.4	-1.9	0.7	-1.1
λ_{11}	1.92	24.4	-7.1	-3.9	24.4	-5.4	0.2	26.7	-7.1	-5.0	-3.9	-1.5	0.4	-4.6	-2.5	1.3	-4.6	-2.0	-0.4
λ_{12}	1.13	153.7	-37.9	-54.8	161.0	-32.4	-28.9	168.1	-36.2	-49.4	-3.5	-0.5	-0.6	-4.3	-1.7	0.5	-4.4	-1.3	-1.4
λ_{13}	2.45	8.3	0.0	-5.4	9.4	-0.6	-3.7	9.5	-0.7	-4.2	-13.2	0.6	-8.6	-13.1	-0.7	-0.8	-12.9	0.7	-6.9
λ_{14}	1.72	4.8	2.9	-7.3	5.5	2.6	-5.2	5.0	2.6	-4.2	-5.6	2.2	-6.1	-4.5	2.7	-2.0	-4.4	2.9	-4.1
λ_{21}	1.92	11.5	0.6	-3.8	12.4	-0.8	-3.3	13.2	-1.1	-3.6	-12.4	0.9	-4.7	-18.1	-5.8	-3.2	-15.7	-2.4	-8.4
λ_{22}	0.62	11.5	0.1	-4.3	11.7	-0.7	-3.3	12.1	-0.5	-3.2	-114.3	6.8	-49.6	-161.3	-48.4	-31.6	-138.9	-17.4	-78.5
λ_{23}	5.17	3.2	0.5	-1.4	3.3	-0.3	-1.6	3.0	0.1	-1.2	-9.7	2.6	-5.1	-11.4	0.0	-2.7	-9.8	1.9	-6.7
λ_{24}	1.80	1.4	0.2	-0.8	1.3	-0.1	-0.8	1.0	0.2	-0.5	-10.1	3.8	-5.9	-10.4	2.5	-2.5	-9.1	3.9	-7.1
λ_{31}	1.20	2.2	0.6	-1.3	5.3	-1.7	-2.2	4.9	-1.0	-1.5	3.1	-1.9	-5.4	-1.6	-6.8	1.3	-8.3	-9.0	-11.0
λ_{32}	0.99	-29.0	-0.2	12.9	-31.3	2.1	10.7	-31.8	1.6	10.3	-15.2	1.9	-9.2	-16.5	-0.2	-1.2	-16.6	0.1	-10.4
λ_{33}	0.56	10.9	1.0	-3.9	10.8	0.3	-3.6	11.3	-0.1	-3.8	-27.6	5.8	-10.8	-25.9	5.3	-3.0	-22.6	6.8	-10.5
λ_{34}	0.18	-21.4	-1.4	8.4	-19.5	-1.1	7.5	-18.1	-1.7	6.6	-41.3	10.6	-13.5	-33.1	14.7	-3.0	-28.5	14.9	-11.0
λ_{41}	1.18	8.9	-0.2	-1.2	3.5	-2.7	-2.9	2.8	-1.3	-0.7	7.0	-0.5	-3.5	5.0	-2.8	0.5	3.4	-2.8	-4.5
λ_{42}	-0.51	-4.7	-0.4	1.3	-5.0	-0.2	1.2	-4.8	-0.2	0.9	-4.3	1.3	-1.9	-5.8	-0.4	-0.6	-5.3	0.3	-2.5
λ_{43}	3.44	198.6	4.6	-49.8	181.0	9.7	-34.6	169.4	9.4	-31.5	-18.5	4.5	-2.6	-20.0	2.7	-1.3	-18.3	3.6	-3.7
λ_{44}	1.23	-13.8	-0.4	3.4	-11.8	-1.1	2.1	-10.8	-1.0	1.9	-49.7	9.5	-6.1	-48.4	9.3	-3.6	-44.6	10.6	-6.3
λ_{51}	0.55										-7.9	-2.4	-2.8	-7.4	-1.8	0.8	-8.6	-2.3	-3.4
λ_{52}	-0.05										28.1	6.0	12.5	26.5	4.7	-0.7	29.2	4.9	14.5
λ_{53}	-0.80										-7.4	-1.2	-3.7	-7.0	-0.9	0.0	-7.5	-0.8	-4.0
λ_{54}	-1.85										6.5	-2.7	5.1	6.7	-1.7	1.7	6.0	-2.7	6.0
λ_{61}	0.93										-13.1	1.0	-7.4	-15.8	-2.4	-1.8	-16.5	-1.6	-9.0
λ_{62}	-0.68										4.1	-0.7	2.1	4.3	-0.4	0.7	4.2	-0.6	2.1
λ_{63}	2.21										-59.5	11.0	-29.6	-61.5	9.0	-9.1	-59.8	10.4	-29.7
λ_{64}	1.07										2.3	-0.9	1.0	2.5	-0.8	0.4	2.4	-0.8	1.1
λ_{71}	-0.28										-36.2	-11.0	-1.3	-44.5	-14.9	-0.1	-44.0	-13.0	-3.4
λ_{72}	-0.67										82.7	20.4	5.0	106.8	38.1	6.0	106.4	32.3	16.4
λ_{73}	-2.53										40.7	8.1	5.9	49.8	15.3	2.2	51.2	14.1	11.9
λ_{74}	-4.80										10.0	1.1	2.2	12.3	3.5	1.2	12.6	3.1	4.0
λ_{81}	-0.35										-2.5	-0.5	-0.8	-2.8	-0.8	0.0	-3.3	-1.0	-1.2
λ_{82}	-1.56										-4.7	-0.4	-2.0	-5.0	-1.1	0.1	-5.7	-1.2	-2.6
λ_{83}	1.48										-5.4	-0.5	-2.8	-5.6	-1.0	0.1	-6.2	-1.1	-3.0
λ_{84}	0.33										-15.3	0.5	-11.5	-14.7	0.0	-1.4	-14.3	1.1	-10.6

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47–76.

Asparouhov, T., & Muthén, B. (2010). Weighted least squares estimation with missing data. *Mplus Technical Appendix*, 1–10.

Bashkov, B. M. (2015). *Examining the performance of the metropolis-hastings robbins-monro algorithm in the estimation of multilevel multidimensional IRT models*. Unpublished doctoral dissertation, Department of Graduate Psychology, James Madison University.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*(2), 135–167.

Bellman, R. (1957). Dynamic programming. *Princeton, USA: Princeton University Press*, *1*(2), 3.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (alt) models a synthesis of two traditions. *Sociological Methods & Research*, *32*(3), 336–383.

Cai, L. (2008). *A metropolis-hastings robbins-monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model*. Unpublished doctoral dissertation, The University of North Carolina at Chapel Hill.

Cai, L. (2010a). Metropolis-hastings robbins-monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Cai, L. (2010b). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612.

Cai, L. (2017). *Flexible multilevel multidimensional item analysis and test scoring [computer software]* (3.51 ed.). Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221.

Cheng, Y., & Yuan, K.-H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika*, *75*(2), 280–291.

Chou, C.-P., Bentler, P. M., & Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(3), 247–266.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*(1), 5–32.

Clifford, D., Bryant, D., Burchinal, M., & Barbarin, O. (2005). National center for early development and learning multi-state study of pre-kindergarten, 2001–2003.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*, *66*(3), 227–268.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Hold, Rinehart, and Winston.

Cudeck, R., & Harring, J. R. (n.d.). Estimating the correlation between two variables when individuals are measured repeatedly. In M. Edwards & R. MacCallum (Eds.), *Current topics in the theory and application of latent variable models.*

Cudeck, R., Harring, J. R., & du Toit, S. H. (2009). Marginal maximum likelihood estimation of a latent variable model with interaction. *Journal of Educational and Behavioral Statistics*, *34*(1), 131–144.

Curran, P. J., Edwards, M. C., Wirth, R., Hussong, A. M., & Chassin, L. (2007). The incorporation of categorical measurement models in the analysis of individual growth. In T. Little, J. Bovaird, & N. Card (Eds.), *Modeling ecological and contextual effects in longitudinal studies of human development* (pp. 89–120). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. *Developmental Psychology*, *44*(2), 365.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society*, 1–38.

Depaoli, S. (2013). Mixture class recovery in gmm under varying degrees of class separation: Frequentist versus bayesian estimation. *Psychological Methods*, *18*(2), 186.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, 49–93.

Duncan, S. C., & Duncan, T. E. (1996). A multivariate latent growth curve analysis of adolescent substance use. *Structural Equation Modeling: A Multidisci-*

plinary Journal, *3*(4), 323–347.

Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*(3), 474–497.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495–515.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*(1), 128–141.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*(1), 1.

Enders, C. K., & Tofighi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Equation Modeling*, *15*(1), 75–95.

Entwisle, D. R., & Alexander, K. L. (1992). Summer setback: Race, poverty, school composition, and mathematics achievement in the first two years of school. *American Sociological Review*, *57*, 72–84.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374.

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. de Gruijter L. van der Kamp (Ed.), (pp. 97–110). New York: John Wiley & Sons.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, *48*(1), 3–26.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika*, *54*(4), 599–624.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 131–155). New York: Springer.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical proceedings*

of the cambridge philosophical society (Vol. 22, pp. 700–725).

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*(3), 275.

Fox, J.-P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika*, *66*(2), 271–288.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436.

Graham, J. W. (2003). Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling*, *10*(1), 80–100.

Han, K. C. T., & Paek, I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement*, 486–498.

Hancock, G. R., & Buehl, M. M. (2008). Second-order latent growth models with shifting indicators. *Journal of Modern Applied Statistical Methods*, *7*(1), 39–55.

Hancock, G. R., Kuo, W., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling*, *8*(3), 470–489.

Hastings, W. K. (1970). Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. John Wiley & Sons.

Hill, C. D. (2006). *Two models for longitudinal item response data*. Unpublished doctoral dissertation, Department of Psychology, University of North Carolina at Chapel Hill.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, *2*(3), 261.

Jeon, M., & Rabe-Hesketh, S. (2015). An autoregressive growth model for longitudinal item analysis. *Psychometrika*, 1–21.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*(1), 82–100.

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

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*(1), 136–153.

Koran, J. (2009). *An integrated item response model for evaluating individual students' growth in educational achievement.* Unpublished doctoral dissertation, University of Maryland, College Park.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, *12*(3), 237–255.

Lee, U. (2003). *Panel attrition in survey data: a literature review* (Vol. 41). Centre for Social Science Research, University of Cape Town.

Li, M. (2015). *Investigating methods of incorporating covariates in growth mixing modeling: A simulation study.* Unpublished doctoral dissertation, Program of Measurement, Statistics and Evaluation, University of Maryland, College Park.

Liu, J. (2012). *A systematic investigation of within-subject and between-subject covariance structures in growth mixture models.* Unpublished doctoral dissertation, Program of Measurement, Statistics and Evaluation, University of Maryland, College Park.

Liu, L., & Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, *62*(1), 261–268.

Liu, L., Hedeker, D., & Mermelstein, R. J. (2013). Modeling nicotine dependence: an application of a longitudinal irt model for the analysis of adolescent nicotine dependence syndrome scale. *Nicotine & Tobacco Research*, *15*(2), 326–333.

Liu, Y., & Yang, J. S. (2017). Bootstrap-calibrated interval estimates for latent variable scores in item response theory. *Psychometrika*, 1–22.

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

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society*, 226–233.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*(4), 344.

Maydeu-Olivares, A., Fairchild, A. J., & Hall, A. G. (2017). Goodness of fit in item factor analysis: Effect of the number of response alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 495–505.

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: a unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 561–614). New York: Plenum.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analyses. , 342–380.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual review of psychology*, *60*, 577–605.

McDonald, R. P. (1982). Linear versus models in item response theory. *Applied Psychological Measurement*, *6*(4), 379–396.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1),

107–122.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551–560.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. *SAGE Focus Editions*, *154*, 205–205.

Muthén, B., du Toit, S. H., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. 1997. *Unpublished technical report*.

Muthén, L., & Muthén, B. (1998-2012). *Mplus user's guide* (Seventh ed.). Los Angeles, CA: Muthén Muthén.

Paek, I., Li, Z., & Park, H.-J. (2016). Specifying ability growth models using a multidimensional item response model for repeated measures categorical ordinal item response data. *Multivariate Behavioral Research*, *51*(4), 569–580.

Patton, J. M., Cheng, Y., Yuan, K.-H., & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement*, *37*(1), 24–40.

Patton, J. M., Cheng, Y., Yuan, K.-H., & Diao, Q. (2014). Bootstrap standard errors for maximum likelihood ability estimates when item parameters are unknown. *Educational and Psychological Measurement*, *74*(4), 697–712.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., et al. (2002). Reliable estimation of

generalized linear mixed models using adaptive quadrature. *The Stata Journal*, *2*(1), 1–21.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Copenhagen: Danish Institute for Educational Research.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods (Second Edition).* Thousand Oaks, CA: Sage Publications.

Reckase, M. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401–412.

Reckase, M. (2009). *Multidimensional item response theory.* London: Springer.

Rijmen, F. (2009). Efficient full information maximum likelihood estimation for multidimensional irt models. *ETS Research Report Series*, *2009*(1), i–31.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *17*.

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 179–200). Washington, DC: American Psychological Association.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53–61.

Stapleton, L. M., Harring, J. R., & Lee, D. (2015). Sampling weight considerations for multilevel modeling of panel data. In G. Hancock, L. Stapleton, & S. Beretvas (Eds.), *Advances in multilevel modeling for educational research:*

Addressing practical issues found in real-world applications (pp. 63–96). IAP.

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

te Marvelde, J. M., Glas, C. A., Van Landeghem, G., & Van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement*, *66*(1), 5–34.

Thissen, D., & Wainer, H. (1990). Confidence envelopes for item response theory. *Journal of Educational and Behavioral Statistics*, *15*(2), 113–128.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, *6*(2), 181.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

Wang, C., Kohli, N., & Henn, L. (2016). A second-order longitudinal model for binary outcomes: item response theory versus structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 455–465.

Wei, G. C., & Tanner, M. A. (1990). A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*(411), 699–704.

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods*, *12*(1), 58.

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44*, 175–188.

Yang, J. S., & Cai, L. (2014). Estimation of contextual effects through nonlinear multilevel latent variable modeling with a metropolis–hastings robbins–monro

algorithm. *Journal of Educational and Behavioral Statistics*, *39*(6), 550–582.

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and psychological measurement*, *72*(2), 264–290.

Young, R., & Johnson, D. R. (2015). Handling missing values in longitudinal panel data with multiple imputation. *Journal of Marriage and Family*, *77*(1), 277–294.