Gaussian Variational Estimation of MIRT and Its Applications in Large-Scale Assessments

Jiaying Xiao

A dissertationsubmitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Chun Wang, Chair

Min Li

David Knight

Program Authorized to Offer Degree:
College of Education

University of Washington

Abstract

Gaussian Variational Estimation of MIRT and Its Applications in Large-Scale Assessments

Jiaying Xiao

Chair of the Supervisory Committee:
Chun Wang
College of Education

Multidimensional Item Response Theory (MIRT) has been widely used in educational and psychological assessments. It estimates multiple constructs simultaneously and models the correlations among latent constructs. While it provides more accurate results, the unidimensional IRT model is still dominant in real applications. One major reason is that the parameter estimation is still challenging because of intractable multidimensional integrals of the likelihood, especially in high dimensions. Several algorithms have been proposed to address the issue, such as adaptive Gaussian quadrature methods, Laplace approximations, and stochastic methods. However, the state-of-the-art algorithms are still time-consuming, especially when the number of latent traits exceeds 5. Recently, the Gaussian variational Expectation Maximization (GVEM) algorithm (Cho et al., 2021) was proposed as an alternative for further improving computational efficiency and estimation accuracy. The general framework allows the closed-form solutions for the expectation-maximization process by introducing a variational lower bound of the likelihood function.

Although prior studies have demonstrated the superiority of the GVEM algorithm over the widely used Metropolis–Hastings Robbins–Monro algorithm (MH-RM) under various conditions, its performance across diverse practical contexts remains relatively unexplored. For instance, there is an immense need for further investigation into the robustness of the GVEM framework across various missing data scenarios. Additionally, efforts should be

directed towards devising methods for estimating standard errors within the GVEM framework. Moreover, the development of an R package to facilitate the application of the GVEM algorithm would significantly augment its accessibility and utility.

The purpose of this dissertation is to extend the applicability of the GVEM algorithm and investigate its performance in diverse scenarios. In the second chapter, a modified GVEM algorithm was proposed by adding the bootstrap bias correction step and denoted it as GVEM-BS. A series of simulation studies and real data analysis were conducted to compare GVEM-BS to MH-RM in terms of estimation precision under different missing data scenarios and assessment designs. The results demonstrated the robustness and precision of GVEM-BS in the context of high missing proportions, especially for missing at completely random conditions. When applying the two methods to different assessment designs, both GVEM-BS and MH-RM yielded comparable results.

In the third chapter, an updated supplemented expectation maximization (USEM) method and a bootstrap method were proposed for GVEM-based SE estimation. These two methods were compared in terms of SE recovery accuracy. The simulation results demonstrated that the GVEM algorithm with bootstrap and item priors (GVEM-BSP) outperformed the other methods, exhibiting less bias and relative bias for SE estimates under most conditions. Although the GVEM with USEM (GVEM-USEM) was the computationally most efficient method, it yielded an upward bias for SE estimates.

In the fourth chapter, an R package, `VEMIRT`, was introduced by offering users efficient computational tools tailored for high-dimensional data under the GVEM framework. This package facilitates both exploratory and confirmatory analyses through the utilization of GVEM models. Additionally, it enables users to compute standard errors of item parameters and implement corrections such as bootstrap sampling and importance sampling, thereby enhancing the accuracy of estimations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DEDICATION

To my family

Chapter 1

INTRODUCTION

Large-scale assessments (LSAs), like the Programme for International Student Assessment (PISA) or the Programme for the International Assessment of Adult Competencies (PIAAC), are designed to assess examinees' multiple competencies as well as the relationships between these competencies and various background variables, including students' socioeconomic backgrounds and attitudes towards learning and their school experiences (Gurkan, 2021). Over the years, LSAs have received significant attention because the findings from LSAs have played a pivotal role in shaping education policies, impacting areas such as pedagogy, teacher training, and instructional hours (Heyneman & Lee, 2014).

Item response theory (IRT), also known as latent trait theory, is a psychometric framework to model the relationship between individuals' responses to test items and their underlying latent constructs or abilities measured by the test. IRT is underpinned by four fundamental assumptions (Lord, 1968): 1) Unidimensionality: the test measures a single latent trait. 2) Local independence: item responses are independent given an examinee's latent trait. 3) Shape of item characteristic curve (ICC): The ICC, reflecting the probability of correctly answering an item as a function of the latent trait, is an S-shaped curve derived from the logistic regression model. This curve increases monotonically as the individual's level of the latent trait rises. 4) Parameter invariance: Item parameters remain constant across samples of examinees drawn from the intended population for whom the test is designed, while ability parameters remain invariant across samples of test items drawn from the population of items assessing the targeted ability.

In practice, the unidimensional assumption becomes untenable when a test measures multiple latent constructs. For instance, consider an English proficiency test that aims to

assess students' abilities in listening, speaking, reading, and writing—all of which are highly related. Similarly, in a life quality questionnaire, items measure several satisfaction domains such as satisfaction with work, family, society, and other relevant factors. Consequently, multidimensional IRT (MIRT) has been developed and gained popularity for handling complex test and survey data. This popularity stems primarily from its capacity to consider the correlations among latent constructs during parameter estimation. Empirical evidence from several studies shows the efficacy of MIRT over unidimensional IRT, particularly in scenarios where the test measures several highly correlated latent constructs with a small number of items (de la Torre & Patz, 2005; W.-C. Wang et al., 2004).

Furthermore, MIRT offers a powerful tool for item analysis, scoring, and calibration within LSAs. However, operational analyses still predominantly rely on unidimensional IRT models. One significant hindrance is the computational challenge posed by large datasets, where the number of items, sample size, and dimensions are all considerable (Ma et al., 2023). For example, the English Language Proficiency Assessment for the 21st Century (ELPA21), encompassing eight competencies assessed by over 600 items, poses significant computational challenges for existing MIRT algorithms. These challenges often lead to prolonged computation times and unreliable estimation outcomes (Huang & Flores, 2018).

Several solutions have been proposed in the literature. For instance, the adaptive Gaussian quadrature method (Cagnone & Monari, 2013) and the Laplace approximation (Lindstrom & Bates, 1988) conduct direct numerical approximations to the integrals. Unfortunately, the former is known to be computationally demanding in high dimensions, whereas the latter, though being computationally efficient, becomes less accurate especially when only a few dichotomous items are measuring each latent trait (Joe, 2008). Other approaches are based on stochastic approximation, such as the Metropolis–Hastings Robbins–Monro (MH-RM) method (Cai, 2010a, 2010b). This method has been implemented in several software programs (Bashkov & DeMars, 2017), but again, it could be computationally intensive since the procedure requires multiple sampling from a posterior distribution. A variational autoencoder (VAE) approach from the deep learning literature has recently gained interest

in fitting MIRT models (T. Liu et al., 2022). Previous studies have shown that VAE-based methods improve item parameter recovery and computational efficiency (Curi et al., 2019; Urban & Bauer, 2021). However, VAE-based methods often lack theoretical support for the consistency of estimators and may underperform with small to medium-sized datasets (Ma et al., 2023).

To address these limitations, Cho et al. (2021) proposed an alternative algorithm, namely, the Gaussian variational expectation–maximization (GVEM) algorithm, to further improve computational efficiency and estimation accuracy. To greatly reduce the computational complexity, the GVEM bypasses calculating intractable integrals by approximating the marginal likelihood with a more tractable variational lower bound, such that both the integration in E-step and solution to the score function in the M-step involve analytic closed forms. Previous research has demonstrated the computational efficiency of GVEM, along with its ability to yield parameter estimates that are comparable to, and sometimes even more accurate than, those produced by the MH-RM algorithm and the constraint joint maximum likelihood estimation (CJMLE) method (Y. Chen et al., 2019) in high-dimensional exploratory item factor analysis models (Cho et al., 2021; Cho et al., 2022). Furthermore, Cho et al. (2021) have proved the consistency of parameter estimates obtained from the GVEM algorithm under high-dimensional settings.

Since the GVEM algorithm has recently been proposed, it is not surprising that there is little literature conducting both methodological and applied investigations of this algorithm. Specifically, while parameter estimation in MIRT poses challenges, addressing missing data simultaneously places higher demands on the robustness of estimation algorithms. Despite the common occurrence of missingness in educational assessments, there has been a notable absence of studies comparing the performance of the MH-RM and GVEM algorithms when the dataset contains missing values across various missing data scenarios. Another critical issue that has yet to be fully addressed in the GVEM literature is the procedure for estimating standard errors (SEs). While point estimation of item parameters provides valuable insights, SEs play an equally crucial role in statistical inference. Understanding SEs is essential for

assessing the precision of parameter estimates and making informed decisions about model fit and test reliability. Furthermore, one potential reason for the limited adoption of applications utilizing the GVEM algorithm is its mathematical complexity, which necessitates a certain level of understanding to grasp the estimation process fully. In contrast, one of the key factors driving the popularity of the MH-RM algorithm is its widespread availability in various software, such as IRTPRO (Cai, Du Toit, et al., 2011), flexMIRT (Cai, 2013), and the `mirt` package (Chalmers, 2012) in R (R Core Team et al., 2013), making it easily applicable to a broader audience.

The primary goals of this dissertation are to address gaps in the existing literature and to expand the applicability of the GVEM algorithm. The structure of the dissertation is outlined as follows. In the following sections of this chapter, a brief overview of the MIRT model is provided and various methods for estimating model parameters within the MIRT framework are discussed. Additionally, an in-depth explanation of the GVEM algorithm is introduced. The second chapter presents a comparative study between a modified version of the GVEM algorithm and the MH-RM algorithm under diverse missing data scenarios for both simulated and real datasets. The third chapter delves into a study focused on standard error procedures within the GVEM framework, exploring methods to enhance the precision of parameter estimates. The fourth chapter introduces the development of the R package, `VEMIRT`, designed to facilitate the application of GVEM algorithms for handling high-dimensional data. By systematically addressing these objectives, the dissertation aims to contribute to the advancement of parameter estimation methods in MIRT and to provide practical tools for researchers in the field. The final chapter summarizes the results findings and provides suggestions for future research.

1.1 Multidimensional Item Response Theory (MIRT)

As an extension of unidimensional IRT, MIRT estimates multiple latent traits simultaneously. MIRT can be classified in several ways, including by response type (dichotomous or polytomous), the number of item parameters (e.g., two-parameter or three-parameter models),

and model structure (between-item or within-item). The primary focus of this dissertation is multidimensional two-parameter logistic (M2PL) model for dichotomous responses, which is one of the most widely used MIRT models in practice (Reckase, 2009) and and serves as the foundation for the methodological and applied investigations conducted herein.

Suppose N examinees respond to J items, resulting in a binary response matrix $\mathbf{Y} = \{\mathbf{Y}_i, i = 1, \ldots, N\}$ where $\mathbf{Y}_i = \{Y_{ij}, j = 1, \ldots, J\}$ refers to the ith examinee's response vector. The item response function of the ith respondent to the jth item is

$$P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i) = \frac{\exp(\boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i - b_j)}{1 + \exp(\boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i - b_j)}, \tag{1.1}$$

where $\boldsymbol{\theta}_i$ denotes the K-dimensional vector of latent ability for examinee i, $\boldsymbol{\alpha}_j$ denotes a $K-$dimensional vector of item discrimination parameters for jth item ($\boldsymbol{\alpha}_j = \{\alpha_{jk}, k = 1, \ldots, K\}$), b_j denotes the corresponding item difficulty parameter. For model identification purposes, the population means and variances of $\boldsymbol{\theta}_i$ are fixed at zero and one, respectively, while the covariance (which refers to correlation) among the latent traits $\boldsymbol{\theta}_i$ is left to be freely estimated. Note that when $\boldsymbol{\alpha}_j$ and $\boldsymbol{\theta}_i$ are unidimensional, the M2PL model becomes a unidimensional 2PL IRT model. For conciseness, let $\mathbf{M}$ denote all model parameters for simplicity, where $\mathbf{M} = \{\mathbf{A}, \mathbf{B}\}$, $\mathbf{A} = \{\boldsymbol{\alpha}_j, j = 1, \ldots, J\}$ and $\mathbf{B} = \{b_j, j = 1, \ldots, J\}$. Considering the local independence assumption in IRT, the marginal log-likelihood function can be expressed as

$$l(\mathbf{M}; \mathbf{Y}) = \sum_{i=1}^{N} \log P(\mathbf{Y}_i \mid \mathbf{M}) = \sum_{i=1}^{N} \log \int \prod_{j=1}^{J} P(Y_{ij} \mid \boldsymbol{\theta}_i, \mathbf{M}) \phi(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i, \tag{1.2}$$

where ϕ denotes the multivariate normal density function of $\boldsymbol{\theta}_i$, with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. The model parameters can be estimated by maximizing this marginal log-likelihood function. However, due to the latent trait structure, this maximization process involves K-dimensional integrals, which makes the estimation process computationally intensive and typically intractable (Cai, 2010a, 2010b). To tackle this issue, several methods have been proposed in the literature over the last few decades, which will be discussed in the following section.

1.2 Parameter Estimation Methods for MIRT

Several full-information methods have been developed to address the computation challenges in MIRT estimation. One such method is adaptive Gaussian quadrature, which approximates the integral directly (Naylor & Smith, 1982). Unlike fixed Gaussian-Hermite quadrature methods (e.g., as discussed by Bock and Aitkin, 1981), adaptive quadrature optimizes the utilization of integration points, thereby requiring fewer quadrature points without compromising the precision of approximation (Schilling & Bock, 2005). However, despite the potential for reducing the number of quadrature points per dimension, the total number of quadrature points still grows exponentially with the number of dimensions. Furthermore, an additional computational step is necessary for each iteration to compute the posterior mode and variance of latent factors, thereby increasing the overall computation costs (Pinheiro & Bates, 1995).

Another method is a fully Bayesian estimation approach, such as Markov chain Monte Carlo (MCMC; Albert, 1992; Patz and Junker, 1999). MCMC is a general approach that integrates the likelihood function of data with prior distributions of model parameters to derive a posterior distribution of the given data. The point and interval estimates of model parameters are obtained by drawing parameter values from sampling distributions and adjusting the draws to approximate the target posterior distribution (Garnier-Villarreal et al., 2021; Lee & Song, 2012; Ulitzsch & Nestler, 2022). MCMC yields precise parameter estimates and allows for flexible model specification, such as customizing MCMC samplers for skewed IRT models (Ulitzsch & Nestler, 2022; X. Zhang et al., 2021). It proves particularly valuable for complex high-dimensional IRT models. Nonetheless, it can be very time-consuming, requiring long chains to converge, especially with complex models (Cho et al., 2022; Ma et al., 2023). This becomes more evident when researchers utilize general-purpose samplers instead of those tailored to the specific model employed (Ulitzsch & Nestler, 2022).

Stochastic approximations, akin to the Bayesian approach, circumvent intractable integrations by sampling from the posterior distributions, with examples including the Metropolis-

Hastings Robbins-Monro algorithm (MH-RM; Cai, 2010a, 2010b). MH-RM employs the Metropolis-Hastings (MH) sampler to draw missing data (referred to as $\boldsymbol{\theta}$ in MIRT) during the stochastic imputation step. Subsequently, the random draws are combined through stochastic approximation to guide the adjustments to the estimates in each iteration via the Robbins-Monro (RM) algorithm. Therefore, MH-RM can be viewed as an extension of the stochastic approximation EM (SAEM) algorithm (Delyon et al., 1999). Unlike Monte Carlo EM (MCEM; McCulloch, 1997), MH-RM demonstrates efficiency in Monte Carlo usage since the simulation size remains fixed and typically small throughout iterations. Additionally, it yields an estimate of the parameter information matrix as a by-product (Cai, 2010b). The MH-RM algorithm has exhibited remarkable stability and efficiency in practical applications. Nonetheless, it may still be computationally intensive for complex high-dimensional models, as a large Monte Carlo sample size is often required, and the posterior distributions generally lack closed forms (Cho et al., 2022).

The Laplace approximation relies on a second-order Taylor expansion of the log-integrand around its mode (Lindstrom & Bates, 1988), resulting in a tractable integral. The first-order Laplace approximation is equivalent to adaptive quadrature with only one quadrature point per dimension. Consequently, the computational demand of the first-order Laplace approximation grows linearly with increasing dimensionality (Andersson & Xin, 2021). Thus, in high-dimensional models, the first-order Laplace approximation is the most computationally efficient method among the three options mentioned. However, it may become less accurate as the number of dimensions increases beyond three or when the sample size is small (Jeon et al., 2017), or the likelihood function is skewed. To enhance computational accuracy, higher-order Laplace approximations can be pursued. However, this approach necessitates a substantial number of higher-order derivatives, which substantially increases the computational expense, particularly in the context of high-dimensional models (Andersson & Xin, 2021; Bianconcini & Cagnone, 2012).

In addition to the aforementioned full-information methods, a recent approach known as constraint joint maximum likelihood estimation (CJMLE) was introduced by Y. Chen et

al. (2019). This method offers higher computational efficiency compared to many marginal maximum likelihood methods by treating the latent abilities as fixed effect parameters rather than random variables, and the estimator is theoretically guaranteed to be consistent under high-dimensional settings. Building upon CJMLE, H. Zhang et al. (2020) proposed a singular value decomposition (SVD) based estimator, which further enhances the performance of CJMLE. These joint maximum likelihood methods are characterized by their low computational cost. However, they sacrifice the flexibility of latent factors by treating them as fixed effects. For instance, it may be challenging conceptually to extend the algorithm to multiple-group conditions where unbiased estimation of group-specific population distributions is often required, as opposed to estimating an individual's latent trait as a fixed effect (Ma et al., 2023).

Recently, variational estimation methods stemming from the machine learning literature have gained increased interest in psychometrics (Cho et al., 2021). The key idea of these methods lies in approximating intractable integrals, such as Equation (1.2), with computationally feasible forms known as a variational lower bound. Rijmen and Jeon (2013) were among the first to develop a variational algorithm for MIRT models, albeit limited to discrete latent variables. Subsequent studies have explored a wide range of variational methods for the estimation of more complex models. For instance, Jeon et al. (2017) introduced the variational maximization-maximization (VMM) algorithm for generalized linear mixed models (GLMMs), which outperformed Laplace approximation in scenarios with small sample sizes. Nonetheless, the reliance on iterative numerical algorithms in each maximization step led to a slow speed in algorithm execution (Ma et al., 2023).

To enhance computational efficiency, researchers have turned to variational autoencoder (VAE), a deep learning-based approach to tackle estimation problems in MIRT models (Curi et al., 2019). Essentially, VAE leverages two neural networks to maximize the variational lower bound: the encoder network maps the data to a probability distribution, which is then sampled from and reconstructed through the decoder network (Kingma & Welling, 2013). Previous studies have shown that VAE-based methods significantly improve item parameter

recovery (Converse et al., 2021; Hasan et al., 2022). As an extension, the importance-weighted VAE has been developed and demonstrates competitive performance compared to other estimation methods, all while achieving faster speeds (T. Liu et al., 2022; Urban & Bauer, 2021). However, it is worth noting that VAE-based methods lack theoretical support for the consistency of estimators and may exhibit suboptimal performance in small to medium-sized datasets (Ma et al., 2023). However, VAE-based methods lack theoretical support for the consistency of estimators and may exhibit suboptimal performance in small to medium-sample data (Ma et al., 2023).

In contrast, Cho et al. (2021) proposed the Gaussian Variational Expectation-Maximization (GVEM) algorithm, which has demonstrated computational speed and produced comparable or more accurate parameter estimates than the MH-RM algorithm and the CJMLE method in high-dimensional exploratory item factor analysis models. Additionally, the literature provided both theoretical and empirical evidence of the consistency of estimated parameters under high-dimensional settings (Cho et al., 2021; Cho et al., 2022). Therefore, this dissertation focuses on investigating the GVEM algorithm.

1.3 Gaussian Variational Expectation Maximization (GVEM)

In this section, I will briefly introduce the GVEM algorithm discussed in Cho et al. (2021). The key idea of the GVEM algorithm is to employ a variational approximation of the intractable marginal log-likelihood function within the EM framework. The derivation is shown as follows. Following notations in Cho et al. (2021), the marginal log-likelihood in Equation 1.2 can be rewritten as

$$\begin{aligned} l(\mathbf{M}; \mathbf{Y}) &= \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log P(\mathbf{Y}_i \mid \mathbf{M}) \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\ &= \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log \frac{P(\mathbf{Y}_i, \boldsymbol{\theta}_i \mid \mathbf{M})}{P(\boldsymbol{\theta}_i \mid \mathbf{Y}_i, \mathbf{M})} \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\ &= \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log \frac{P(\mathbf{Y}_i, \boldsymbol{\theta}_i \mid \mathbf{M})}{q_i(\boldsymbol{\theta}_i)} \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i + KL\{q_i(\boldsymbol{\theta}_i) \| P(\boldsymbol{\theta}_i \mid \mathbf{Y}_i, \mathbf{M})\}, \end{aligned}$$

where $KL\{q_i(\boldsymbol{\theta}_i)\|P(\boldsymbol{\theta}_i \mid \mathbf{Y}_i, \mathbf{M})\} = \int_{\boldsymbol{\theta}_i} \log \frac{q_i(\boldsymbol{\theta}_i)}{P(\boldsymbol{\theta}_i|\mathbf{Y}_i,\mathbf{M})} \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$ denotes the Kullback-Leibler (KL) divergence between the posterior distribution $P(\boldsymbol{\theta}_i \mid \mathbf{Y}_i, \mathbf{M})$ and an arbitrary probability density function $q_i(\boldsymbol{\theta}_i)$. Note that $KL\{q_i(\boldsymbol{\theta}_i)\|P(\boldsymbol{\theta}_i \mid \mathbf{Y}_i, \mathbf{M})\} \geq 0$. so a lower bound of the marginal log-likelihood can be obtained as

$$l(\mathbf{M}; \mathbf{Y}) \geq \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log P(\mathbf{Y}_i, \boldsymbol{\theta}_i \mid \mathbf{M}) \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i - \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log q_i(\boldsymbol{\theta}_i) \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \tag{1.3}$$

Since the equality in (1.3) holds if and only if $q_i(\boldsymbol{\theta}_i) = P(\boldsymbol{\theta}_i \mid Y_i, \mathbf{M})$ for $i = 1, \ldots, N$, the best choice of $q_i(\boldsymbol{\theta}_i)$ is the the posterior distribution $P(\boldsymbol{\theta}_i \mid Y_i, \mathbf{M})$. However, it is not practically applicable considering that the posterior distribution $P(\boldsymbol{\theta}_i \mid Y_i, \mathbf{M})$ is unknown. Alternatively, $q_i(\boldsymbol{\theta}_i)$ can be chosen from a normal distribution and a local variational method (Bishop, 2006; Jordan et al., 1999) is employed to obtain a closed-form lower bound expression of the expected log-likelihood with respect to $q_i(\boldsymbol{\theta}_i)$. Following the derivations in Cho et al. (2021), the optimal choice of $q_i(\boldsymbol{\theta}_i)$ is $q_i(\boldsymbol{\theta}_i) \sim N(\boldsymbol{\theta}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, and its mean and covariance are

$$\boldsymbol{\mu}_i = \boldsymbol{\Sigma}_i \times \sum_{j=1}^{J} \left\{ 2\eta(\xi_{i,j}) b_j + Y_{ij} - \frac{1}{2} \right\} \boldsymbol{\alpha}_j^\top, \tag{1.4}$$

$$\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} + 2 \sum_{j=1}^{J} \eta(\xi_{i,j}) \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j^\top, \tag{1.5}$$

where $\xi_{i,j}$ denotes a variational parameter indexed by i and j, and $\eta(\xi_{i,j}) = (2\xi_{i,j})^{-1}[e^{\xi_{i,j}}/(1 + e^{\xi_{i,j}}) - 1/2]$.

Let $E^{(t)}(\mathbf{M}, \boldsymbol{\xi})$ denote the tth iteration's lower bound expression of the expected log-likelihood, and now it has a closed form expression

$$\begin{aligned} E^{(t)}(\mathbf{M}, \boldsymbol{\xi}) = & \sum_{i=1}^{N} \sum_{j=1}^{J} \Bigg(\log \frac{e^{\xi_{i,j}^{(t)}}}{(1 + e^{\xi_{i,j}^{(t)}})} + (\frac{1}{2} - Y_{ij}) b_j^{(t)} + (Y_{ij} - \frac{1}{2}) \boldsymbol{\alpha}_j^{(t)\top} \boldsymbol{\mu}_i^{(t)} - \frac{1}{2} \xi_{i,j}^{(t)} \\ & - \eta(\xi_{i,j}^{(t)}) \{ b_j^{(t)2} - 2 b_j^{(t)} \boldsymbol{\alpha}_j^{(t)\top} \boldsymbol{\mu}_i^{(t)} + \boldsymbol{\alpha}_j^{(t)\top} [\boldsymbol{\Sigma}_i^{(t)} + (\boldsymbol{\mu}_i^{(t)})(\boldsymbol{\mu}_i^{(t)})^\top] \boldsymbol{\alpha}_j^{(t)} - \xi_{i,j}^{(t)2} \} \Bigg) \\ & + \frac{N}{2} \log |\Sigma_{\boldsymbol{\theta}}^{(t)-1}| - \sum_{i=1}^{N} \frac{1}{2} Tr(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(t)-1} [\boldsymbol{\Sigma}_i^{(t)} + (\boldsymbol{\mu}_i^{(t)})(\boldsymbol{\mu}_i^{(t)})^\top]). \end{aligned} \tag{1.6}$$

In every E step, the expectation function is updated iteratively with all recently updated model parameters. In every M step, the $E^{(t)}(\mathbf{M}, \boldsymbol{\xi})$ is maximized to estimate the parameters $(\mathbf{M}, \boldsymbol{\xi})$. This is achieved by setting the derivative of $E^{(t)}(\mathbf{M}, \boldsymbol{\xi})$ with respect to $(\mathbf{M}, \boldsymbol{\xi})$ to be zero, leading to the following updated equations:

$$\begin{aligned}
\boldsymbol{\alpha}_j &= \frac{1}{2}\Big[\sum_{i=1}^{N} \eta(\xi_{i,j})\boldsymbol{\Sigma}_i + \eta(\xi_{i,j})\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top\Big]^{-1} \sum_{i=1}^{N}\Big[\Big(Y_{ij} - \frac{1}{2} + 2b_j\eta(\xi_{i,j})\Big)\boldsymbol{\mu}_i^\top\Big], \qquad (1.7)\\
b_j &= \frac{\sum_{i=1}^{N}\Big[(\frac{1}{2} - Y_{ij}) + 2\eta(\xi_{i,j})\boldsymbol{\alpha}_j^\top\boldsymbol{\mu}_i\Big]}{\sum_{i=1}^{N} 2\eta(\xi_{i,j})}, \qquad (1.8)\\
\xi_{i,j}^2 &= b_j^2 - 2b_j\boldsymbol{\alpha}_j^\top\mu_i + \boldsymbol{\alpha}_j^\top[\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top]\boldsymbol{\alpha}_j. \qquad (1.9)
\end{aligned}$$

Note that for the population covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, in the exploratory model estimation, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ can be treated as an identity matrix during the GVEM estimation and then later proper rotation is conducted to produce non-zero correlations; in the confirmatory model estimation, $\Sigma_{\boldsymbol{\theta}}$ is updated by

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \frac{1}{N}\sum_{i=1}^{N}[\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top]. \qquad (1.10)$$

Because it is necessary to fix the diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ during estimation to fix the scale, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ should be rescaled after the convergence of the M-step by

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^* = [\sqrt{\text{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})}^{-1}]^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}[\sqrt{\text{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})}^{-1}], \qquad (1.11)$$

where $\text{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ is a column vector whose elements are the values on the main diagonal of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. The item discrimination parameter needs to be rescaled accordingly by $\boldsymbol{\alpha}_j^* = \boldsymbol{\alpha}_j\sqrt{\text{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})}$.

In light of the preceding discussion, the GVEM algorithm for the M2PL model can be summarized as follows.

Algorithm 1: GVEM algorithm

Input: Binary response matrix $\mathbf{Y}$

Initialize $\boldsymbol{M}^{(0)} = \{\boldsymbol{A}^{(0)}, \boldsymbol{B}^{(0)}\}, \boldsymbol{\xi}^{(0)}$;

while *not converged* **do**

E step: For t-the iteration, update $\boldsymbol{\mu}_i^{(t)}$ and $\boldsymbol{\Sigma}_i^{(t)}$ according to equations (1.4) and (1.5);

M step: Update $\boldsymbol{M}^{(t)}$ and $\boldsymbol{\xi}^{(t)}$ according to equations (1.7), (1.8), and (1.9). For the confirmatory model estimation, update $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(t)}$ according to (1.10), and rescale $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(t)}$ and $\boldsymbol{\alpha}_j^{(t)}$;

end

Output: $\hat{\boldsymbol{M}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$

Chapter 2

A COMPARISON OF PARAMETER ESTIMATION ALGORITHMS FOR MIRT MODELS WITH MISSING DATA

The existence of missingness is common in large-scale assessments (LSAs) and can stem from various sources. For instance, respondents might inadvertently skip items, run out of time to answer certain questions, or intentionally omit items due to uncertainty about the correct answers (Xiao & Bulut, 2020). Additionally, item nonresponse may arise from the assessment design itself, where certain items are not administered to respondents (Köhler et al., 2017). Item nonresponse can undermine the accurate scaling of competencies, particularly when it pertains to unobserved responses—meaning the true value on the item if it had been observed (Mislevy & Wu, 1996). Understanding the mechanisms behind missingness is crucial as it fosters a deeper comprehension of variations in test-taking behavior and enables the consideration of these differences when making inferences about examinee competencies (Ulitzsch, 2020). Additionally, it informs decisions regarding the treatment of missing data in subsequent procedures.

While multidimensional item response theory (MIRT) offers a robust framework for item analysis, scoring, and calibration within LSAs, estimating item parameters for high-dimensional MIRT models can pose computational challenges due to intractable integrals in the likelihood function (Andersson & Xin, 2021). To address this, several algorithms have been proposed, with the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010a) being a popular choice. Recently, the Gaussian variational EM (GVEM) algorithm (Cho et al., 2021) was proposed as an alternative for further improving computational efficiency and estimation precision. While previous research showed the GVEM outperformed the MH-RM with regard to item parameter recovery under different exploratory factor anal-

ysis conditions, no studies have compared them when the dataset has missing values.

The robustness of a statistical method hinges on its ability to provide reasonable statistical inference and withstand the impact of outliers or missing data (Zhu et al., 2018). Therefore, it is worthwhile to investigate the robustness of MH-RM and GVEM algorithms when data have missing issues. The MH-RM algorithm can handle missing data by using the full information maximum likelihood estimation (Cai, 2010a). To apply GVEM to the high missing proportion conditions, the current study proposed a modified version of the GVEM algorithm by adding a bootstrap bias correction step (Pfeffermann & Correa, 2012) and denoted the new algorithm as GVEM-BS.

The goals of the current study are twofold. Firstly, it proposes a modified version of the GVEM algorithm to enhance its robustness under missing data scenarios. Secondly, it investigates the performance of two GVEM-based methods compared to the MH-RM algorithm under various missing data scenarios. The structure of this paper is as follows: first, we provide a brief discussion on missing data in LSAs. Next, we review several approaches to handling missing data in LSAs. Following this, we present two simulation studies. The first study evaluates the recovery of item parameters for three methods across different conditions, while the second study examines the performance of each method when missing data arises due to the assessment design. Finally, we illustrate the application of these methods using two real datasets.

2.1 Missing Data in Large-Scale Assessments

Missing data is a prevalent issue in LSAs, often categorized as either planned or unplanned. Planned missing values arise from the assessment design itself, such as balanced incomplete block (BIB) design or multistage adaptive testing (MST), where certain items are non-administered (Köhler et al., 2017). A BIB design splits the item pool into a set of blocks and these blocks are assigned to booklets. The split need not be random but may be based on some practical issues, such as to match the completion time for each booklet (Van der Linden et al., 2004). Booklets are spiraled across students randomly. The BIB design guarantees

that each block appears with the same frequency and the positional effect is controlled. Table 2.1 illustrates a typical BIB design. In LSAs utilizing BIB designs like PISA, the missing data pattern is often recognized as missing completely at random (MCAR). It is because booklets are randomly allocated to respondents, and the missingness is independent of both observed and unobserved values (Mislevy & Wu, 1996). Let $\boldsymbol{W}$ denote as the missing data indicator matrix, which has the same dimensions as the response matrix $\boldsymbol{Y}$. W_{ij} indicates whether response Y_{ij} is missing, where $W_{ij} = 1$ if it is missing, and 0 otherwise. Let $\boldsymbol{\psi}$ refer to the unknown parameters of the distribution of $\mathbf{W}$. For all possible values of $\boldsymbol{\psi}$, MCAR means that missingness does not depend on the response dataset $\mathbf{Y}$, whether missing or observed. Mathematically, the MCAR mechanism can be expressed as follows (Rubin, 1976): $\mathrm{p}(\mathbf{W} \mid \mathbf{Y}, \boldsymbol{\psi}) = \mathrm{p}(\mathbf{W} \mid \boldsymbol{\psi})$ for all $\mathbf{Y}$ and $\boldsymbol{\psi}$.

By contrast, MST refers to a group sequential design, in which items are grouped into modules that are matched to an examinee's provisional ability estimates (Chang, 2015). A simple two-stage MST design is presented in Figure 2.1. LSAs utilizing MST, such as PIAAC, employ an adaptive routing system where examinees are directed to item blocks based on their performance in the previous stage. Consequently, more proficient examinees are more likely to encounter a more challenging set of items (Yamamoto et al., 2018). The incomplete response data stemming from MST designs can be considered a missing at random (MAR) scenario (Mislevy & Wu, 1988; C. Wang et al., 2020), aligning with the definition of MAR, which asserts that missingness depends on observed responses rather than unobserved ones (Roth, 1994). That can be expressed as: $\mathrm{p}(\mathbf{W} \mid \mathbf{Y}, \boldsymbol{\psi}) = \mathrm{p}(\mathbf{W} \mid \mathbf{Y}_{obs}, \boldsymbol{\psi})$ for all $\mathbf{Y}_{mis}$ and $\boldsymbol{\psi}$.

These planned missing designs are commonly employed to optimize study costs or enhance data quality by alleviating participant burden. Importantly, researchers have direct control over planned missing data, ensuring that its presence does not introduce systematic bias into statistical analyses (Imbriano, 2018). Therefore, it is possible to ignore planned missing values due to assessment designs (Köhler et al., 2017). In other words, there is no need to integrate the missingness mechanism into models for observed data processes (Holman & Glas, 2005).

Table 2.1: A balanced incomplete block design

Booklet	Block		
1	A	B	C
2	B	C	D
3	C	D	E
4	D	E	F
5	E	F	G
6	F	G	A
7	G	A	B

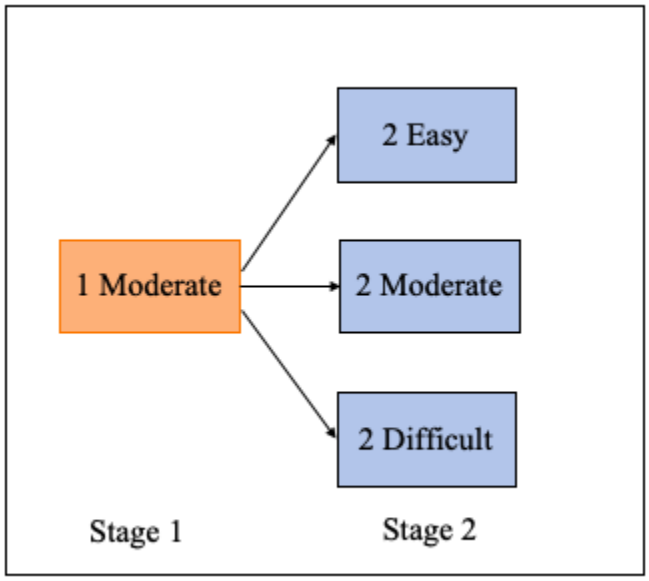

Figure 2.1: A two-stage MST design

In LSAs, unplanned missing data can be categorized into two types: omitted and not-

reached items, both of which present challenges to the scaling and item calibration process (Köhler et al., 2017; Ulitzsch, 2020). Omitted items refer to skipped responses where the examinee has seen an item but opted not to respond, which can occur at any section of the test (Mislevy & Wu, 1996). In contrast, not-reached items occur when examinees fail to attempt a sequence of items presented towards the end of a test (Mislevy & Wu, 1996). Unplanned missing responses are typically considered missing not at random (MNAR), as their occurrence is not determined by either the test developer or the observed responses but may depend on unobserved factors (Rose et al., 2010). MNAR can be expressed as: $\mathrm{p}(\mathbf{W} \mid \mathbf{Y}, \boldsymbol{\psi}) \neq \mathrm{p}(\mathbf{W} \mid \mathbf{Y}_{obs}, \boldsymbol{\psi})$. This violation of ignorability assumptions necessitates the establishment of a model for the processes underlying missing data in order to draw unbiased inferences about the parameters of interest (Rubin, 1976).

The current study will focus on discussions of planned missing data, that is, MCAR and MAR.

2.2 Missing Data Handling Approaches

This section provides a brief overview of various approaches proposed in the literature to handle missing values in LSAs. Some classical approaches, like treating missing values as incorrect responses or partially correct, have been implemented in LSAs such as the National Assessment of Educational Progress (NAEP) and PISA (Allen et al., 1999; Ray, Margaret, et al., 2003). When missing responses are scored as (partially) incorrect, it assumes that respondents who skip some items on the test lack the proficiency to find the correct answer (Xiao & Bulut, 2020). However, research has revealed biases in item and person parameter estimates when missing values were scored as incorrect (De Ayala et al., 2001; Finch, 2008; Köhler et al., 2017; Rose et al., 2010). Even the method of fractionally correct scoring, which performed slightly better, still resulted in bias, particularly when missing values were MNAR (De Ayala et al., 2001; Finch, 2008; Köhler et al., 2017). Other imputation techniques, such as mean substitution or regression imputation, may also introduce bias. These biases affect both the estimates themselves and the associated significance tests, as they reduce standard

errors and introduce artificial certainty into the estimates (Little et al., 2014).

Multiple imputation (MI) has emerged as a widely-used method to address missing data (Graham et al., 2006; Rubin, 1976). MI employs a two-step process. Firstly, multiple imputed datasets are generated, where missing values are replaced with imputed values. Subsequently, standard complete case analysis procedures are applied to each imputed dataset. The results from these analyses are then combined to produce a single outcome, akin to results obtained from complete cases analyses (Graham et al., 2006). Various adaptations of MI exist in the literature, with one popular alternative proposed by Raghunathan et al. (2001) known as MI using chained equations (MICE). The MICE approach assumes that the data are drawn from a multivariate distribution, allowing each incomplete variable to be imputed through iterative sampling from a conditional distribution (Van Buuren & Oudshoorn, 1999; Xiao & Bulut, 2020). A notable advantage of MICE is its flexibility in specifying conditional distributions, as it allows the conditional distributions of the variables (such as item scores in the context of educational assessments) to be modeled according to their specific characteristics (Sinharay, 2021). Comparative studies have shown that MICE could yield accurate results, but it could be computationally demanding (Sinharay, 2021; Xiao & Bulut, 2020).

Full-Information Maximum Likelihood (FIML) is another popular method for addressing missing data. Unlike imputation approaches, FIML utilizes all available data to estimate parameters directly, without replacing or imputing missing values (Eekhout et al., 2015). In FIML estimation, missing responses are simply ignored during parameter estimation for specific items, while information from observed responses is utilized to estimate the parameters for those items. This is achieved through casewise log-likelihoods, allowing FIML to infer the overall model structure from observed data alone, without requiring knowledge of the missing responses (Little et al., 2014; Xiao & Bulut, 2020). A key advantage of FIML is its ability to simultaneously estimate parameters and their standard errors in a single step, enhancing efficiency compared to data imputation methods (Graham, 2009). Research indicates that FIML tends to produce unbiased parameter estimates under both MCAR and MAR condi-

tions (Enders & Bandalos, 2001; Xiao & Bulut, 2020). Moreover, FIML commonly serves as the default missing data technique in many IRT software programs, facilitating its practical application (J. M. Edwards & Finch, 2018). Given these benefits, both the MH-RM and GVEM methods discussed in this study employed FIML to handle missing data effectively.

2.3 Simulation Study 1

2.3.1 Simulation Design

A simulation study was conducted under the between-item multidimensional two-parameter logistic (M2PL) framework. The true item parameters were selected from the item bank of the NAEP combined national and state assessments for grade 8 in 2013. It should be noted that item parameters were fixed to be the same across different replications. Six manipulated factors were considered: (1) analysis type (both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA)), (2) missing data mechanisms (MCAR and MAR), (3) missing proportions (0 to 0.9 with an interval of 0.1), (4) the test length (30 and 45), (5) the correlations among dimensions (0.1 and 0.7), (6) item parameter estimation algorithms (GVEM, GVEM-BS, and MH-RM) The sample size was fixed at 1000 and the number of dimensions was 3.

In EFA, there were no constraints on the loading matrix $\mathbf{A}$ and the covariance matrix $\boldsymbol{\Sigma_\theta}$ was set to be an identity matrix during estimation. The oblimin rotaion was applied here to allow factors to be correlated. Different from EFA, CFA has a pre-specified item factor loading structure, thereby constraining many item discrimination parameters (or loading parameters) in $\mathbf{A}$ to 0. To generate MCAR, the desired proportions of missing data were randomly generated by replacing original responses with missing values. This process ensured that each response had an equal chance of being missing. For MAR, missing values in one item were generated using a logistic model based on other observed variables (i.e., responses to other items). Specifically, missing values on item 1 were initially generated depending on the remaining items, where all items carried the same weight/coefficient. Subsequently,

missing values on other items were generated in a similar manner. When the test length was 30, 10 items loaded onto each factor, similar to the condition where 15 items loaded onto each factor for the 45-item test. Under all conditions, the true latent traits were simulated from $MVN(\mathbf{0}, \mathbf{\Sigma_\theta})$, where $\mathbf{\Sigma_\theta}$ is a covariance matrix with variance of 1 and $r = 0.1$ or 0.7 in the off-diagonal.

2.3.2 Estimation Algorithms

The MH-RM algorithm was conducted by using the *mirt* package (Chalmers, 2012) in R. In this study, we used 5 as the default number of Metropolis-Hastings draws per iteration in the *mirt* package, which was considered sufficient by Cai (2010a).

GVEM algorithm is more computational efficient than the MH-RM algorithm since GVEM uses a variational lower bound to derive closed-form updates for all model parameters in Expectation-Maximization (EM) steps (Cho et al., 2021). When missing values exist in the response matrix, the expectation formula could be rewritten as

$$
\begin{aligned}
E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}) = \sum_{\substack{i=1 \\ i,j:W_{ij}=0}}^{N} \sum_{j=1}^{J} \Bigg(& \log \frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} + (\frac{1}{2} - Y_{ij}) b_j + (Y_{ij} - \frac{1}{2}) \boldsymbol{a}_j^\top \boldsymbol{\mu}_i - \frac{1}{2} \xi_{i,j} \\
& - \eta(\xi_{i,j}) \{ b_j^2 - 2 b_j \boldsymbol{a}_j^\top \boldsymbol{\mu}_i + \boldsymbol{a}_j^\top [\Sigma_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top] \boldsymbol{a}_j - \xi_{i,j}^2 \} \Bigg) \\
& + \frac{N}{2} \log |\mathbf{\Sigma}_\theta^{-1}| - \sum_{i=1}^{N} \frac{1}{2} Tr(\mathbf{\Sigma}_\theta^{-1} [\mathbf{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top])
\end{aligned} \tag{2.1}
$$

Note here we assume missing data can be ignorable. When the response matrix has no missing values, equations (1.6) and (2.1) are the same.

According to our preliminary simulation studies, we found GVEM would produce biased estimates for item parameters since it estimated a lower bound function instead of the real likelihood function. Therefore, we adopted the bootstrap procedure (Pfeffermann & Correa, 2012) to correct the bias and denoted it as GVEM-BS. The procedure is straightforward. For replication r, GVEM algorithm is implemented to estimate item parameters (denoted as

$\mathbf{MR}_r, r = 1, 2...R$, where R is the total number of replications). Based on $\mathbf{MR}_r$, we simulate some bootstrap datasets ($\boldsymbol{B}_{r1}, \boldsymbol{B}_{r2}, ...\boldsymbol{B}_{rB}$), where B refers to the boostrap sampler) and estimate item parameters by using GVEM ($\mathbf{MB}_{r1}, \mathbf{MB}_{r2}, ...\mathbf{MB}_{rB}$). The corrected item parameters $\mathbf{MC}_r$ for rth replication can be computed as

$$\mathbf{MC}_r = 2 \times \mathbf{MR}_r - \frac{\sum_{b=1}^{B} \mathbf{MB}_{rb}}{B} \tag{2.2}$$

In our study, 10 bootstrap datasets were simulated for each replication. The average bias, root mean squared error (RMSE) were computed based on the results of 50 replications.

2.3.3 Study 1 Results

Figures 2.2 and 2.2 compare the bias and RMSE values for each method when missing data mechanism is MCAR. It should be noted that MH-RM failed to converge under some replications and these unsuccessful replications were excluded for further analysis. Table 2.2 summarizes the number of unsuccessful replications with missing values for MH-RM. Overall, GVEM-BS produced comparable or superior accuracy in item parameter estimation compared to both MH-RM and GVEM. In scenarios with low missing proportions, all methods yielded unbiased estimates for b parameters. However, MH-RM tended to overestimate a parameters, whereas GVEM-BS and GVEM tended to underestimate them. As the missing proportions increased, a parameters were severely biased for all methods accordingly, with GVEM-BS showcasing superior performance over MH-RM and GVEM. Interestingly, GVEM-based methods maintained accurate estimates for b parameters even under high missing proportion conditions, while MH-RM exhibited underestimation. Notably, both GVEM and GVEM-BS performed better in EFA conditions than CFA conditions, especially for a parameters. Another important observation was the improvement of GVEM-BS's performance with high factor correlation, whereas MH-RM and GVEM exhibited biased outcomes, particularly under high missing proportions. The impact of test length appeared negligible on both bias and RMSE results. While all methods demonstrated comparable RMSE values

under low missing proportions, GVEM-BS outperformed MH-RM and GVEM as missing proportions increased. Specifically, MH-RM generated considerably larger RMSE values for both a and b parameters, whereas GVEM-BS maintained RMSE values below 0.75, indicating its greater robustness under MCAR conditions. Furthermore, MH-RM displayed better performance in CFA conditions compared to EFA conditions. The high factor correlation interacted with high missing proportions, resulting in improved performance of GVEM-BS and deteriorated performance of MH-RM.

Figures 2.4 and 2.5 illustrate the performance of three methods under MAR conditions. In comparison to MCAR scenarios, all methods exhibited less accuracy in parameter estimation, particularly noticeable with higher missing proportions. Regarding bias, MH-RM provided unbiased results for a parameters at low or medium missing proportions but tended to overestimate b parameters across most conditions. Conversely, both GVEM-BS and GVEM tended to underestimate a parameters and overestimate b parameters, with this pattern becoming more pronounced as missing proportions increased. Notably, two GVEM methods showed better accuracy for b parameters compared to a parameters, especially under conditions of high missing proportions and factor correlations. Additionally, increasing the test length improved model performance for both methods. GVEM-BS yielded larger RMSE values for a parameters under conditions of medium missing proportions and high factor correlations, but it outperformed MH-RM and GVEM for both a and b parameters with high missing proportions. As missing proportions increased, MH-RM produced larger RMSE values for a parameters compared to b parameters, and this method tended to fail to converge with high missing proportions. Table 2.3 also summarizes the number of unsuccessful replications for MH-RM. Overall, GVEM-BS exhibited more consistent patterns of change compared to MH-RM, suggesting its greater robustness in handling missing data. However, GVEM-BS yielded smaller RMSE values for a parameters in EFA analysis compared to CFA analysis. Consistent with MCAR findings, GVEM-BS performed better under $r = 0.7$ conditions than $r = 0.1$ conditions. The test length appeared to have negligible effects on parameter estimation accuracy.

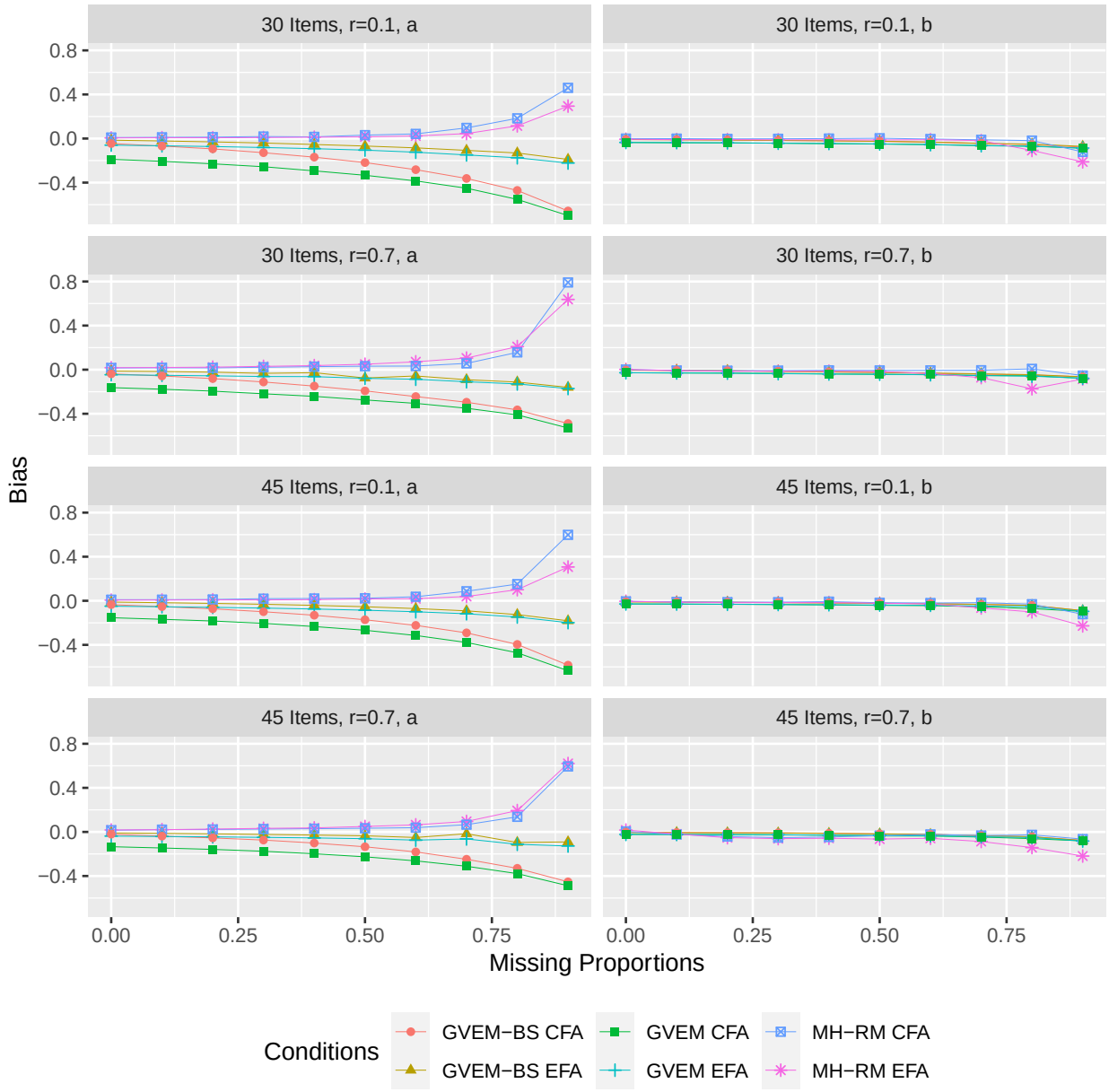

Figure 2.2: Bias values across the different simulation conditions when missing data is MCAR

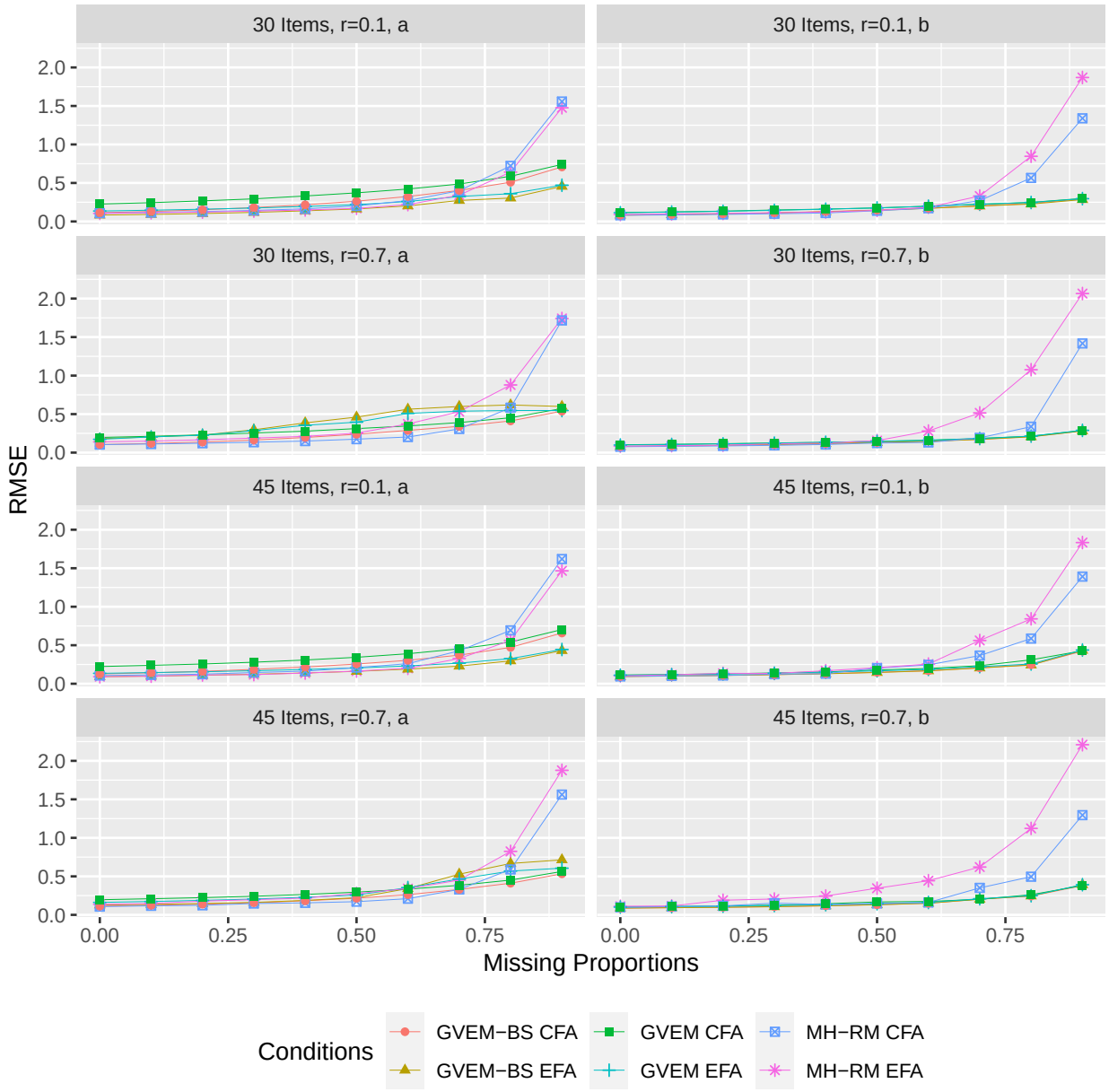

Figure 2.3: RMSE values across the different simulation conditions when missing data is MCAR

2.4 Simulation Study 2

2.4.1 Simulation Design

The second simulation study aimed to explore the impact of assessment design. CFA was implemented since the item loading structure was often known in large-scale assessments.

Table 2.2: The number of unsuccessful replications for MH-RM under MCAR conditions.

Factor Correlation	Analysis Type	Test Length	Missing Proportions	Replications
0.1	EFA	30	0.9	1
0.1	EFA	45	0.9	18
0.1	EFA	45	0.8	4
0.1	EFA	45	0.7	3
0.7	EFA	30	0.9	2
0.7	EFA	45	0.9	15
0.7	EFA	45	0.8	1
0.7	EFA	45	0.7	1
0.1	CFA	30	0.9	1
0.1	CFA	45	0.9	18
0.1	CFA	45	0.8	3
0.1	CFA	45	0.7	3
0.1	CFA	45	0.4	1
0.7	CFA	30	0.9	2
0.7	CFA	45	0.9	15
0.7	CFA	45	0.8	3
0.7	CFA	45	0.7	5
0.7	CFA	45	0.6	2

True item parameters were derived from a response dataset obtained from a NAEP MST Grade 8 math assessment conducted in 2011. This dataset, utilized in subsequent empirical analyses, comprises 74 items covering 5 math domains, with a total sample size of 8,401 students. Among these, approximately 40% ($N = 3,344$) were part of the experiment sample, undergoing the two-stage MST (similar to Figure 2.1), while the remaining 60% ($N =$

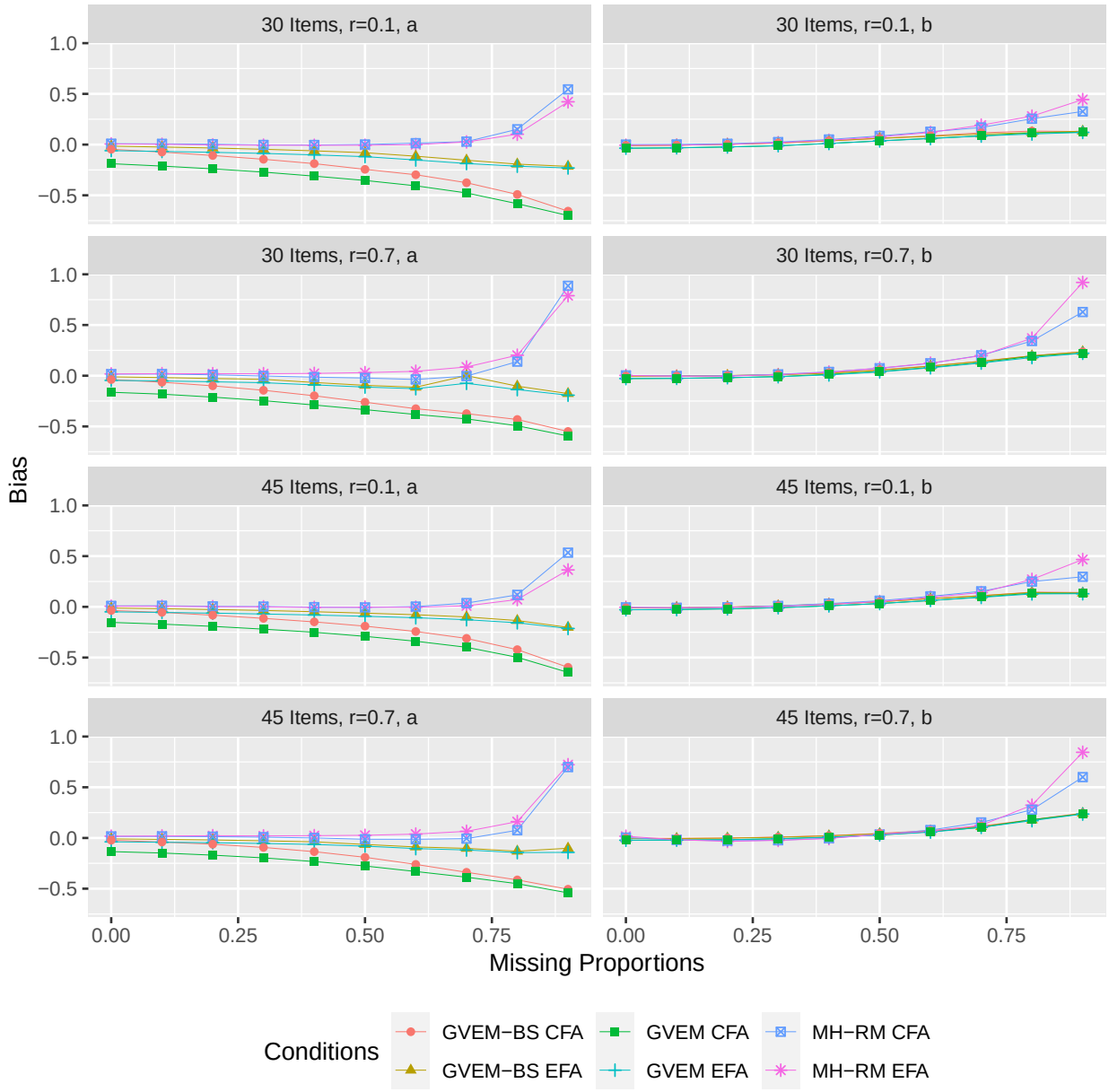

Figure 2.4: Bias values across the different simulation conditions when missing data is MAR

5, 057) constituted the calibration sample undergoing random routing. Since the original item parameters rely on unidimensional IRT models, the parameters were recalibrated using between-item M2PL model, and the resulting factor correlations were considered as the true

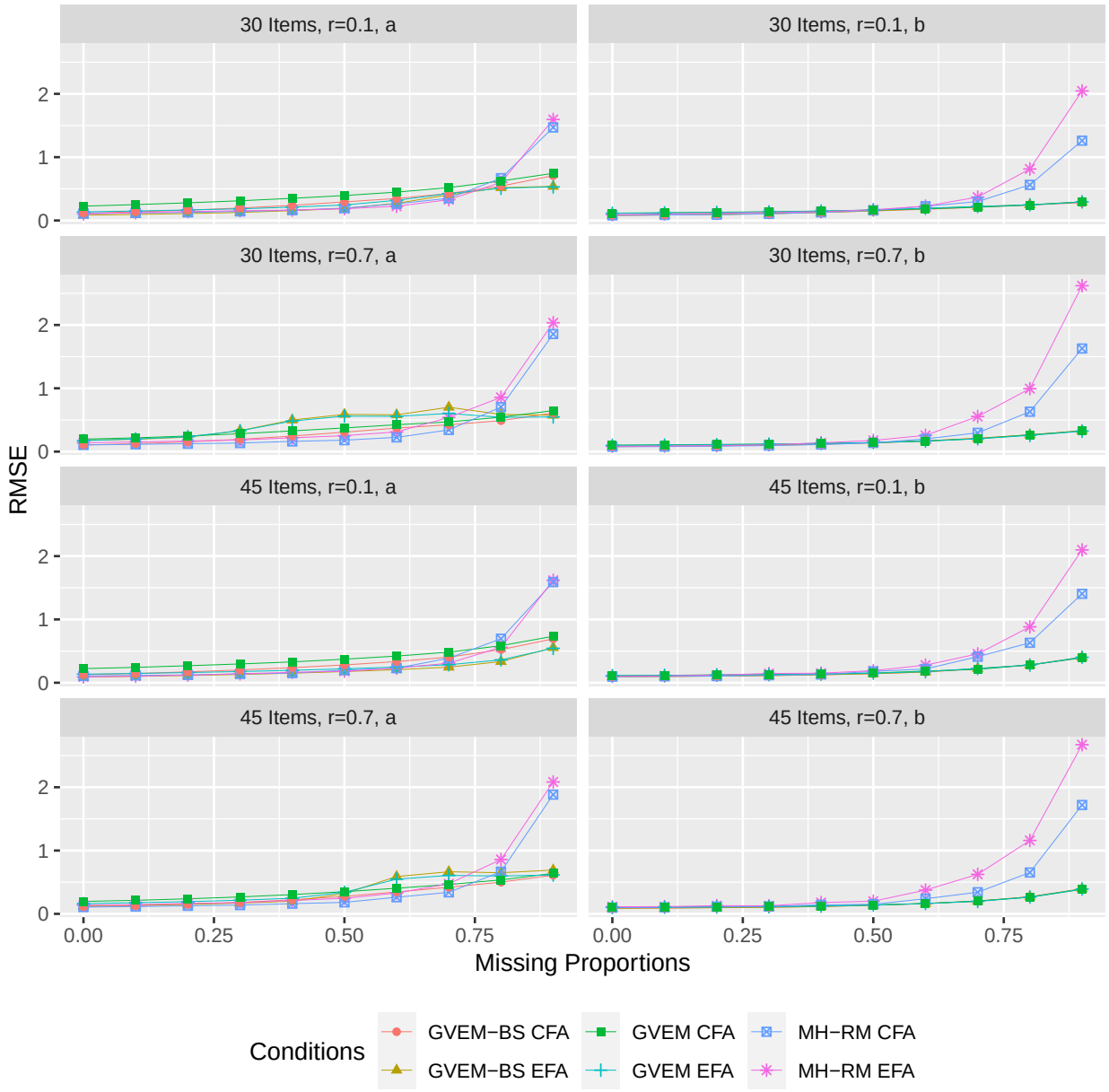

Figure 2.5: RMSE values across the different simulation conditions when missing data is MAR

item parameters. Table 2.4 provides details regarding the number of items per content domain and the mean values of true item parameters based on the MIRT model. The factor

Table 2.3: The number of unsuccessful replications for MH-RM under MAR conditions.

Factor Correlation	Analysis Type	Test Length	Missing Proportions	Replications
0.1	EFA	30	0.9	1
0.1	EFA	45	0.9	12
0.1	EFA	45	0.8	2
0.7	EFA	45	0.9	6
0.7	EFA	45	0.8	3
0.1	CFA	30	0.9	1
0.1	CFA	45	0.9	12
0.1	CFA	45	0.8	3
0.7	CFA	45	0.9	6

correlations ranged from 0.82 to 0.91.

Table 2.4: Descriptive statistics of items per domain.

Domain	#Items	Mean(a)	Mean(b)
Number properties and operations	14	1.15	-0.79
Measurement	12	1.36	-0.57
Geometry	15	1.12	-0.69
Data analysis and probability	9	1.02	-0.38
Algebra	24	1.30	-0.37

In the current study, two commonly used assessment designs, MST and BIB, were examined. The MST simulation design was similar to C. Wang et al. (2020). In the routing stage, two parallel forms were administered, with examinees randomly assigned to one of

the two forms, resulting in missing data that is MCAR, with missing proportions per item at approximately 50%. Subsequently, the remaining items were categorized based on their difficulty parameter into three equivalent sets to form easy, moderate, and difficult modules, respectively. Following the routing stage, each examinee's score was estimated using the expected a posteriori (EAP) method and ordered, after which approximately one-third of the examinees were routed to one of the three modules in the second stage. Consequently, the missing data in the second stage was MAR, with missing proportions per item at approximately 67%. Table 2.5 provides details regarding the number of items per form and per module. The sample size was fixed as 4200 and the missing proportions in the entire dataset were about 59%.

Table 2.5: Number of items per domain and per form.

Domain	Routing		Target		
	Form 1	Form 2	Easy	Medium	Difficult
Number properties and operations	3	4	2	2	3
Measurement	3	3	2	2	2
Geometry	3	3	3	3	3
Data analysis and probability	2	2	2	1	2
Algebra	6	5	5	5	3
Total number of items	17	17	14	13	13

As for BIB design, we employed an automated test assembly (ATA) technique using xxIRT package (Luo & Luo, 2017) in R to generate 7 parallel blocks based on the item bank. The ATA process adhered to specific requirement: firstly, each block consisted of 10 or 11 items; secondly, each block maximized the test information across ability values ranging from -1 to 1; and thirdly, each block included at least one item from the data analysis domain. Based on the criterion shown in Table 2.1, 7 booklets were generated, each comprising

three blocks. The test information plots for 7 booklets are presented below. Simulees were divided into three groups randomly and completed one booklet. As a result, the missing data mechanism was MCAR, and the missing proportions were about 57%, aligning with the MST design. The sample size was also fixed as 4200.

Same as Study 1, the bootstrap sampler was fixed as 10 for GVEM-BS and the number of replications was 50. Bias and RMSE were compared for each method.

2.4.2 Study 2 Results

Tables 2.6 and 2.7 summarize the item parameter recovery for all items and items within each module. Consistent with previous findings, we found both GVEM methods underestimated a parameters, whereas MH-RM overestimated them. One exception was for the medium module where MH-RM underestimated a parameters. Overall, all three methods performed similarly across each module, with MH-RM yielding nearly unbiased estimates regarding a parameter recovery, followed by GVEM-BS. As for b parameters, MH-RM tended to underestimate them, whereas both GVEM methods tended to overestimate them in most conditions, except for the difficult module. Specifically, all three methods also yielded downward bias results for b parameter recovery in the difficult module. GVEM-BS generally performed better than MH-RM regarding b parameter recovery, except for the medium module. In this case, MH-RM yielded nearly unbiased results. Furthermore, RMSE results revealed that MH-RM generally excelled in item parameter recovery, outperforming both GVEM methods, with the easy module being an exception where GVEM showcased superior accuracy, followed by GVEM-BS and MH-RM. As a result, average RMSE results of all items showed that GVEM-BS performed the best in terms of a parameter recovery, whereas GVEM outperformed regarding b parameter recovery.

Table 2.8 presents average bias and RMSE results of item parameters for each method for BIB design. Similar to the results for the MST design, both GVEM methods tended to underestimate a parameters but overestimate b parameters. In contrast, MH-RM overestimated a parameters but underestimated b parameters. Under the BIB design, when the missing data

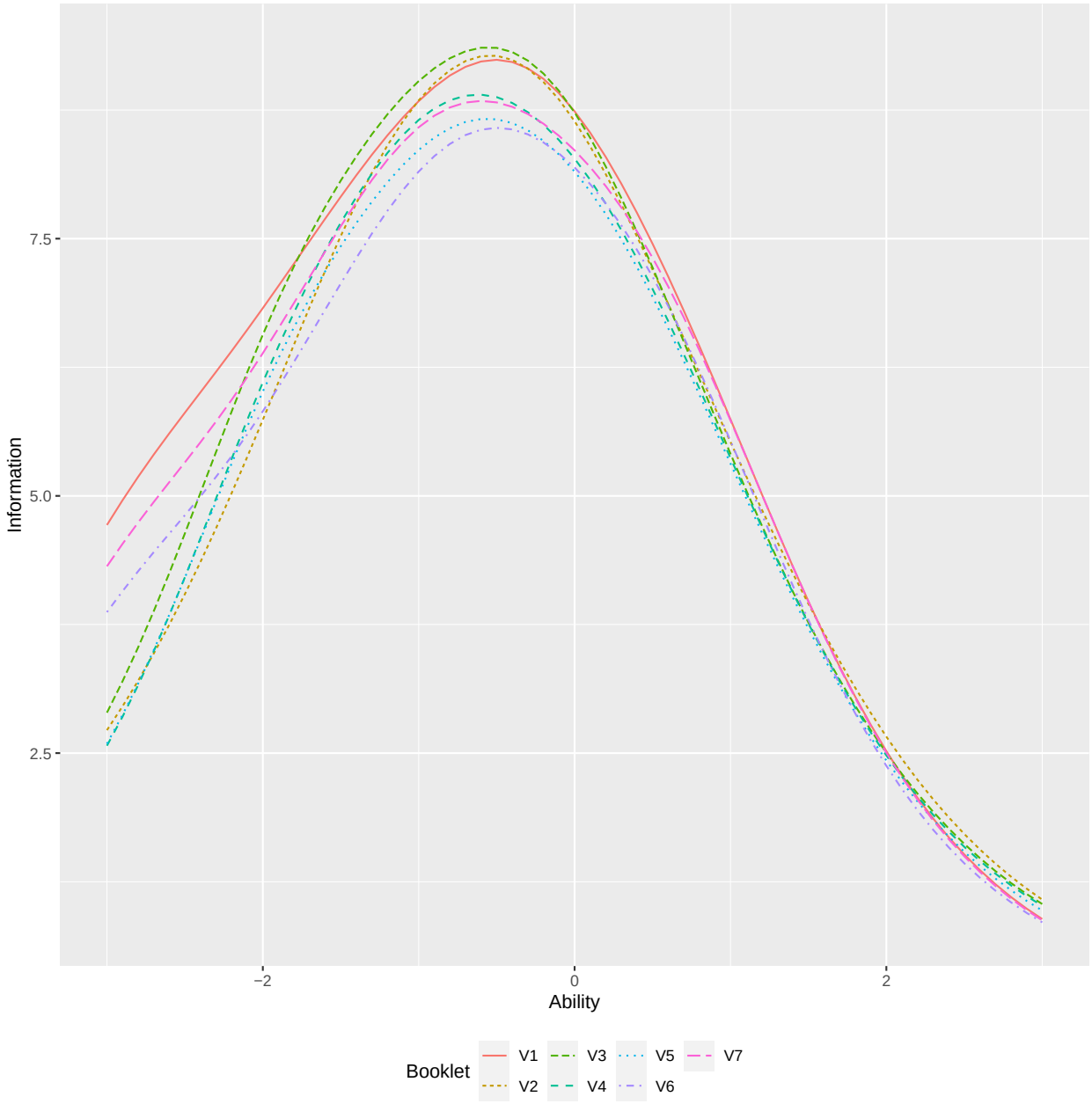

Figure 2.6: Test Information Function Plots for Booklets

were MCAR, MH-RM yielded the smallest bias and RMSE values for a parameter recovery, followed by GVEM-BS and GVEM. In terms of b parameter recovery, GVEM-BS produced

Table 2.6: Average bias results of item parameters for MST design.

Method	All		Routing		Easy		Medium		Difficult	
	a	b	a	b	a	b	a	b	a	b
GVEM	-0.156	0.055	-0.146	0.037	-0.168	0.136	-0.211	0.089	-0.114	-0.020
GVEM-BS	-0.118	0.039	-0.116	0.026	-0.090	0.080	-0.176	0.084	-0.095	-0.018
MH-RM	0.015	-0.036	0.011	-0.031	0.058	-0.086	-0.013	-0.001	0.009	-0.029

Table 2.7: Average RMSE results of item parameters for MST design.

Method	All		Routing		Easy		Medium		Difficult	
	a	b	a	b	a	b	a	b	a	b
GVEM	0.248	0.267	0.178	0.086	0.374	0.566	0.297	0.185	0.148	0.057
GVEM-BS	0.235	0.273	0.149	0.078	0.379	0.582	0.280	0.188	0.128	0.057
MH-RM	0.252	0.337	0.087	0.072	0.491	0.737	0.253	0.193	0.084	0.062

the smallest bias results, while MH-RM and GVEM-BS produced comparable RMSE values.

Table 2.8: Average bias and RMSE results of item parameters for BIB Design.

Method	Bias		RMSE	
	a	b	a	b
GVEM	-0.159	0.048	0.200	0.112
GVEM-BS	-0.123	0.036	0.162	0.095
MH-RM	0.015	-0.047	0.099	0.091

2.5 Real Data Analysis

In this section, both GVEM-BS and MH-RM methods were applied to two response datasets obtained from a NAEP MST Grade 8 math assessment conducted in 2011, as detailed in the previous section. The experiment dataset originated from a two-stage MST, while the calibration dataset was derived from random routing. Consequently, the missing data in the experiment dataset were identified as MAR, while the missing data in the calibration dataset were considered MCAR. The missing proportions for both datasets were approximately 59%. Again, the bootstrap sampler was fixed as 10 for GVEM-BS.

Figure 2.7 presents the scatter plots of the estimated item parameters for the GVEM-BS and MH-RM methods based on two datasets. The black line denotes a benchmark line indicating identical estimated results between the two methods. While both datasets employed the same set of items, the estimated a parameters ranged from approximately 0.2 to 2.2 based on the calibration dataset, with values increasing to the range of 0.5 to 2.5 for the experiment dataset. Notably, the estimation results from both datasets indicated that most items had a parameters falling within the range of 0.8 to 1.5. Furthermore, an important observation was that MH-RM tended to yield larger estimated values compared to GVEM-BS for the same item, consistent with previous simulation studies' findings that GVEM-BS often underestimated a parameters while MH-RM tended to overestimate them. This trend was more apparent for the calibration dataset, representing an MCAR scenario. In contrast, for the experiment dataset, an MAR scenario, GVEM-BS occasionally produced larger estimated a parameters while other times estimating lower values compared to MH-RM. The correlations between GVEM-BS and MH-RM in terms of estimated a values for the calibration and experiment datasets were 0.978 and 0.931, respectively.

In terms of b parameter estimates, the two methods demonstrated more consistent results, particularly for the calibration dataset. The estimated b parameters fell within the range of -4 to 1 for both datasets. Additionally, we observed that both methods consistently estimated the smallest b parameters for the easy block and the largest ones for the

difficult block, which aligns with expectations given that the blocks were defined based on the values of b parameters. Interestingly, we noted that GVEM-BS tended to yield larger estimates compared to MH-RM for items in the easy block. However, some inconsistencies were observed in the estimated values for items in the difficult block based on the experiment sample. The correlations between GVEM-BS and MH-RM in terms of estimated b values for the calibration and experiment datasets were 0.999 and 0.991, respectively, indicating a high degree of agreement between the two methods.

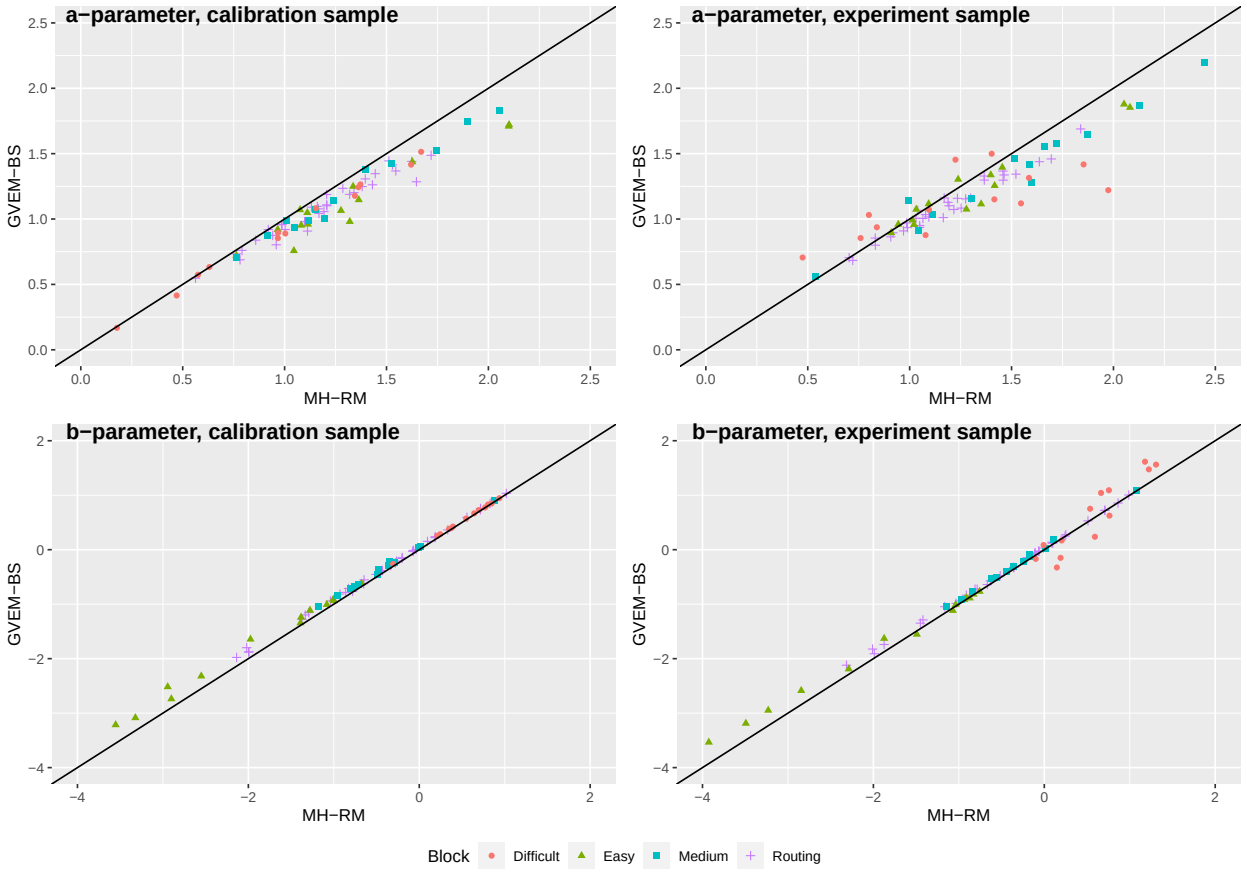

Figure 2.7: Scatter plots of the estimated item parameters from two methods based on calibration and experiment datasets

2.6 Discussion

While the parameter estimation in MIRT models is challenging, handling missing data in the meantime puts higher request on the robustness of the estimation algorithms. While previous research has demonstrated GVEM's superiority over MH-RM in item parameter recovery under various exploratory factor analysis conditions (Cho et al., 2021; Cho et al., 2022), the comparison remains incomplete without considering datasets with missing values. Addressing missing data is crucial for accurate competency scaling, as it directly influences inferences about examinee competencies (Mislevy & Wu, 1996; Ulitzsch, 2020). Thus, the current study proposed a modified GVEM by adding the bootstrap bias correction step and denoted it as GVEM-BS. We compared its estimation precision to original GVEM and MH-RM algorithms, under different missing data scenarios.

GVEM-BS showed promising results in handling missing data, exhibiting comparable or superior accuracy to MH-RM and traditional GVEM algorithms. Incorporating bootstrap bias correction improved GVEM's accuracy and robustness, as indicated by both simulation and empirical results. Simulation study 1 highlighted GVEM-BS's superior performance under conditions of high factor correlation, while MH-RM's accuracy declined. This underscores the importance of considering the interplay between model specifications and missing data characteristics. However, MH-RM outperformed in simulation study 2, which examined different assessment designs, albeit with a notable limitation of unsuccessful convergence under high missing data conditions. This highlights the need for further investigation into MH-RM's convergence properties when handling missing data.

The current study can be extended in several directions. Firstly, the performance of GVEM-BS could be further evaluated under a more complex missing data mechanism, missing not at random (MNAR). In LSAs, MNAR instances like omitted and not-reached items are common, posing challenges to scaling and item calibration (Ulitzsch, 2020). Evaluating the robustness of GVEM-BS in such scenarios would offer insights into its real-world applicability, where missing data patterns may be more intricate.

Secondly, to enhance GVEM methods' performance in MAR scenarios with high missing proportions, employing multiple imputation to impute missing values prior to parameter estimation could be beneficial. Additionally, exploring alternative modifications to the GVEM algorithm to improve the performance tailored to address specific challenges posed by missing data may yield further improvements. For instance, previous research found applying importance weighted variational inference technique to GVEM could correct the bias in a parameter estimates (Ma et al., 2023). Future studies could investigate its performance under missing data scenarios to assess its effectiveness in mitigating bias and improving estimation accuracy.

Thirdly, the study could be expanded to investigate the impact of different model specifications on the performance of GVEM-BS and MH-RM under missing data scenarios. By varying multidimensional structures (between-item M2PL or within-item M2PL), the number of dimensions (low or high), or response types (dichotomous or polytomous), researchers can gain a better understanding of how these factors interact with missing data characteristics and influence parameter estimation accuracy. This comprehensive analysis would provide insights into the conditions under which each algorithm performs optimally and guide researchers in selecting the most appropriate method for their specific modeling needs.

Furthermore, the current study only considers parametric bootstrap sampling. In the future, non-parametric bootstrap techniques could be taken into account. Non-parametric bootstrap methods do not assume a specific distribution for the data, making them more flexible and potentially more robust in various practical scenarios (Coskun et al., 2013). Exploring non-parametric bootstrap could provide additional robustness, particularly in complex or less well-understood missing data contexts. This extension would further enhance the applicability and reliability of the GVEM-based methods in a broader range of educational assessment settings.

Chapter 3

STANDARD ERRORS FOR GAUSSIAN VARIATIONAL ESTIMATION IN MIRT MODELS

For multidimensional item response theory (MIRT), accurate estimates of item parameters, as well as their standard errors (SEs), or more generally error variance–covariance matrices, are important prerequisites for many applications (P. Chen & Wang, 2021; C. Wang & Zhang, 2019), including but not limited to multidimensional computerized adaptive testing (Yao, 2012), item parameter calibration (P. Chen, 2017), limited-information goodness-of-fit testing (Cai et al., 2006), as well as differential item functioning (Woods et al., 2013). Several methods have been proposed for the item parameter estimates under MIRT models, and the Gaussian Variational Expectation Maximization (GVEM) as an alternative further improves the computational efficiency and estimation accuracy (Cho et al., 2021). However, the GVEM framework does not produce the standard error (SE) estimation procedure. Moreover, studies related to the SEs estimation procedure are scarce (Paek & Cai, 2014), especially for the MIRT models. One reason is that the EM algorithm is a first-order iterative scheme, which does not inherently provide second-order derivative matrices or error variance–covariance matrices as an automatic byproduct (McLachlan & Krishnan, 2007). Since the point estimation of item parameters only tells one side of the story and SEs are key elements for statistical inference, the current study complements the work by Cho et al. (2021) by focusing on the SE estimation procedure under the GVEM framework. Two estimation methods are proposed: USEM and bootstrap. Their performance is evaluated via simulation studies.

The rest of the chapter is organized as follows. First, we discuss the importance of SE estimates. Next, the existing methods for SE estimates are reviewed. Then, two methods for

the SEs of item parameters under the GVEM framework are introduced: the updated version of supplemented EM (USEM; Tian et al., 2013) and the bootstrap, denoted as GVEM-USEM and GVEM-BS, respectively. Simulation studies are presented to illustrate the performance of two proposed procedures, followed by a discussion of the results summary.

3.1 Importance of SE Estimation

According to its definition, the standard error (SE) represents the standard deviation of a sampling distribution of an estimator. In the context of item parameter estimation under the MIRT framework, SE serves to quantify the accuracy of a point estimate of an item parameter. A smaller SE value signifies a more precise estimation of the item parameter. Furthermore, SE plays a pivotal role in constructing the confidence interval for item parameter estimation.

The SE estimation procedure is important because the uncertainty of item parameter estimation can be carried over in subsequent analysis of operational psychometric tasks (Lin, 2018; C. Wang & Zhang, 2019). For example, test scoring estimates an individual's latent trait(s) or ability parameter(s) when item parameters are fixed. The uncertainties of item parameter estimates transfer to the person parameter estimation implicitly. The SEs of person parameter estimates can later serve as one of the most commonly used stopping rules in computerized adaptive testing (Magis et al., 2017). Another applied scenario is the fixed-item parameter calibration. Accurate estimation of the fixed (common) items' parameters and their standard errors is crucial, as it enables precise estimation of the item parameters and their standard errors for new items. Furthermore, it allows us to quantify the influence of the parameter estimation uncertainty in the fixed items on the calibration accuracy of the new items (P. Chen & Wang, 2021).

In addition to test scoring and item parameter calibration, estimating SEs is a prerequisite for calculating an asymptotic distribution for limited-information goodness-of-fit test statistics. This type of test proves more powerful than full-information tests (e.g., Pearson's statistic) to evaluate the model fit, as it only relies on lower-order marginal probabilities

(Cai & Hansen, 2013). The approach has found application across various item response theory (IRT) models, including hierarchical multidimensional item response theory (MIRT) models (Cai, Yang, et al., 2011), item bifactor models (Cai & Hansen, 2013), and two-tier models(Cai, 2010c). Lastly, SE estimation is crucial for differential item functioning (DIF) analysis, a commonly studied topic in the IRT literature. One of the widely used DIF methods is Wald's χ^2 test (Wald, 1941). When employing a Wald test to detect DIF in a specific item, the parameters of all other items are typically constrained to be equal between focal and reference groups, while the parameters of the item under scrutiny are freely estimated. The comparison of the difference in parameter estimates between the two groups relies on the information of standard errors. Therefore, the accuracy of SE estimation significantly impacts the performance of the Wald test.

3.2 Existing SE Estimation Techniques

In this section, an overview of SE estimation procedures in the IRT literature is given. The details of USEM and bootstrap methods under the GVEM framework will be introduced in subsequent sections.

A gold standard approach to estimating SE is to take the square root of the diagonal entries of expected Fisher's information matrix (FIS). The FIS approach computes the information matrix using the negative expectation of the Hessian matrix, i.e., the second-order derivatives of the log-likelihood function:

$$\mathbf{I}_{\text{FIS}} = -E(H(\mathbf{M})) = -\frac{\partial^2}{\partial \mathbf{M} \partial \mathbf{M}^T} l(\mathbf{M}; \mathbf{Y}) \tag{3.1}$$

The square roots of the diagonal elements of $\mathbf{I}_{\text{FIS}}^{-1}$ are SE estimates for model parameters. While the FIS is considered a golden approach, computing the Hessian matrix can be computationally demanding. The computation of expected Fisher information matrix involves a summation loop over the total number of possible response patterns, which increases exponentially as the test length increases (Paek & Cai, 2014; C. Wang & Zhang, 2019). This exponential increase in computation cost poses a significant challenge, particularly for tests

with a large number of items or response categories. Despite its time-consuming nature, the expected Fisher information approach yields the most accurate standard errors because it directly aligns with the definition of Fisher information (Lin, 2018). Some IRT software programs, such as IRTPRO (Cai, Du Toit, et al., 2011) and flexMIRT (Cai, 2013), offer this method as an option for standard error estimation.

Alternatively, the cross-product (XPD) approach offers a more computationally efficient method to express the information matrix, utilizing the cross-product of gradients:

$$\mathbf{I}_{\text{XPD}} = E[\frac{\partial l(\mathbf{M};\mathbf{Y})}{\partial \mathbf{M}} \cdot \frac{\partial l(\mathbf{M};\mathbf{Y})}{\partial \mathbf{M}^T}], \tag{3.2}$$

where $\frac{\partial l(\mathbf{M};\mathbf{Y})}{\partial \mathbf{M}}$ denotes the gradient of the log-likelihood function. Equation (3.2) can be approximated using the sample cross-product of gradients, which relies solely on the observed response patterns in the data, making it much easier to compute. Furthermore, when employing the EM algorithm for estimation, empirical XPD is considerably more tractable than the observed information matrix (Chalmers et al., 2017). This is attributed to equation (3.2), which exclusively involves the first derivatives of the log-likelihood function, whereas Fisher's equation (3.1) entails the computation of the Hessian matrix. In practice, empirical XPD also has shown superior computational efficiency compared to other methods (Paek & Cai, 2014). Consequently, most IRT software applications offer XPD as an option for producing standard errors. Note that Equations (3.1) and (3.2) are asymptotically equivalent under correct model specification (Chalmers et al., 2017). However, under misspecification, neither of the two methods can provide consistent SE estimates (Falk & Monroe, 2018; White, 1982). Furthermore, XPD could exhibit some upward bias when the test length is long and the sample size is relatively small (Paek & Cai, 2014).

Except for the FIS and XPD approaches, prior research suggests a robust SE estimation using a sandwich-type variance-covariance matrix, which can be considered as a weighted combination of the observed information and cross-product (Yuan et al., 2014). The variance-covariance matrix $\mathbf{V}_{\text{SW}}$ can be expressed as follows:

$$\mathbf{V}_{\text{SW}} = \mathbf{I}_{\text{FIS}}^{-1}\mathbf{I}_{\text{XPD}}\mathbf{I}_{\text{FIS}}^{-1} \tag{3.3}$$

SEs derived from the sandwich-type variance-covariancee matrix have been integrated into structural equation modeling (SEM) software for both continuous data, utilizing the asymptotic covariance matrix of sample covariances, and ordinal data, employing the asymptotic covariance matrix of polychoric correlations (Bentler, 1995; Muthén & Muthén, 2009; Yuan et al., 2014). Calculating the expected Fisher information matrix, necessary for constructing the sandwich covariance matrix, poses a challenge, yet one can opt to compute the observed version as an alternative.

Additionally, several alternative methods are available for obtaining the error variance-covariancee matrix when using the EM algorithm. These methods all involve numerically differentiating what is known as the EM map function during the process. Let $\mathbf{V_M}$ denote the $p \times p$ dimensional error variance-covariance matrix for model parameters $\mathbf{M}$, which can be computed as the inverse of the observed-data information $\mathbf{I}_o$. According to the missing information principle (Orchard & Woodbury, 1972), the observed-data information is the difference between the complete-data information $\mathbf{I}_c$ and the missing-data information $\mathbf{I}_m$. Therefore, $\mathbf{V_M}$ can be derived as

$$\mathbf{V_M} = \mathbf{I}_o^{-1} = (\mathbf{I}_c - \mathbf{I}_m)^{-1} = \mathbf{I}_c^{-1}(\mathbf{I}_p - \mathbf{I}_m\mathbf{I}_c^{-1})^{-1} = \mathbf{I}_c^{-1}(\mathbf{I}_p - \mathbf{\Delta})^{-1}, \tag{3.4}$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix, $\mathbf{I}_c$ is the M-step information matrix as a byproduct in each EM iteration, $\mathbf{\Delta} = \mathbf{I}_m\mathbf{I}_c^{-1}$ is the fraction of missing information. The key part is to compute $\mathbf{\Delta}$. Cai (2008) indicated $\mathbf{\Delta}$ governs the rate of convergence of the EM process and can be computed as

$$\mathbf{M}^{(t+1)} - \hat{\mathbf{M}} \approx \mathbf{\Delta}(\mathbf{M}^{(t)} - \hat{\mathbf{M}}), \tag{3.5}$$

where $\hat{\mathbf{M}}$ is the maximum likelihood estimate (MLE) of $\mathbf{M}$, $\mathbf{\Delta} = \frac{\partial \mathcal{F}(\hat{\mathbf{M}})}{\partial \mathbf{M}}\Big|_{\mathbf{M}=\hat{\mathbf{M}}} = (\Delta_{ij})_{p\times p}$ is the Jacobian matrix evaluated at $\hat{\mathbf{M}}$, and $\mathcal{F}$ is a vector-valued mapping function which is referred to the EM map in the literature: $\mathbf{M}^{(t+1)} = \mathcal{F}(\mathbf{M}^{(t)})$ and $\hat{\mathbf{M}} = \mathcal{F}(\hat{\mathbf{M}})$.

The elements of $\mathbf{\Delta}$ can be obtained using numerical differentiation methods. One approach is based on the forward difference method (FDM) proposed by Jamshidian and Jen-

nrich (2000), which is among the common finite difference approximations used for NDM. Specifically, the FDM adds a small positive value $e = \eta \max(|\hat{\mathbf{M}}, 1|)$ to the MLE and keeps the rest of the estimated values unchanged, where η is a small perturbation constant. Let $\mathbf{v}_i$ be a p-dimensional vector with all zeros except for the $i$the element equal to 1. The resulting vector of this operation can be expressed as $\hat{\mathbf{M}} + e\mathbf{v}_i = (\hat{M}_1, \ldots, \hat{M}_{i-1}, \hat{M}_i + e, \hat{M}_{i+1}, \ldots, \hat{M}_p)$. Therefore, the derivative is approximated by using the forward difference as

$$\Delta_{ij} = \frac{\partial \mathcal{F}_j(\hat{\mathbf{M}})}{\partial M_i} = \lim_{\eta \to 0} \frac{\mathcal{F}_j(\hat{\mathbf{M}} + e\mathbf{v}_i) - \mathcal{F}_j(\hat{\mathbf{M}})}{e}. \tag{3.6}$$

Using similar perturbations, the Richardson extrapolation method (REM) approximates the derivative using the first-order Richardson extrapolation of the central difference as

$$\Delta_{ij} = \lim_{\eta \to 0} \frac{\mathcal{F}_j(\hat{\mathbf{M}} - 2e\mathbf{v}_i) - 8\mathcal{F}_j(\hat{\mathbf{M}} - e\mathbf{v}_i) + 8\mathcal{F}_j(\hat{\mathbf{M}} + e\mathbf{v}_i) - \mathcal{F}_j(\hat{\mathbf{M}} + 2e\mathbf{v}_i)}{12e}. \tag{3.7}$$

A notable distinction between the FDM and the REM resides in their computational complexity. REM is approximately four times more intricate, requiring the evaluation of four functions, entailing the evaluation of four distinct functions, while FDM necessitates the assessment of only one. Jamshidian and Jennrich (2000) advocated for REM over FDM in scenarios where computational efficiency is when computational speed is not a concern, as they discovered that despite the slight increase in complexity, REM could offer substantially improved precision.

While FDM or REM are noniterative, they rely on selecting a perturbation constant for use in the finite difference approximation denominator. Unfortunately, this choice can significantly impact the error of numerical differentiation, thereby affecting the accuracy of the final item parameter standard error estimates derived from the numerical derivatives of the EM map (Tian et al., 2013). Instead of choosing a constant perturbation, the supplemented expectation maximization (SEM) approach (Cai, 2008; Meng & Rubin, 1991) reuses the EM iterative history to choose the perturbation adaptively. Specifically, it calculates the elements of $\Delta_i j$ based on

$$\Delta_{ij} = \frac{\partial \mathcal{F}_j(\hat{\mathbf{M}})}{\partial M_i} = \lim_{M_i \to \hat{M}_i} \frac{\mathcal{F}_j(\hat{M}_1, \ldots, \hat{M}_{i-1}, M_i, \hat{M}_{i+1}, \ldots, \hat{M}_p) - \mathcal{F}_j(\hat{\mathbf{M}})}{M_i - \hat{M}_i}$$
$$= \lim_{t \to \infty} \frac{\mathcal{F}_j(M_{(i)}^{(t)}) - \hat{M}_i}{M_i^{(t)} - \hat{M}_i} = \lim_{t \to \infty} \Delta_{ij}^{(t)}. \tag{3.8}$$

where $\mathbf{M}_{(i)}^{(t)} = (\hat{M}_1, \ldots, \hat{M}_{i-1}, M_i^{(t)}, \hat{M}_{i+1}, \ldots, \hat{M}_p)$ equals to $\hat{\mathbf{M}}$ except that the ith element is replaced by $M_i^{(t)}$, which is the estimate of M_i at the tth EM iteration. Compared with other SE techniques, the SEM approach has demonstrated its ease of implementation with the EM algorithm and its robust performance under various conditions (Cai, 2008). This approach has been incorporated into several IRT software programs like IRTPRO (Cai, Du Toit, et al., 2011) and flexMIRT (Cai, 2013). However, previous research found that the accuracy of SEM may be compromised when the EM algorithm converges slowly (Jamshidian & Jennrich, 2000). Considering there is room to improve the computational efficiency and stability of the SEM, Tian et al. (2013) proposed a modified version, the updated SEM (USEM) approach. The current study integrates the USEM with GVEM to estimate SEs of item parameters.

The aforementioned methods necessitate computing the first or second derivative of the likelihood function; however, in some cases, the likelihood is intractable analytically or challenge to obtain. To deal with this issue, the bootstrap method is an alternative for estimating SE, which has been applied in a wide range of statistical models (i.e., Efron & Tibshirani, 1986; Gonçalves & White, 2005). In the educational measurement field, the bootstrap approach is commonly used for estimating SE of IRT equating methods (i.e., Y. Liu et al., 2008; Tsai et al., 2001; Z. Zhang & Zhao, 2019), and providing accurate SEs for person parameter estimates (i.e., Fitzpatrick & Yen, 2001; Liou & Yu, 1991; Patton et al., 2014). Its application to SE estimation of item parameters in MIRT, however, is rarely discussed. While in general, bootstrap is computationally demanding due to resampling, by integrating it into the GVEM framework, we aim to maintain a decent computational cost.

3.3 USEM Approach under the GVEM Framework

In GVEM, $\mathbf{I}_c = -\frac{\partial^2}{\partial \mathbf{M} \partial \mathbf{M}^T} E(\mathbf{A}, \mathbf{B}, \xi)$. Since the GVEM could produce closed-form solutions, the second derivatives with respect to $\boldsymbol{\alpha}_j, b_j$ are

$$\begin{aligned}
\frac{\partial^2}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\alpha}_j^T} E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}) &= \sum_{i=1}^{N} -2\eta(\xi_{i,j})[\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top], \\
\frac{\partial^2}{\partial \boldsymbol{\alpha}_j \partial b_j} E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}) &= \sum_{i=1}^{N} 2\eta(\xi_{i,j}) \boldsymbol{\mu}_i, \\
\frac{\partial^2}{\partial b_j^2} E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}) &= \sum_{i=1}^{N} -2\eta(\xi_{i,j})
\end{aligned} \tag{3.9}$$

Other elements in $\mathbf{I}_c$ are 0. In general, $\mathbf{I}_c$ can be expressed as

$$\begin{bmatrix}
-\frac{\partial^2}{\partial \alpha_1 \partial \alpha_1^T} E(\mathbf{A}, \mathbf{B}, \xi) & -\frac{\partial^2}{\partial \alpha_1 \partial b_1} E(\mathbf{A}, \mathbf{B}, \xi) & \cdots & 0 & 0 \\
-\frac{\partial^2}{\partial \alpha_1 \partial b_1} E(\mathbf{A}, \mathbf{B}, \xi) & -\frac{\partial^2}{\partial b_1^2} E(\mathbf{A}, \mathbf{B}, \xi) & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & -\frac{\partial^2}{\partial \alpha_j \partial \alpha_j^T} E(\mathbf{A}, \mathbf{B}, \xi) & -\frac{\partial^2}{\partial \alpha_j \partial b_j} E(\mathbf{A}, \mathbf{B}, \xi) \\
0 & 0 & \cdots & -\frac{\partial^2}{\partial \alpha_j \partial b_j} E(\mathbf{A}, \mathbf{B}, \xi) & -\frac{\partial^2}{\partial b_j^2} E(\mathbf{A}, \mathbf{B}, \xi)
\end{bmatrix}$$

To obtain $\boldsymbol{\Delta}$, the "forced-EM" process (Meng & Rubin, 1991) is conducted. Specifically, run one iteration of GVEM code to get $\mathcal{F}(\mathbf{M}^{(t)})$ for some t, and use it to calculate $\Delta_{ij}^{(t)}$ for all $i, j = 1, 2, \ldots, p$ The step is repeated for $t = 1, 2, \ldots, T^*$until the entire matrix converges (i.e.,$|\Delta_{ij}^{(T^*)} - \Delta_{ij}^{(T^*-1)}| \leq \epsilon$ for all of the elements, where $T^* \leq T$, T is the number of iterations for GVEM). The major difference between the USEM and SEM is that the USEM adopts a row-wise convergence criterion (Tian et al., 2013). That is, for a given i, if $|\Delta_{ij}^{(T^*)} - \Delta_{ij}^{(T^*-1)}| \leq \epsilon$ holds for all $j = 1, 2, \ldots, p$, then the i-th row is considered converged and is no longer involved in the "forced-EM" process. By doing so, USEM converges faster compared to SEM.

3.4 Bootstrap Approach under the GVEM Framework

The bootstrap is an alternative approach to SEs estimates of item parameters when the SEs computation is mathematically intractable (Efron & Tibshirani, 1986). It is a resampling procedure that generates a large number of bootstrap samples in either a parametric fashion or a nonparametric fashion (Z. Zhang & Zhao, 2019). From each bootstrap sample, the statistic of interest is calculated. The sampling distribution of the replications could be obtained to make some inferences related to the accuracy of the statistic (Patton et al., 2014). The current study implemented parametric bootstrap sampling under the GVEM framework (GVEM-BS) and the steps are outlined below.

1. B bootstrap datasets are simulated based on GVEM estimates $\hat{\mathbf{M}}$.

2. The GVEM is conducted to estimate item parameters for each bootstrap dataset, and these item parameter estimates are denoted as $\hat{\mathbf{M}}^1, \hat{\mathbf{M}}^2, \ldots \hat{\mathbf{M}}^B$.

3. The SE estimates are simply the sample standard deviations of the estimated item parameters. The SE estimate for any item parameter M_j (such as α_{jk} or b_j) could be expressed as

$$\sigma_{\hat{M}_j} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{M}_j^b - \hat{M}_j)^2}. \tag{3.10}$$

While the number of bootstrap samples B is usually taken relatively large, our current study finds that $B = 50$ provides good and stable numerical results, and our pilot study results indicated even using as few as 5 bootstrap samples could produce similar estimates. Note if $\hat{\mathbf{M}}$ is obtained by GVEM with prior information, then in the bootstrap samplers, the GVEM would utilize the same prior distributions.

Besides, the current study modifies the expectation of the variational density function by adding prior beliefs on item parameters. That is, we can assume some prior distributions of $\boldsymbol{\alpha}_j \sim N(\boldsymbol{\mu}_{\boldsymbol{\alpha}}^{(0)}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{(0)})$ and $b_j \sim N(\mu_b^{(0)}, \sigma_b^{2(0)})$, and the expectation could be rewritten as

$$
\begin{aligned}
E^{(t)}(\mathbf{M}, \boldsymbol{\xi}) \quad = \quad & \sum_{i=1}^{N}\sum_{j=1}^{J}\Bigg(\log \frac{e^{\xi_{i,j}^{(t)}}}{(1+e^{\xi_{i,j}^{(t)}})} + (\frac{1}{2} - Y_{ij}) b_j^{(t)} + (Y_{ij} - \frac{1}{2}) \boldsymbol{\alpha}_j^{(t)\top} \boldsymbol{\mu}_i^{(t)} - \frac{1}{2}\xi_{i,j}^{(t)} \\
& -\eta(\xi_{i,j}^{(t)}) \{ b_j^{(t)2} - 2 b_j^{(t)} \boldsymbol{\alpha}_j^{(t)\top} \boldsymbol{\mu}_i^{(t)} + \boldsymbol{\alpha}_j^{(t)\top} [\boldsymbol{\Sigma}_i^{(t)} + (\boldsymbol{\mu}_i^{(t)})(\boldsymbol{\mu}_i^{(t)})^\top] \boldsymbol{\alpha}_j^{(t)} - \xi_{i,j}^{(t)2} \} \Bigg) \\
& + \frac{N}{2} \log |\Sigma_{\boldsymbol{\theta}}^{(t)-1}| - \sum_{i=1}^{N} \frac{1}{2} Tr(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(t)-1} [\boldsymbol{\Sigma}_i^{(t)} + (\boldsymbol{\mu}_i^{(t)})(\boldsymbol{\mu}_i^{(t)})^\top]) \\
& - N \sum_{j=1}^{J} \Bigg(\frac{(\boldsymbol{\alpha}_j^{(t)} - \boldsymbol{\mu}_{\boldsymbol{\alpha}}^{(0)})^\top (\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{(0)-1} (\boldsymbol{\alpha}_j^{(t)} - \boldsymbol{\mu}_{\boldsymbol{\alpha}}^{(0)})}{2} - \frac{(b_j^{(t)} - \mu_b^{(0)})^2}{2\sigma_b^{(0)2}} \Bigg). \qquad (3.11)
\end{aligned}
$$

This is similar to the Bayes modal estimation approach presented by Tierney and Kadane (1986). Imposing these priors could prevent deviant parameter estimates and help the algorithm to produce more accurate and stable parameter estimates (Cho et al., 2022). The current study applied the bootstrap approach to the SEs estimation procedure and denoted this version with prior information as GVEM with Bootstrap Sampling and Prior (GVEM-BSP). Note if $\hat{\mathbf{M}}$ was obtained by GVEM with prior information, then in the bootstrap samplers, the GVEM would utilize the same prior distributions.

3.5 Simulation Study

A simulation study was conducted to evaluate the performance of three proposed SE estimation procedures (GVEM-USEM, GVEM-BS, GVEM-BSP). The manipulated factors included: (1) the test length was fixed at 45 or 30; (2) the factor correlations were simulated from $Unif(0.1, 0.3)$ or $Unif(0.5, 0.7)$ to represent the low and high correlation conditions respectively; (3) the multidimensional structure was either between-item M2PL or within-item M2PL; the number of factors was either 3 or 5 (i.e., $K = 3, 5$), representing either low or high dimensionality. In the between-item M2PL, with a test length of 45, 15 items were loaded onto each factor for $K = 3$, and 9 items were loaded onto each factor for $K = 5$. Similarly, at a test length of 30, 10 items were loaded per factor for $K = 3$ and 6 items for $K = 5$. For the within-item M2PL, when the test length was 45 and $K = 3$, about 60%,

24% and 16% items were loaded onto one, two, and three factors respectively. For $K = 5$, the proportions of items loaded onto one, two, and three factors were about 56%, 22% and 22%, respectively. When the test length was 30 and $K = 3$, there were about 60%, 20%, and 20% were loaded onto one, two, and three factors respectively. For $K = 5$, about 33% items loaded onto one, two, and three factors, respectively. Under all conditions, true item discrimination parameters were drawn from $Unif(1, 2)$ and true item difficulty parameters were drawn from $N(0, 1)$, which is the same settings as Cho et al. (2021). The sample size was fixed at 1000. The ability parameters $\boldsymbol{\theta}_i$ were drawn from a multivariate normal distribution, $N \sim (\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is a variance-covariance matrix whose diagonal elements were 1 and the off-diagonal elements depended on the factor correlation conditions.

As alluded to above, both GVEM-BS and GVEM-BSP generated 50 bootstrap samples per replication. For GVEM-BSP, the prior distribution for $\boldsymbol{\alpha}_j$ was set as $N \sim (\mathbf{1.5}, \boldsymbol{I}_3)$, which was informative since the mean value of the true distribution was 1.5. On the other hand, the prior distribution for b_j was $N \sim (0, 10)$, following the setting from Sinharay (2005). Fifty replications were conducted for each condition. The true item parameters held constant across all replications within each condition. The empirical standard deviations of the estimated item parameters across replications per condition served as "true" SEs for each method. Note that the current design was based on a confirmatory model, hence there were many $\boldsymbol{\alpha}_j$ known as $\mathbf{0}$. Therefore, the evaluation criterion below excluded these parameters and focused on non-zero model parameters. Given a non-zero item parameter β_j (which is used as a general notation to represent α_{jk} or b_j), the empirical standard deviation σ_{β_j} can be computed as

$$\sigma_{\beta_j} = \frac{1}{R-1} \sum_{r=1}^{R} (\hat{\beta}_j^r - \beta_j)^2, \tag{3.12}$$

where $\hat{\beta}_j^r$ is the rth replication of item parameter estimate for β_j and R is the total number of replications. For the overall comparisons of different SE estimation procedures, bias, relative bias, and estimated SEs were computed for each non-zero model parameter across all items

within a condition. For each replication, they can be defined by

$$\text{Bias} = \frac{1}{J^*}\sum_{j=1}^{J^*}(\hat{\sigma}^r_{\beta_j} - \sigma_{\beta_j}), \tag{3.13}$$

$$\text{Relative Bias} = \frac{1}{J^*}\sum_{j=1}^{J^*}\frac{\hat{\sigma}^r_{\beta_j} - \sigma_{\beta_j}}{\sigma_{\beta_j}}, \tag{3.14}$$

$$\text{Estimated SEs} = \frac{1}{J^*}\sum_{j=1}^{J^*}\hat{\sigma}^r_{\beta_j}, \tag{3.15}$$

where J^* denotes the number of non-zero model parameters, and $\hat{\sigma}^r_{\beta_j}$ denotes the rth replication of estimated SE for β_j.

3.6 Results

It should be noted that the GVEM-USEM algorithm failed to estimate some SEs of item parameters under some replications and these values were excluded for further analysis. Table 3.1 summarizes the number of unsuccessful replications with missing values for the GVEM-USEM method. The simulation results under various conditions are presented by boxplots to show the distribution of bias, relative bias, and estimated SEs. All three methods (GVEM-BSP, GVEM-BS, GVEM-USEM) were consistently presented in the same order under each manipulated condition. Figures 3.1 and 3.2 illustrate bias and relative bias results for $K = 3$. In low-dimensional settings, the GVEM-USEM yielded nearly unbiased results for SE estimates of item discrimination parameters under the simplest condition ($K = 3, J = 45$, low factor correlations and between-item structure). However, it tended to overestimate SEs for item difficulty parameters. Both bootstrap methods, GVEM-BSP and GVEM-BS, slightly underestimated the SEs for item parameters. As conditions became more complex, GVEM-BSP and GVEM-BS methods performed better by producing close to 0 bias and relative bias results for SE estimates. Particularly, GVEM-BSP provided approximately unbiased SE estimates under most conditions. In contrast, the GVEM-USEM method consistently

overestimated SEs of item parameters. Under the most complex condition (i.e., short test length, high latent correlations, and within-item structure), both GVEM-BS and GVEM-USEM methods overestimated item discrimination parameters and yielded some outliers. In contrast, the GVEM-BSP method yielded unbiased results for SE estimates of item discrimination parameters in terms of bias and relative bias results, but slightly underestimated SEs for item difficulty parameters. The bias and relative bias varied more for the within-item conditions than the between-item conditions. This was expected since the within-item model had a more complex loading structure.

Table 3.1: The number of unsuccessful replications for the GVEM-USEM.

Test Length	Factor Correlation	Model Structure	Number of Factors	Replications
45	0.1-0.3	Within-Item	3	1
45	0.5-0.7	Between-Item	3	3
45	0.5-0.7	Within-Item	3	4
30	0.5-0.7	Within-Item	3	1

Figures 3.3 and 3.4 illustrate bias and relative bias results for $K = 5$. In high-dimensional conditions, both GVEM-BSP and GVEM-BS methods consistently outperformed the GVEM-USEM method, yielding approximately zero bias and relative bias results for SE estimates under all conditions, except when factor correlations were high and the item structure was within-item M2PL, for both test lengths of 45 and 30. Under these two conditions, all methods exhibited increased variability in SE estimates. The GVEM-BS method consistently yielded positive bias and relative bias results. When $J = 45$, the GVEM-BSP overestimated SEs, whereas it underestimated them when $J = 30$. The GVEM-USEM method yielded more outliers in scenarios where $J = 30$, factor correlations were high and the item structure was within-item M2PL. Under other conditions, it tended to overestimate SEs, consistently with our findings for $K = 3$. The general trend of bias and relative bias results remained

the same as in low-dimensional conditions. That is, the SE estimates became more challenging with high factor correlations and a within-item structure, although the GVEM-BSP consistently demonstrated superior performance across all conditions.

Figures 4.1 and 3.6 present estimated SEs for each method. It is important to note that different scales are employed for the left and right columns to enhance the visibility of differences between methods. In general, GVEM-BS and GVEM-BSP method produced comparable results in terms of estimating SEs smaller than 0.15. Three exceptions were observed: 1) when $K = 3$, $J = 30$, factor correlations with high factor correlations and a within-item M2PL structure; 2) when $K = 5$, $J = 45$, with high factor correlations and a within-item M2PL structure; 3) when $K = 5$, $J = 30$, with high factor correlations and a within-item M2PL structure. Under these three conditions, the GVEM-BS estimated larger SEs for item discrimination parameters, possibly contributing to increased variability in bias and relative bias results under high latent correlation and within-item M2PL conditions. The GVEM-USEM method estimated smaller SEs than the GVEM-BS under these three conditions, while generally estimating relatively larger SEs than the two bootstrap methods and presenting some outliers in other scenarios.

The elapsed time for three GVEM methods per replication under two conditions is also shown in Table 3.2. In the first condition where the test length was 45, the GVEM-BS method required approximately 0.87 minutes per replication. Conversely, the GVEM-BSP method exhibited a slightly longer processing time, averaging 1.20 minutes per replication. Notably, the GVEM-USEM method displayed superior computational efficiency, completing each replication in a mere 0.08 minutes. Under the second condition with a reduced test length of 30 items, the elapsed time for all three GVEM methods increased marginally. Specifically, the GVEM-BS method recorded an average of 1.17 minutes per replication, while the GVEM-BSP method required 1.32 minutes. Remarkably, the GVEM-USEM method again demonstrated its efficiency, completing each replication in just 0.04 minutes. These results indicate the consistent superiority of the GVEM-USEM method in terms of computational efficiency, as evidenced by substantially shorter elapsed times compared to the other

two methods across both test length conditions.

Table 3.2: The elapsed time for three GVEM methods per replication under two conditions.

Condition	Method	Elapsed time (mins)
Between-item M2PL, Low correlations, $K = 5$, $J = 45$	GVEM-BS	0.87
	GVEM-BSP	1.20
	GVEM-USEM	0.08
Between-item M2PL, Low correlations, $K = 5$, $J = 30$	GVEM-BS	1.17
	GVEM-BSP	1.32
	GVEM-USEM	0.04

CPU: 2.40 GHz 20-Core Intel Xeon; RAM: 1.00TB 2133 MHz DDR4

3.7 Discussion

The recent work of Cho et al. (2021) has shown the superiority of the GVEM algorithm in terms of computation efficiency and item parameter estimation accuracy, but its SE estimation is not fully discussed. Since obtaining accurate SEs is also an important prerequisite for many applications, the current study applied the USEM and bootstrap methods within the GVEM framework and compared the performance of three GVEM methods (GVEM-BSP, GVEM-BS, GVEM-USEM) with respect to SE recovery. The simulation results showed that the GVEM-BSP performed the best under most conditions, because adding a prior could make parameter estimation more stable and robust. Although computationally more efficient, GVEM-USEM tended to exhibit an upward bias. GVEM-BS method demonstrated comparable performance to GVEM-BSP under conditions with low factor correlations and between-item structure, yet displayed increased variability under scenarios involving high factor correlations and within-item conditions.

The GVEM-BSP is a promising method to estimate SE. Moreover, our pilot study finds that it can produce accurate item parameter estimates, making it suitable for score reporting. Besides, the GVEM-BSP can be extended to the Variational Bayesian (VB) estimation (Bishop, 2006), an alternative approximation technique to solve intractable integrals by specifying variational distributions of item parameters. The main advantage of VB is that SEs of item parameters can be derived with closed-form solutions. However, our pilot study shows that interestingly, it underestimates SE consistently, and we will defer a detailed examination of this method to a future study.

Although the literature demonstrates several great benefits of the USEM method in terms of SE estimation, incorporating it into the GVEM algorithm has several drawbacks. First, unlike traditional EM algorithms, GVEM requires additional derivation (e.g., the inverse of the complete-data information matrix $\mathbf{I}_c^{-1}$) to compute SE. Second, the USEM method relies on the information matrix based on the variational lower bound $E(\mathbf{M}, \boldsymbol{\xi})$ of the marginal log-likelihood, which could incur some bias of SE estimates. Third, the information matrix in the current study is only a sub-block of the entire information matrix induced by $(\mathbf{M}, \boldsymbol{\xi})$, and hence could be ill-conditioned. This also explains the occurrence of unsuccessful replications (see Table 3.1).

The current study can be extended in the following directions. First, all of the conditions were in a confirmatory mode, meaning that the factor loading structure was assumed to be known. For unknown cases such as exploratory factor analysis, the proposed SE estimation procedures can be combined with, e.g., the GVEM with adaptive lasso penalty, which has been shown to accurately recover the model parameters and the loading structure for such purpose (Cho et al., 2021; Cho et al., 2022). Second, the current study was based on a two-parameter MIRT model. It is desirable to investigate the performance of different SE estimation procedures within the GVEM framework under other types of MIRT models, such as the M3PL model including guessing behaviors, or even the M4PL considering inattention situations. Lastly, the current simulation design was more of an ideal scenario where the response dataset did not have any missing values. However, missing data is ubiquitous in

practice, which could result in biased parameter estimates and inflated SEs (Kalkan et al., 2018). Future studies are necessary to explore the performance of GVEM with different SE estimation procedures under various missing data scenarios.

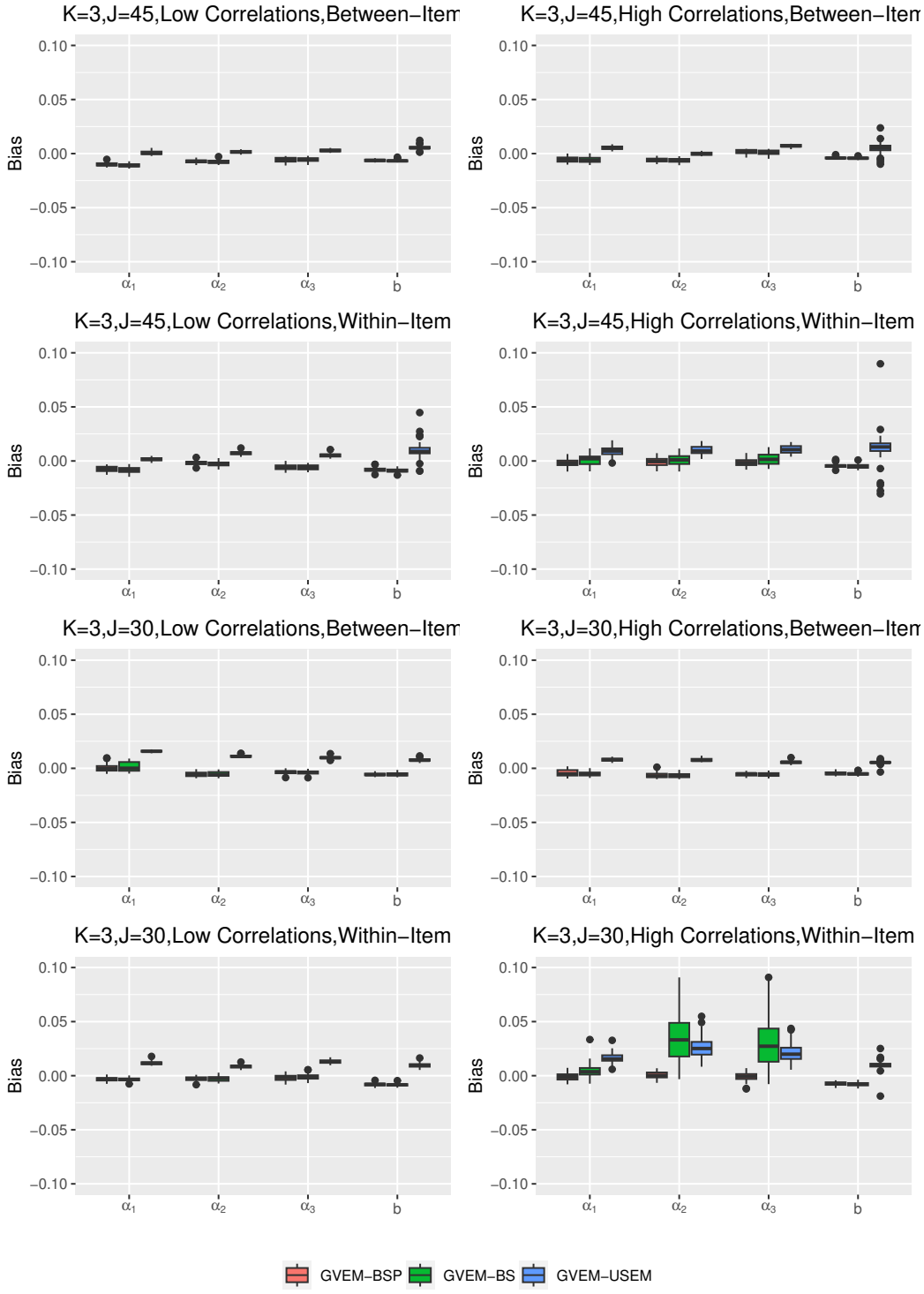

Figure 3.1: Bias Comparison for Item Parameters' Standard Errors Estimates When $K = 3$

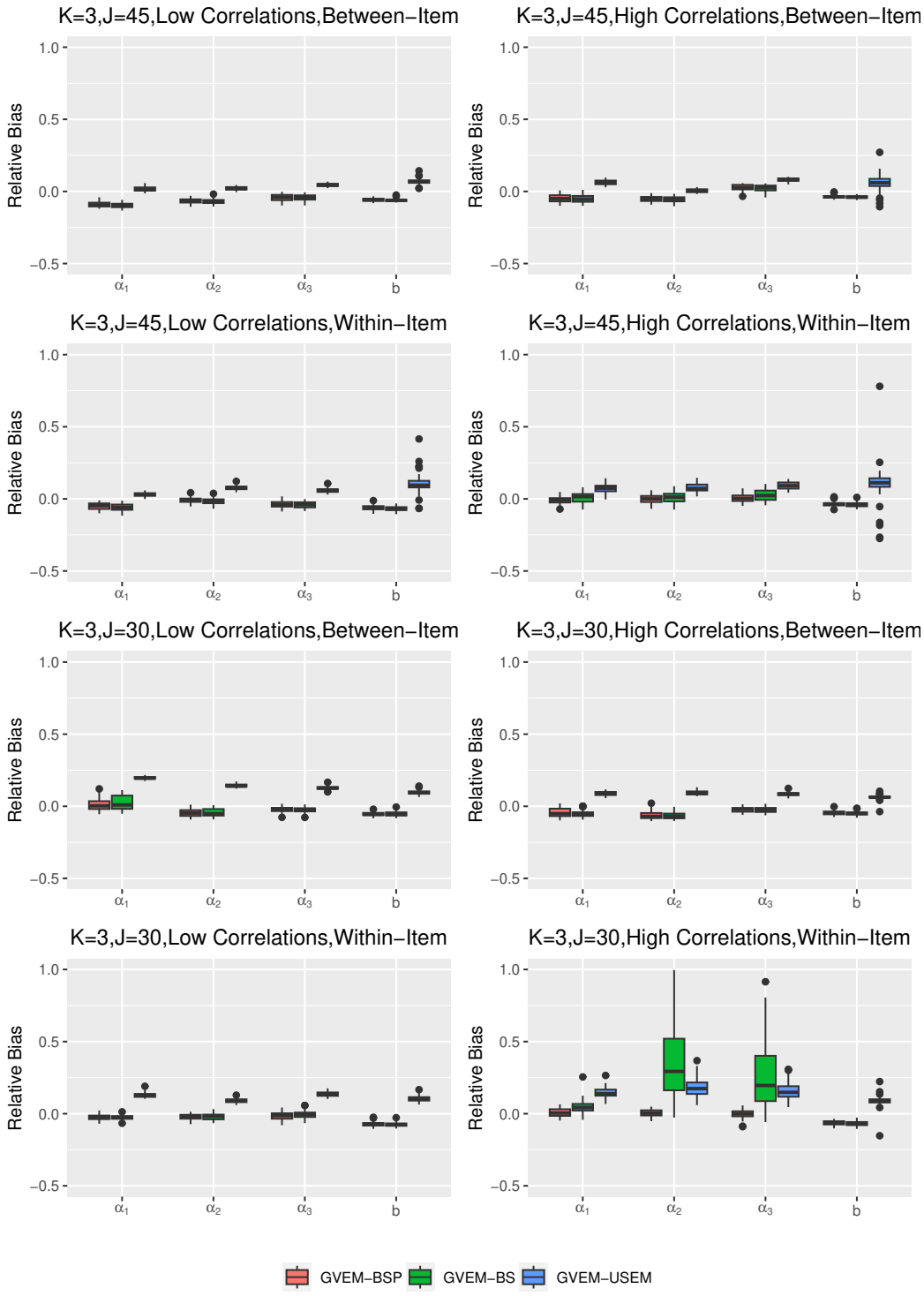

Figure 3.2: Relative Bias Comparison for Item Parameters' Standard Errors Estimates When $K = 3$

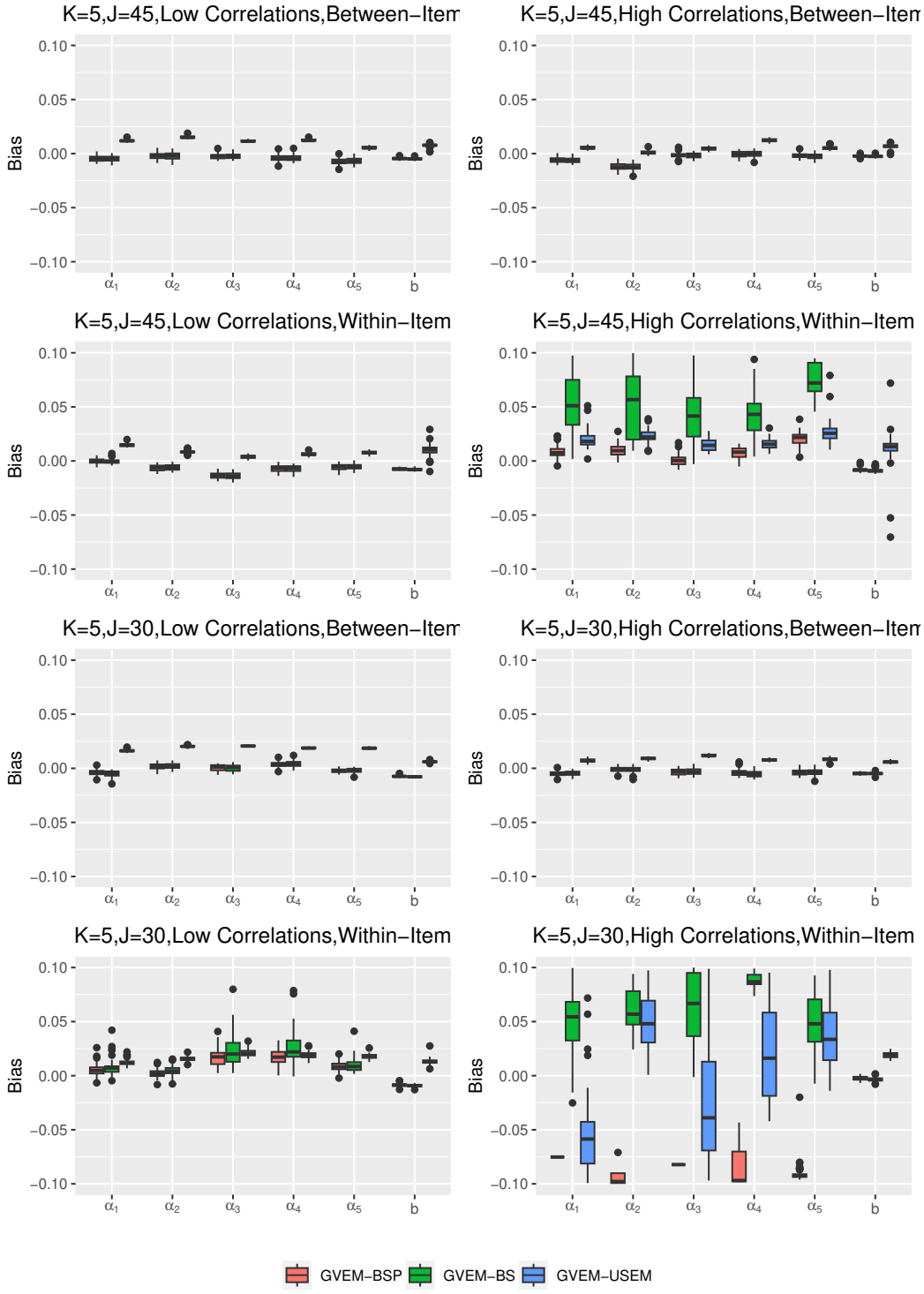

Figure 3.3: Bias Comparison for Item Parameters' Standard Errors Estimates When $K = 5$

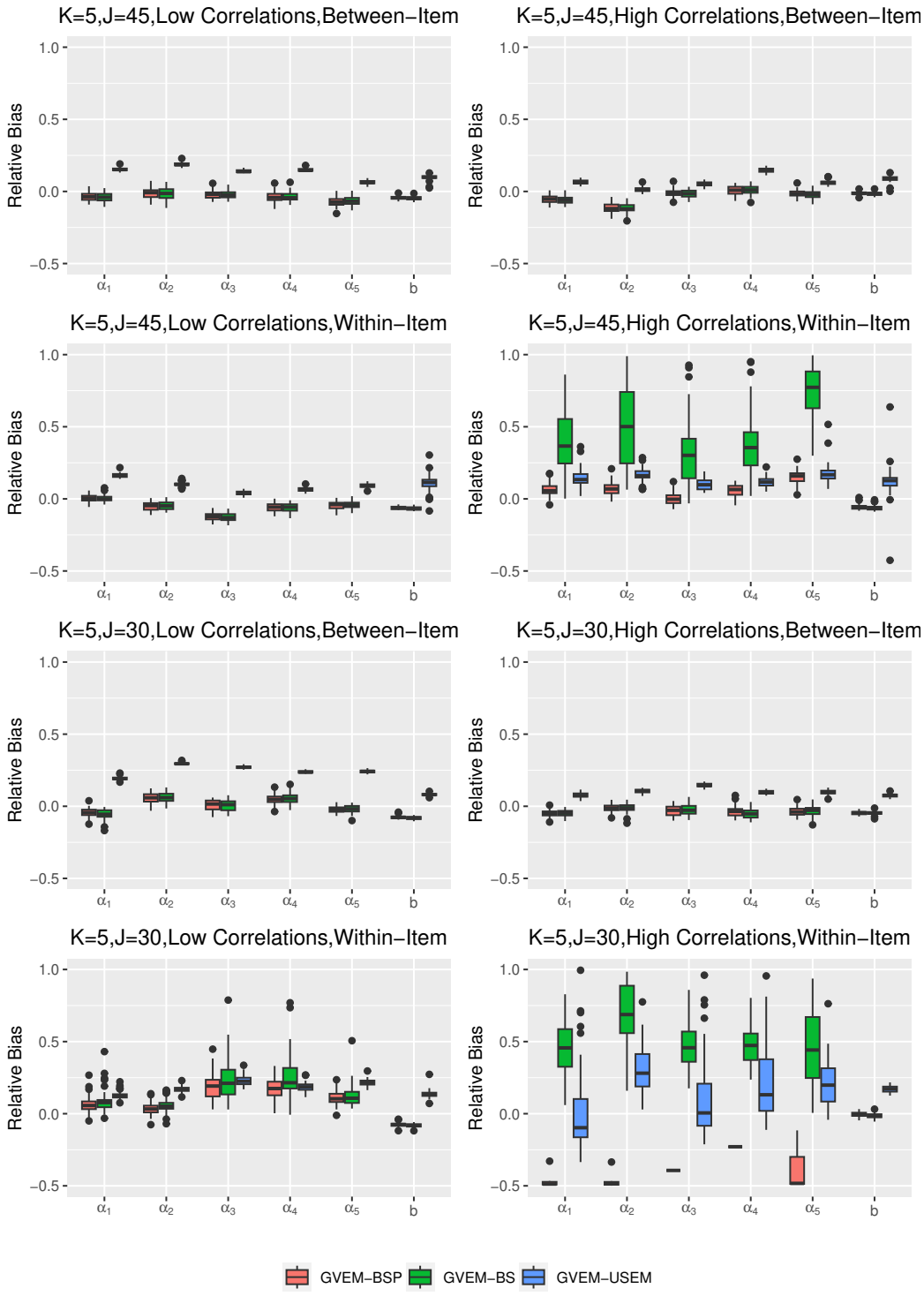

Figure 3.4: Relative Bias Comparison for Item Parameters' Standard Errors Estimates When $K = 5$

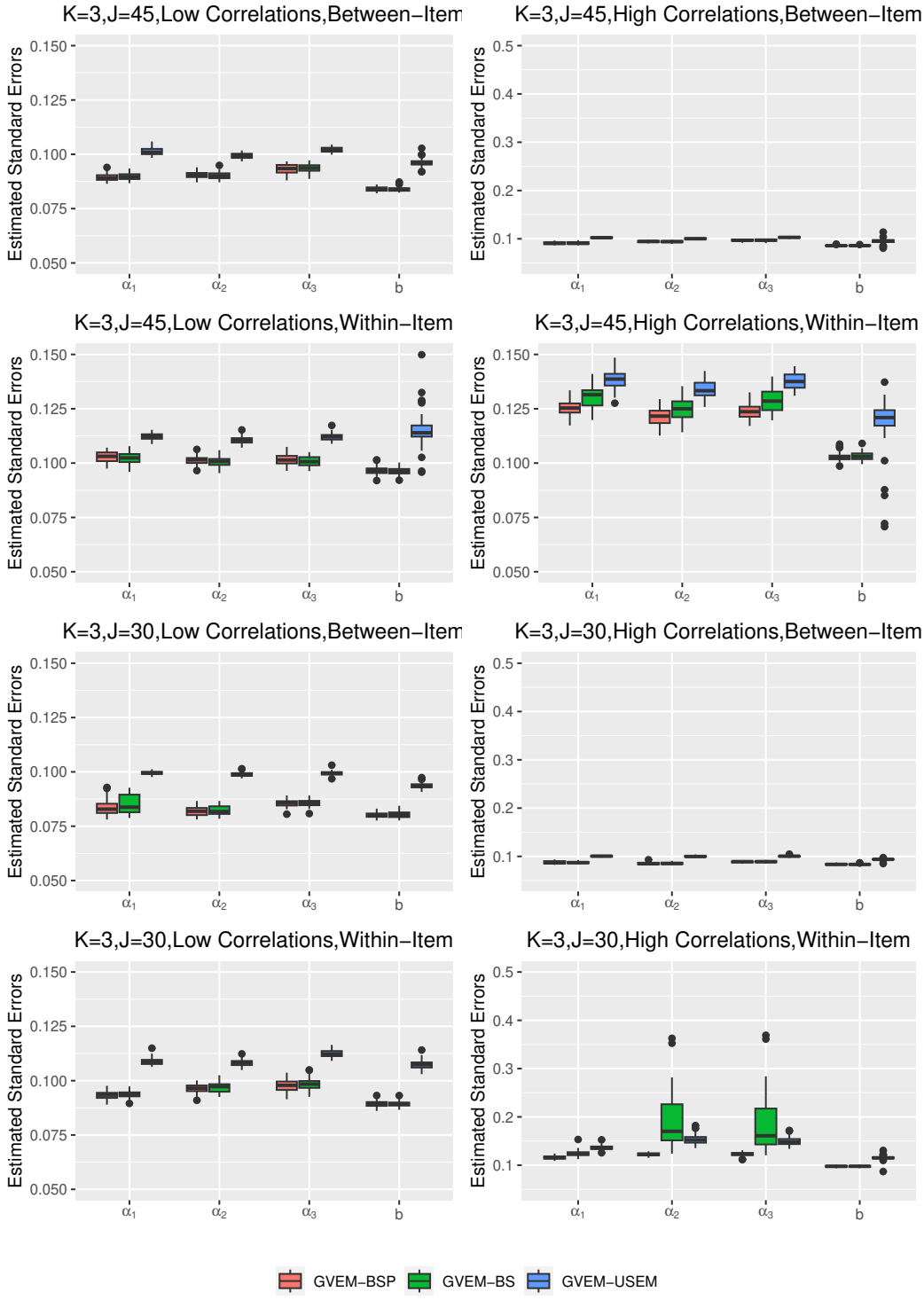

Figure 3.5: Estimated Standard Errors Comparison When $K = 3$

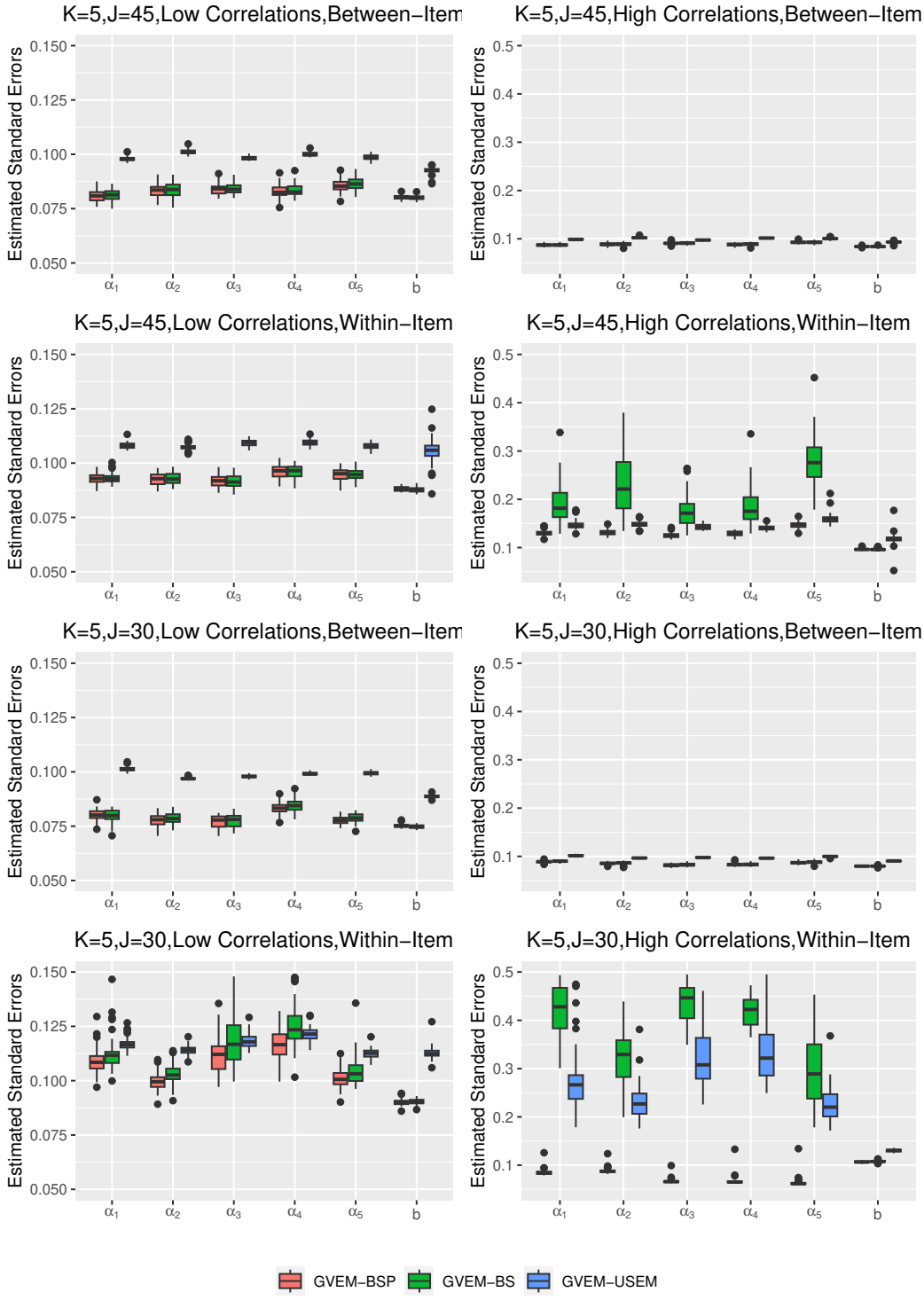

Figure 3.6: Estimated Standard Errors Comparison When $K = 5$

Chapter 4

VEMIRT: AN R PACKAGE FOR HIGH-DIMENSIONAL IRT MODELS

When assessments or surveys aim to measure multiple constructs, multidimensional item response theory (MIRT) models provide a comprehensive framework and statistical tool for item calibration, scoring, and analysis (Ma et al., 2023). However, a major challenge for implementing MIRT is the computational complexity arising from intractable integrals in the likelihood function when estimating item parameters in higher dimensional space (Andersson & Xin, 2021). Despite this challenge, advancements in statistical models and computational power have have propelled MIRT research forward, rendering MIRT models increasingly viable for practical applications (Chalmers, 2012; M. C. Edwards, 2010).

Numerous solutions have been proposed in the literature to address the intractable integral issue in MIRT, including but not limited to the adaptive Gaussian quadrature method (Cagnone & Monari, 2013), the Laplace approximation (Lindstrom & Bates, 1988), Metropolis–Hastings Robbins–Monro (MH-RM) method (Cai, 2010a, 2010b). These methods are versatile and can fit various MIRT models, often available in commercial software packages or R packages like Mplus(Muthén & Muthén, 2009), IRTPRO (Cai, Du Toit, et al., 2011), flexMIRT (Cai, 2013), `mirt` package(Chalmers, 2012), `lamle` package (Andersson & Xin, 2021), `MCMCpack` package (Martin et al., 2011), and `brms` package (Bürkner, 2019).

Recently, Cho et al. (2021) introduced a Gaussian variational expectation–maximization (GVEM) algorithm to further enhance computational efficiency and estimation accuracy. GVEM approximates the intractable likelihood within the EM framework by proposing a variational lower bound (Cho et al., 2021; Cho et al., 2022). Previous research has showcased its superior performance in both exploratory and confirmatory MIRT models, including mul-

tidimensional two-parameter logistic (M2PL) models and multidimensional three-parameter logistic (M3PL) models (Cho et al., 2021). Additionally, Cho et al. (2022) proposed a regularized GVEM algorithm to efficiently and accurately recover item factor loading structures. The new GVEM algorithm incorporates an (adaptive) L_1 penalty to the variation lower bound to shrink certain factor loadings to 0. Moreover, to address bias in item discrimination parameters during confirmatory MIRT model estimation, (Ma et al., 2023) proposed an importance-weighted version of GVEM (IW-GVEM) to improve the estimation accuracy with only a modest increase of computation time.

Given the superior performance of GVEM, ensuring direct accessibility to researchers and practitioners is essential. Thus, this chapter introduces a `VEMIRT` R package, which implements GVEM algorithms for item parameter estimation across various MIRT models, including exploratory M2PL, confirmatory M2PL, exploratory M3PL, and confirmatory M3PL. Beyond item parameter recovery, the package empowers users to compute standard errors of item parameters and implement corrections such as bootstrap sampling and importance sampling, thereby enhancing the estimation accuracy.

This study serves three primary goals. Firstly, it offers a comprehensive conceptual overview of the GVEM framework. Secondly, it elucidates the implementation of this framework within statistical software. Lastly, through practical examples, it illustrates how the software can be applied to address real-world inquiries. The rest of this chapter is structured as follows. First, we review the existing software and packages developed to estimate MIRT models. Then, a brief description of the statistical background for GVEM models is provided. Next, the main functionalities of `VEMIRT` with some illustrative toy examples are presented. Lastly, the chapter concludes with a summary and further discussion.

4.1 Commercial Software and Packages for MIRT Models

Over the decades, numerous commercial software packages have emerged to facilitate item analysis within the MIRT framework, each varying in their supported statistical methods (Bürkner, 2019; Mair & Hatzinger, 2016). While providing an overview of all such software

packages would be exhaustive and perhaps not entirely relevant to the study goal, this section focuses on introducing six widely used commercially available software and packages for conducting MIRT analyses. These include Mplus (Muthén & Muthén, 2017), IRTPRO(Cai, Du Toit, et al., 2011), flexMIRT (Cai, 2013), `mirt` package (Chalmers, 2012), `lamle` package (Andersson & Xin, 2021), `brms` package (Bürkner, 2019).

4.1.1 Mplus

Mplus stands out as a latent variable modeling program offering a diverse range of models for data analysis, including but not limited to structural equation modeling, regression and path analysis, and multilevel mixture modeling (Muthén & Muthén, 2017). A distinctive attribute of Mplus is its flexibility in accommodating various types of analyses within a unified modeling framework. For example, it allows users to combine exploratory factor analysis with growth modeling by using an exploratory factor analysis measurement model in a growth model (Asparouhov & Muthén, 2023). It supports an extensive range of IRT models, encompassing 1PL, 2PL, 3PL, 4PL (including an upper asymptote parameter), graded response (Samejima, 1976), generalized partial credit (Muraki & Muraki, 2016), and nominal response models (Thissen et al., 2011). However, it's important to note that Mplus may exhibit limitations in accurately estimating parameters for M3PL and M4PL models (Sims, 2017). To address this issue, Mplus allows users to incorporate prior information for model parameters, enhancing identifiability and improving estimation accuracy (Asparouhov & Muthén, 2016).

When it comes to model estimation methods, Mplus offers three main options: maximum likelihood (ML) estimation, weighted least square (WLSMV) estimation, and Bayesian estimation (Paek et al., 2018). ML estimation is the most commonly used approach and employs various techniques to overcome computational challenges associated with MIRT models, such as rectangular integration, Gauss-Hermite quadrature, and Monte Carlo (MC) integration (Asparouhov & Muthén, 2012; Han & Paek, 2014). While Gauss-Hermite quadrature can effectively handle models with up to three or four latent factors, Monte Carlo integration proves beneficial for both confirmatory and exploratory factor analyses in high-dimensional

conditions. When employing WLSMV estimation in Mplus, it's essential to apply additional constraints to guarantee model identification (Asparouhov & Muthén, 2016). Specifically, Mplus imposes constraints such as setting the residual variance to 1 (referred to as the "theta" parameterization) or fixing the variance of latent traits to 1 (known as the "delta" parameterization). These constraints help ensure that the model remains identifiable and that parameter estimates are reliable. In addition to item parameter estimation, Mplus provides estimates for bootstrap standard errors for item parameters.

While Mplus offers remarkable flexibility and a multitude of functions for analyzing MIRT data, it should be noted that different conversion formulas are necessary to translate factor analysis parameters into IRT parameters when employing WLSMV estimation (C. Wang & Zhang, 2019). This underscores the importance of researchers and practitioners being well-versed in the relationships between IRT and item factor analysis (FA) parameterizations. Moreover, Mplus's user interface primarily consists of a text editor, where users directly input syntax commands and manipulate input files (Han & Paek, 2014). While it does provide a basic syntax generator through dialog boxes, users are expected to have some coding background to navigate and utilize the software effectively. It is worth noting that Mplus is not available for free, as academic users are required to purchase a single-user license priced at US$595. Despite this cost, the software's extensive capabilities and robust features make it a valuable tool for latent variable modeling and item response theory analysis.

4.1.2 IRTPRO

IRTPRO is an advanced software for both unidimensional and multidimensional item calibration and test scoring (Cai, Du Toit, et al., 2011). Built upon a highly generalized IRT framework, IRTPRO accommodates multiple groups, response categories, and dimensions (Han & Paek, 2014). As a result, it can analyze various IRT models, including 1PL, 2PL, 3PL, 4PL, graded response, generalized partial credit, and nominal response models. Similar to Mplus, IRTPRO allows users to combine these models within a test or scale in any configuration, enabling the specification of equality constraints or fixed values for parameters as

needed (Cai, Du Toit, et al., 2011).

For MIRT estimation, IRTPRO offers four different methods (Cai, Du Toit, et al., 2011; Han & Paek, 2014): the Bock–Aitkin approach utilizing the expectation–maximization algorithm (BAEM; Bock & Aitkin, 1981), the adaptive quadrature approach (Schilling & Bock, 2005) with several options for numerical integrations (i.e., Gauss–Hermite, Monte Carlo, and Latin Hypercube),and the MH-RM method (Cai, 2010a, 2010b), and the Makov chain Monte Carlo (MCMC) based on the Patz-Junker's blocked Metropolis algorithm (Junker, 1999). The tutorial suggests that for two-dimensional models, both the Bock-Aitkin approach and adaptive quadrature are effective. For three- to four-dimensional models, adaptive quadrature and MH-RM are preferred. Higher dimensional models are best handled using MH-RM and MCMC. In addition to item parameter estimation, IRTPRO implements several standard error estimation methods for item parameters: SEM (Cai, 2008), XPD, sandwich covariance matrix and M-step values, accumulation (Cai, 2010a), and Monte Carlo (Diebolt & Ip, 1995).

Unlike Mplus, IRTPRO offers a mouse-driven dialog box-based graphical user interface (GUI) as the primary mode of interacting with program features for most users. Furthermore, users have the option of creating syntax files from scratch or reusing syntax files for repeated runs, which may be common for operational data analysis (Cai, Du Toit, et al., 2011). However, IRTPRO does not support the estimation of 4PL models. Additionally, for academic users, a new single-user license costs $495.

4.1.3 flexMIRT

flexMIRT is recognized as the most advanced IRT software, capable of fitting a variety of IRT models, including multilevel, multidimensional, and multiple group models, for item analysis and test scoring (Cai, 2013). It can fit a wide variety of IRT models including 1PL, 2PL, 3PL, graded response, generalized partial credit models, and their combinations. One notable feature of flexMIRT is its utilization of parallel processing, which accelerates computations by automatically distributing the workload across multiple available cores or

processing units (Houts & Cai, 2013).

By default, flexMIRT utilizes the BAEM algorithm (Bock & Aitkin, 1981) to estimate MIRT models. However, for high-dimensional cases, BAEM may suffer from computational limitations, leading to exceedingly long computation times or failure to complete (Houts & Cai, 2013). To address this issue, flexMIRT offers two alternative estimation methods: MH-RM (Cai, 2010a, 2010b) and MCMC (M. C. Edwards, 2010). Additionally, flexMIRT supports numerous methods for estimating item parameters' standard errors, including SEM, XPD, Fisher (expected) information matrix, Richardson extrapolation, sandwich covariance matrix and forward difference (Cai, 2013).

Similar to Mplus, flexMIRT primarily relies on a syntax-based interface without graphical UI elements. Users must manually write syntax commands, requiring some basic coding knowledge (Han & Paek, 2014). Like IRTPRO, flexMIRT does not support the estimation of 4PL models. Additionally, flexMIRT operates on a subscription-based model, priced at US$125 per year for academic users, with up to three installations permitted (Han & Paek, 2014).

4.1.4 R Packages for MIRT Analysis

Apart from the aforementioned commercial software, several free and open-source packages have been developed to facilitate MIRT analysis within the R programming language (R Core Team et al., 2013). Among these, the `mirt` package stands out as a widely used tool (Chalmers, 2012). The `mirt` package is capable of analyzing both dichotomous and polytomous response data, employing ML estimation methods within both unidimensional and multidimensional IRT frameworks. It can analyze various IRT models, including 1PL, 2PL, 3PL, 4PL, graded response, generalized partial credit, nominal response models, and their combinations. Within the `mirt` package, users can estimate MIRT models using two primary methods: the BAEM algorithm (Bock & Aitkin, 1981) and the MH-RM algorithm (Cai, 2010a, 2010b). The BAEM algorithm offers several integration grid options, including rectangular, quasi-Monte Carlo integration grids (QMCEM), and stochastic techniques

(MCEM) (Chalmers, 2012). While the rectangular grid method is effective for models with fewer than three factors, the developer recommends utilizing QMCEM, MCEM, or MH-RM for models with three or more dimensions. In addition to MIRT model estimation, the `mirt` package provides functionality for estimating item parameters' standard errors, including Richardson, XPD, sandwich covariance matrix, Oakes method, SEM, Fisher information matrix, and MH-RM (Chalmers, 2012). While the mirt package offers a flexible framework capable of fitting various models and employing diverse estimation algorithms, it may become time-consuming when handling high-dimensional MIRT analysis. For instance, when employing the MH-RM method for both the mirt package and flexMIRT, flexMIRT demonstrates superior efficiency.

In addition to the `mirt` package, the `lamle` R package offers adaptive quadrature or Laplace approximations to address the intractable integrals in the likelihood function of confirmatory MIRT models (Andersson & Xin, 2021) or generalized linear latent variable models (Andersson et al., 2023). This package supports a variety of models, including the generalized partial credit model (Muraki & Muraki, 2016), the graded response model (Samejima, 1976), and generalized linear latent variable models for negative-binomial, Poisson, and normal distributions. It accommodates different data types, such as binary, ordinal, count, continuous, and their combinations. However, this package has limited documentation and fewer user-friendly features compared to `mirt` package. However, this package has limited documentation and fewer user-friendly features compared to the `mirt` package. For example, the documentation does not specify whether it supports M2PL or M3PL models. This package provides standard error estimates but does not specify the method applied.

On the other hand, the `MCMCpack` R package utilizes the MCMC approach to estimate 2PL MIRT models (normal, heteroscedastic, and robust estimation) (Martin et al., 2011). Despite its compilation of C++ code to enhance computational speed, the package remains computationally intensive and demands significant memory storage (Chalmers, 2012). Additionally, proficiency in Bayesian inference is required to effectively use this package. Furthermore, the `MCMCpack` package lacks flexibility in fitting complex MIRT models, such as M3PL and

M4PL models, and it cannot handle polytomous response data.

Another R package employing Bayesian statistics is `brms` (Bürkner, 2019), which leverages Stan (Carpenter et al., 2017) for model estimation. It employs MCMC sampling via adaptive Hamiltonian Monte Carlo (Hoffman, Gelman, et al., 2014), an efficient and stable algorithm suitable for high-dimensional, highly correlated parameter spaces. By incorporating prior knowledge in the form of prior distributions, `brms` can fit a wide range of Bayesian IRT models, including 1PL, 2PL, 3PL, 4PL, graded response, and generalized partial credit models (Bürkner, 2019). Therefore, it can handle binary, categorical, and ordinal data. Similar to the `MCMCpack` package, `brms` package necessitates an understanding of Bayesian statistics and prior distributions. Furthermore, if the objective is to provide estimates in real-time scenarios, such as for adaptive testing purposes, full Bayesian inference may be too slow unless specifically tuned for such tasks (Bürkner, 2019; van der Linden & Ren, 2015).

4.2 GVEM Methods

In the preceding chapters, we have delved into the foundational GVEM framework (see section Gaussian Variational Expectation Maximization (GVEM) in Chapter 1) and its adaptations for handling missing data scenarios, as seen in GVEM-BS in Chapter 2, applicable to both exploratory and confirmatory M2PL models, along with techniques for standard error (SE) estimation (such as GVEM-BS, GVEM-BSP, GVEM-USEM in Chapter 3) tailored to confirmatory M2PL models. Building upon this groundwork, this section will elucidate additional GVEM methods designed for exploratory scenarios, including their modifications within the M3PL framework. Furthermore, we will introduce the importance-weighted version of GVEM (denoted as IW-GVEM), specifically aimed at mitigating bias in confirmatory M2PL models.

4.2.1 GVEM for Exploratory Analysis

MIRT models can be employed in either exploratory or confirmatory manners. In a confirmatory MIRT model, the item factor loading structure is predetermined, leading to the

constraint of many item discrimination parameters $\boldsymbol{\alpha}_j$ (or loading parameters) to 0. Conversely, an exploratory MIRT model assumes that the item factor loading structure, akin to the sparsity structure of $\boldsymbol{\alpha}_j$, is unknown. To identify the item factor loading structure, a two-step approach can be utilized under the GVEM framework. Initially, all item factor loadings are estimated freely to satisfy identifiability constraints. Subsequently, a post-hoc rotation (Browne, 2001), which may involve methods like promax (Hendrickson & White, 1966) or CF-Quartimax rotation combined with an arbitrary cutoff for the rotated factor loading, yields a sparsity factor loading matrix.

To avoid setting an arbitrary cutoff, Cho et al. (2022) added a regularization penalty to the GVEM algorithm to estimate factor loading structure and model parameters simultaneously. Specifically, they proposed both Lasso and adaptive Lasso versions to produce accurate factor loading structure recovery. For Lasso penalization, the following optimization problem should be solved:

$$(\hat{\mathbf{A}}_\lambda, \hat{\mathbf{B}}_\lambda) = \text{argmax}_{\mathbf{A},\mathbf{B}} l(\mathbf{A}, \mathbf{B}; \mathbf{Y}) - \lambda \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{w}_{jk} |\alpha_{jk}| \tag{4.1}$$

where $\hat{w}_{jk} = 1/|\hat{\alpha}_{jk}|^\gamma$, and $\gamma > 0$ and $\lambda > 0$ refer to tuning parameters. While Lasso offers computational efficiency in estimating MIRT models, its bias may hinder consistent variable selection and model estimation (Cho et al., 2022). An extension known as adaptive Lasso addresses this issue by adjusting the penalty parameter for each parameter α_{jk}. Instead of using a constant penalty parameter λ, adaptive Lasso employs penalization weights $\lambda \hat{w}_{jk} = \lambda/|\hat{\alpha}_{jk}|^\gamma$. This adjustment ensures that smaller values of $\hat{\alpha}_{jk}$ are penalized more than larger values, improving the accuracy of the estimation process. Generalized information criterion (GIC) (Fan & Tang, 2013) is implemented to choose the optimal penalty parameter λ. The parameter that minimizes GIC will be considered optimal. To ensure identifiability, Cho et al. (2022) suggested two constraint settings for the item factor loading structure: either setting a $K \times K$ identity sub-matrix or a triangular sub-matrix with ones on the diagonal. Both constraints are available in the package and more details will be introduced in later

sections.

4.2.2 IW-GVEM for Confirmatory M2PL Models

Previous research has indicated that the GVEM algorithm can result in relatively large bias on item discrimination parameters ($\boldsymbol{\alpha}_j$) in confirmatory MIRT models, particularly when the factor correlations are high and the sample size is not large (Ma et al., 2023). This bias issue is common for variational algorithms (Bishop, 2006). To address this bias, Ma et al. (2023) proposed an importance-weighted GVEM algorithm for confirmatory M2PL models. This approach incorporates an importance-weighted variational inference technique to create a tighter variational lower bound to the target, which is otherwise intractable, marginal likelihood. Specifically, for each examinee i, D samples are drawn from the variational distribution $q_i(\boldsymbol{\theta}_i)$ for S times: $\boldsymbol{\theta}_i^{(s,d)} \sim q_i(\boldsymbol{\theta}_i)$, for $s = 1, \ldots, S, d = 1, \ldots, D, i = 1, \ldots, N$, where $q_i(\boldsymbol{\theta}_i^{(s,d)}) \sim N(\boldsymbol{\theta}_i^{(s,d)} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Then, a tighter lower bound $\mathcal{L}_M(\boldsymbol{Y})$ can be approximated by

$$\mathcal{L}_D(\boldsymbol{Y}) \approx \sum_{i=1}^{N} \left(\frac{1}{S} \sum_{s=1}^{S} \left[\log \frac{1}{D} \sum_{d=1}^{D} w_i^{(s,d)} \right] \right) \tag{4.2}$$

where $w_i^{(s,m)} = p(Y_i, \boldsymbol{\theta}_i^{(s,d)} \mid \boldsymbol{A}, \boldsymbol{B}) / q_i(\boldsymbol{\theta}_i^{(s,d)})$, and the joint distribution function of $p(Y_i, \boldsymbol{\theta}_i^{(s,d)} \mid \boldsymbol{A}, \boldsymbol{B})$ is

$$\begin{aligned} & \log p(Y_i, \boldsymbol{\theta}_i^{(s,d)} \mid \boldsymbol{A}, \boldsymbol{B}) = \log p(Y_i \mid \boldsymbol{\theta}_i^{(s,d)}, \boldsymbol{A}, \boldsymbol{B}) + \log \phi(\boldsymbol{\theta}_i^{(s,d)}) \\ = & \sum_{j=1}^{J} \left\{ Y_{ij} (\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i^{(s,d)} - b_j) + \log \frac{1}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i^{(s,d)} - b_j)} \right\} + \log \phi(\boldsymbol{\theta}_i^{(s,d)}). \end{aligned}$$

The gradient ascent method is implemented to update model parameters, and the Adaptive moment estimation (Adam) method (Kingma & Ba, 2014) is applied to adjust the learning rate.

4.2.3 GVEM for the M3PL Model

The M2PL model, discussed in earlier sections, is widely recognized as a cornerstone in MIRT modeling. Compared with the M2PL model, the M3PL model is particularly suitable for

multiple-choice scenarios, as it integrates an additional parameter c_j to assess the guessing probability of answering item j. The item response function for the M3PL model is expressed as follows:

$$P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i) = c_j + (1 - c_j)\frac{\exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}. \tag{4.3}$$

The inclusion of the guessing parameter in the M3PL model complicates the estimation process because the item response function no longer belongs to the exponential family (Cho et al., 2022; Thissen & Wainer, 1982). To address this issue, Cho et al. (2021) utilized an equivalent representation of the M3PL model with an auxiliary latent variable Z_{ij}, which indicates whether the ith individual correctly answers the jth item based on their latent ability, or guesses it. Here, $Z_{ij} = 1$ if the ith individual solves item j based on their latent ability, and $Z_{ij} = 0$ if they guess item j correctly (von Davier, 2009). Let $\boldsymbol{Z}i = Zi1, Z_{i2}, \ldots, Z_{iJ}$ and its distribution be denoted as $p(\boldsymbol{Z}i) = \prod j = 1^J p(Z_{ij})$, where $Z_{ij} \sim$ Bernoulli$(1-c_j)$. Then, the complete data likelihood of the ith subject is: Then the complete data likelihood of the ith subject is

$$\begin{aligned}
& \log P(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \\
= \; & \log P(Y_i \mid \boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \log \phi(\boldsymbol{\theta}_i) + \log p(\boldsymbol{Z}_i) \\
= \; & \sum_{j=1}^{J} \left\{ Y_{ij} Z_{ij} (\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j) + Z_{ij} \log \frac{1}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)} + (1 - Z_{ij}) \log I(Y_{ij} = 1) \right\} \\
& + \log \phi(\boldsymbol{\theta}_i) + \log p(\boldsymbol{Z}_i).
\end{aligned} \tag{4.4}$$

Using the same variational lower bound as in M2PL, the variational lower bound in M3PL could be defined as

$$
\begin{aligned}
& \log P(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \\
\geq\ & \sum_{j=1}^{J} Z_{ij} \log \frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} + \sum_{j=1}^{J} Z_{ij} Y_{ij} (\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j) + \sum_{j=1}^{J} \frac{1}{2} Z_{ij} (b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - \xi_{i,j}) \\
& - \sum_{j=1}^{J} Z_{ij} \eta(\xi_{i,j}) \{ (b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2 - \xi_{i,j}^2 \} + \sum_{j=1}^{J} \{ (1 - Z_{ij}) \log I(Y_{ij} = 1) \} \\
& + \log \phi(\boldsymbol{\theta}_i) + \log p(\boldsymbol{Z}_i) \\
=:\ & l(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{\xi}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}). \qquad (4.5)
\end{aligned}
$$

The variational lower bound of the marginal likelihood with respect to variational distributions q_i and r_i of the latent variables $\boldsymbol{\theta}_i$ and Z_{ij} can be defined as

$$
E^{(t)}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}) := \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \left[\sum_{\boldsymbol{Z}_i} l(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{\xi}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \times r_i^{(t)}(\boldsymbol{Z}_i) \right] \times q_i^{(t)}(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \qquad (4.6)
$$

Appropriate choices of the variational distributions could lead to a closed-form expression of the lower bound expressed in (4.6). However, the GVEM algorithm in M3PL becomes computationally inefficient as the sample size increases (Cho et al., 2021). To address this, (Cho et al., 2021) introduced a stochastic optimization approach for the variational approximation in the E step (Hoffman et al., 2013). Specifically, for each iteration $t \geq 1$, a subset of data U_t with a desired size is chosen, and model parameters are updated based on this subset. With updated variational parameters partially for $i \in U_t$, a noisy estimate of the expected variational lower bound for the tth iteration, denoted as $\hat{Q}_t$, is calculated as follows:

$$
\hat{Q}_t = \sum_{i \in U_t} \int_{\boldsymbol{\theta}_i} \left[\sum_{\boldsymbol{Z}_i} l(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{\xi}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \times r_i^{(t)}(\boldsymbol{Z}_i) \right] \times q_i^{(t)}(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i
$$

A stochastic approximation of the variational lower bound is then obtained by taking a weighted average of the noisy estimates of the lower bound from the previous and current steps, i.e., $(1 - \epsilon_t)\hat{Q}t - 1 + \epsilon t \hat{Q}_t$, where ϵ_t denotes a decreasing step size.

The GVEM with Lasso penalization can also be extended for exploratory M3PL models (Cho et al., 2022). The following optimization problem should be solved:

$$(\hat{\mathbf{A}}_\lambda, \hat{\mathbf{B}}_\lambda, \hat{\mathbf{C}}_\lambda, \hat{\boldsymbol{\xi}}) = \text{argmax}_{\mathbf{A},\mathbf{B},\mathbf{C},\boldsymbol{\xi}} \; E^{(t)}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}) - P_\lambda(\mathbf{A}) + P(\mathbf{B}) + P(\mathbf{C}) \tag{4.7}$$

Here, $P_\lambda(\mathbf{A})$ denotes a penalty function on $\mathbf{A}$, $P(\mathbf{B}) = \sum_{j=1}^{J} \log N(b_j|\mu_b, \sigma_b^2)$, and $P(\mathbf{C}) = \sum_{j=1}^{J} \log Beta(c_j|\alpha_c, \beta_c)$ are prior distributions of $\mathbf{B}$ and $\mathbf{C}$, respectively. These priors are incorporated to enhance the stability and robustness of the estimation process. Similar to the regularized GVEM approach for M2PL models, there are two constraints available to ensure model identification in M3PL models.

4.2.4 Other Functions

In addition to GVEM-based methods, the `VEMIRT` package offers parallel analysis to determine the number of factors (Horn, 1965). Parallel analysis involves generating multiple polychoric correlation matrices based on a binary or polytomous real dataset. The average eigenvalues from these random correlation matrices are then compared to the eigenvalues from the correlation matrix of the real data. Factors corresponding to observed eigenvalues greater than the average random eigenvalues are retained (Hayton et al., 2004). This can serve as a prior step for exploratory factor analysis.

4.3 Implementation in the `VEMIRT` *package*

Two simulated datasets, **exampleData_2pl** and **exampleData_3pl**, are employed to showcase the functionality of the main functions for M2PL and M3PL analysis in the `VEMIRT` package. These datasets adhere to the simulation criteria outlined by Cho et al. (2021) and are readily accessible within the package. Tables 4.1 and 4.2 show the true item parameters for both datasets. Specifically, **exampleData_2pl** is generated based on a between-item M2PL model with 5 factors, encompassing responses from 2000 examinees to 75 items. On the other hand, **exampleData_3pl** is simulated based on a within-item M3PL model with 3 factors, comprising responses from 2000 examinees to 45 items. Both datasets contain no missing values, and the true factor correlations are set to 0.1. It is important to note that

the data passed to all estimation functions must only include dichotomous responses, with missing values coded as NA. Table 4.3 summarizes features for two example datasets.

Table 4.1: True item parameters for exampleData_2pl dataset

Item	α_{j1}	α_{j2}	α_{j3}	α_{j4}	α_{j5}	b_j
1	1.496	0	0	0	0	1.122
2	2.091	0	0	0	0	1.462
3	1.479	0	0	0	0	2.142
4	1.343	0	0	0	0	-1.700
5	1.734	0	0	0	0	1.326
6	0.969	0	0	0	0	-3.366
7	2.601	0	0	0	0	2.329
8	1.632	0	0	0	0	1.343
9	2.465	0	0	0	0	-0.969
10	2.261	0	0	0	0	-0.017
11	1.768	0	0	0	0	-0.221
12	2.006	0	0	0	0	-0.459
13	2.227	0	0	0	0	-0.748
14	1.751	0	0	0	0	0.340
15	1.105	0	0	0	0	-2.822
16	0	0.833	0	0	0	-2.975
17	0	1.581	0	0	0	-0.493
18	0	1.394	0	0	0	0.357
19	0	1.581	0	0	0	1.632
20	0	1.632	0	0	0	1.139
21	0	1.989	0	0	0	0.374
22	0	2.499	0	0	0	0.833

23	0	1.938	0	0	0	-1.258
24	0	2.567	0	0	0	0.867
25	0	2.210	0	0	0	-0.102
26	0	1.003	0	0	0	-4.896
27	0	2.108	0	0	0	-1.564
28	0	0.782	0	0	0	0.884
29	0	1.326	0	0	0	-0.731
30	0	1.428	0	0	0	-2.074
31	0	0	1.734	0	0	1.190
32	0	0	2.244	0	0	0.561
33	0	0	1.530	0	0	-1.207
34	0	0	1.190	0	0	-2.601
35	0	0	1.173	0	0	-2.414
36	0	0	1.394	0	0	1.207
37	0	0	1.054	0	0	1.377
38	0	0	1.513	0	0	1.666
39	0	0	1.360	0	0	0.595
40	0	0	1.564	0	0	3.196
41	0	0	3.043	0	0	1.751
42	0	0	3.162	0	0	1.768
43	0	0	1.921	0	0	0.782
44	0	0	1.445	0	0	0.697
45	0	0	3.621	0	0	2.329
46	0	0	0	1.309	0	-0.289
47	0	0	0	1.547	0	0.289
48	0	0	0	1.020	0	-2.023
49	0	0	0	0.510	0	3.961

50	0	0	0	0.986	0	1.309
51	0	0	0	1.156	0	2.244
52	0	0	0	1.156	0	0.017
53	0	0	0	1.003	0	-1.343
54	0	0	0	1.445	0	-1.241
55	0	0	0	2.261	0	0.663
56	0	0	0	1.802	0	0.153
57	0	0	0	1.768	0	-0.612
58	0	0	0	1.683	0	0.799
59	0	0	0	2.057	0	0.204
60	0	0	0	1.768	0	1.717
61	0	0	0	0	2.023	1.122
62	0	0	0	0	1.734	2.771
63	0	0	0	0	2.125	0.680
64	0	0	0	0	2.023	-0.102
65	0	0	0	0	1.156	0.306
66	0	0	0	0	1.190	1.207
67	0	0	0	0	1.717	0.170
68	0	0	0	0	2.958	1.547
69	0	0	0	0	1.411	-0.204
70	0	0	0	0	1.802	0.918
71	0	0	0	0	1.615	0.952
72	0	0	0	0	1.462	2.363
73	0	0	0	0	2.465	4.063
74	0	0	0	0	2.499	-0.561
75	0	0	0	0	1.292	-0.187

Table 4.2: True item parameters for exampleData_3pl dataset

Item	α_{j1}	α_{j2}	α_{j3}	b_j	c_j
1	1.109	0.000	0.000	0.689	0.200
2	1.735	0.000	0.000	0.554	0.200
3	1.261	0.000	0.000	-0.062	0.200
4	1.854	0.000	0.000	-0.306	0.200
5	1.926	0.000	0.000	-0.380	0.200
6	0.807	0.000	0.000	-0.695	0.200
7	1.410	0.000	0.000	-0.208	0.200
8	1.866	0.000	0.000	-1.265	0.200
9	1.439	0.000	0.000	2.169	0.200
10	0.000	1.570	0.000	1.208	0.200
11	0.000	1.636	0.000	-1.123	0.200
12	0.000	1.430	0.000	-0.403	0.200
13	0.000	1.493	0.000	-0.467	0.200
14	0.000	1.111	0.000	0.780	0.200
15	0.000	0.934	0.000	-0.083	0.200
16	0.000	1.954	0.000	0.253	0.200
17	0.000	1.878	0.000	-0.029	0.200
18	0.000	1.613	0.000	-0.043	0.200
19	0.000	0.000	1.332	1.369	0.200
20	0.000	0.000	1.082	-0.226	0.200
21	0.000	0.000	1.822	1.516	0.200
22	0.000	0.000	0.807	-1.549	0.200
23	0.000	0.000	1.303	0.585	0.200
24	0.000	0.000	1.749	0.124	0.200

25	0.000	0.000	0.902	0.216	0.200
26	0.000	0.000	1.451	0.380	0.200
27	0.000	0.000	1.008	-0.502	0.200
28	1.321	1.744	0.000	-0.333	0.200
29	1.946	0.781	0.000	-1.019	0.200
30	1.317	1.347	0.000	-1.072	0.200
31	1.597	1.698	0.000	0.304	0.200
32	1.466	0.000	0.909	0.448	0.200
33	0.879	0.000	1.692	0.053	0.200
34	1.875	0.000	1.869	0.922	0.200
35	1.058	0.000	1.218	2.050	0.200
36	0.000	1.021	1.581	-0.491	0.200
37	0.000	1.148	0.869	-2.309	0.200
38	0.000	1.040	1.230	1.006	0.200
39	0.803	0.929	1.093	-0.709	0.200
40	1.160	1.268	1.768	-0.688	0.200
41	1.943	1.267	1.311	1.026	0.200
42	1.862	1.211	1.763	-0.285	0.200
43	1.616	0.941	1.765	-1.221	0.200
44	1.551	0.924	1.743	0.181	0.200
45	1.993	1.041	1.300	-0.139	0.200

4.3.1 Confirmatory Factor Analysis for the M2PL Model

The `VEMIRT` package has three primary functions for conducting confirmatory factor analysis (CFA) for M2PL models: **gvem_2PLCFA**, **bs_2PLCFA**, **importance Sampling**. The latter two functions, **bs_2PLCFA** and **importance Sampling**, leverage bootstrap or importance sampling techniques, respectively, to rectify any bias inherent in the estimates

Table 4.3: A summary of features for two example datasets

Dataset	Model Structure	Domain	Test Length	Sample Size
exampleData_2pl	Between-item M2PL	5	75	2000
exampleData_3pl	Within-item M3PL	3	45	2000

generated by **gvem_2PLCFA**. Among these, only **gvem_2PLCFA** and **bs_2PLCFA** are capable of producing standard error (SE) estimates for item parameters. The SE estimates are derived using different methodologies: **gvem_2PLCFA** utilizes the supplemented EM (SEM) method (Cai, 2008), while **bs_2PLCFA** employs the bootstrap method.

Since the item factor loading structure is assumed to be known for CFA, an indicator matrix representing this structure must be provided for CFA functions. This indicator matrix has the same dimensions as the item discrimination matrix and comprises binary values (0s and 1s). Specifically, a value of 1 indicates that the item is loaded onto the corresponding factor, while a value of 0 indicates otherwise. By specificying the dataset **exampleData_2pl** and corresponding factor loading indicator matrix **exampleIndic_cfa2pl**, the CFA for a M2PL model using GVEM algorithm is conducted:

```
R> CFA_result <- gvem_2PLCFA(u=exampleData_2pl, indic=exampleIndic_cfa2pl, SE.
   est = TRUE)
```

Here we set `SE.est = TRUE` to estimate SEs of item parameters using the SEM method. By default, this function can be disabled with `SE.est = FALSE`. The object `CFA_result` includes various results, such as `ra` and `rb` for item discrimination and item difficulty estimates, `rsigma` for population variance-covariance matrix estimate, `GIC`, `AIC`, `BIC` for model fit index, and `SE` for SE estimates of item parameters.

The bootstrap method could be implemented to correct bias for item parameters in the object `CFA_result` and calculate SE as well. By executing the following command, we draw

5 bootstrap samples:

```
R> bs_2PLCFA(gvem_result=CFA_result, boots = 5)
```

The output includes corrected item discrimination and item difficulty parameters, along with their standard errors. Alternatively, the IW-GVEM method can be employed to correct bias from GVEM estimation:

```
R> importanceSampling(u=exampleData_2pl,gvem_result=CFA_result)
```

By default, the number of samples D and the number of times to draw samples S are both set to 10, but they can be modified as needed. Except for original outputs in the object `CFA_result`, the object produced by `importanceSampling` function includes corrected item discrimination and item difficulty parameters, the population variance-covariance matrix, the optimal learning rate, and the lower bound value. Table 4.4 compares the bias and root-mean-square deviation (RMSE) results for item parameter recovery, and execution time for each function. Note that all computations were performed on a desktop workstation equipped with a 2.4GHz-Core Intel Xeon CPU, 1TB RAM, and 2133MHz DDR4 memory. It was expected that both GVEM-BS and IW-GVEM yielded more accurate item parameter estimates, albeit with some sacrifice of computation time.

Table 4.4: Item parameter recovery and execution time for each CFA function

Function	Bias of $\boldsymbol{\alpha}_j$	Bias of b_j	RMSE of $\boldsymbol{\alpha}_j$	RMSE of b_j	Execution Time (min)
GVEM	-0.259	-0.050	0.350	0.154	0.372
GVEM-BS	-0.077	-0.001	0.176	0.105	1.570
IW-GVEM	-0.145	-0.017	0.271	0.125	2.583

4.3.2 Exploratory Factor Analysis for the M2PL Model

The same dataset is used to illustrate exploratory factor analysis (EFA) for an M2PL model in the `VEMIRT` package. There are three functions in the package to conduct EFA for the M2PL model: `gvem_2PLEFA_rot` employing GVEM with post-hoc rotation, `gvem_2PLEFA_lasso` utilizing GVEM with Lasso penalty, and `gvem_2PLEFA_adaptlasso` employing GVEM with adaptive Lasso penalty. It should be noted that the current EFA functions do not estimate SEs of item parameters.

`gvem_2PLEFA_rot` applies the promax rotation method by default, but CF-Quartimax rotation is also available by specifying the `rot` input:

```
R> gvem_2PLEFA_rot(u=exampleData_2pl, domain=5)
R> gvem_2PLEFA_rot(u=exampleData_2pl, domain=5, rot= ``cfQ")
```

Except for the outputs similar to those in CFA functions, a distinctive result in EFA compared to CFA is the factor loading structure, denoted as `Q_mat`.

Different from `gvem_2PLEFA_rot`, there are three additional parameters for both GVEM penalty functions (`gvem_2PLEFA_lasso` and `gvem_2PLEFA_adaptlasso`): `constrain`, `indic`, and `non_pen`. The parameter `constrain` determines the constraint setting, which must be specified as either "C1" or "C2" to ensure identifiability. Accordingly, the parameter `indic` imposes constraints on the sub-matrix of the factor loading structure. It matches the dimensions of the item discrimination matrix. Under "C1", a $K \times K$ sub-matrix of `indic` becomes an identity matrix, designating K items that load solely on each factor. Conversely, "C2" sets the sub-matrix as a lower triangular matrix with ones on the diagonal, indicating items associated with each factor, but potentially with others as well. Nonzero entries, excluding diagonal ones, incur penalties during estimation. For example, for $K = 3$, the "C2" constraint requires the following sub-matrix in the indicator matrix: $C2 = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{smallmatrix}\right)$ While the first two items are straightforward to identify in the indicator matrix, the last item appears to be similar to the remaining items that load on every factor. Since the definition

assumes this item must load on the last factor, with others dependent on regularization results, the parameter `non_pen` is necessary to identify this specific item. Although "C2" is weaker than "C1", it ensures empirical identifiability. By default, "C1" is applied. During estimation, both "C1" and "C2" constraints constrain population means and variances to 0 and 1, respectively. Additionally, `gvem_2PLEFA_adaptlasso` needs to specify the tuning parameter `gamma`, which is set to 2 by default. Users can freely change this value although we recommend `gamma`=2 based on our simulation results. Two example `indic` matrices for the dataset `exampleData_2pl` can be found in Tables 4.5 and 4.6. The following commands could be implemented to run GVEM penalty functions:

```
R> gvem_2PLEFA_lasso(u=exampleData_2pl, indic=exampleIndic_efa2pl_c1,
    constrain= ''C1")
R> gvem_2PLEFA_lasso(u=exampleData_2pl, indic=exampleIndic_efa2pl_c2,
    constrain= ''C2", non_pen=61)
R> gvem_2PLEFA_adaptlasso(u=exampleData_2pl, indic=exampleIndic_efa2pl_c1,
    constrain= ''C1", gamma=2)
R> gvem_2PLEFA_adaptlasso(u=exampleData_2pl, indic=exampleIndic_efa2pl_c2,
    constrain= ''C2", non_pen=61, gamma=2)
```

Table 4.5: An example indicator matrix for C1 constraint

Item	F_1	F_2	F_3	F_4	F_5
1	1	0	0	0	0
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1

7	1	1	1	1	1
8	1	1	1	1	1
9	1	1	1	1	1
10	1	1	1	1	1
11	1	1	1	1	1
12	1	1	1	1	1
13	1	1	1	1	1
14	1	1	1	1	1
15	1	1	1	1	1
16	0	1	0	0	0
17	1	1	1	1	1
18	1	1	1	1	1
19	1	1	1	1	1
20	1	1	1	1	1
21	1	1	1	1	1
22	1	1	1	1	1
23	1	1	1	1	1
24	1	1	1	1	1
25	1	1	1	1	1
26	1	1	1	1	1
27	1	1	1	1	1
28	1	1	1	1	1
29	1	1	1	1	1
30	1	1	1	1	1
31	0	0	1	0	0
32	1	1	1	1	1
33	1	1	1	1	1

34	1	1	1	1	1
35	1	1	1	1	1
36	1	1	1	1	1
37	1	1	1	1	1
38	1	1	1	1	1
39	1	1	1	1	1
40	1	1	1	1	1
41	1	1	1	1	1
42	1	1	1	1	1
43	1	1	1	1	1
44	1	1	1	1	1
45	1	1	1	1	1
46	0	0	0	1	0
47	1	1	1	1	1
48	1	1	1	1	1
49	1	1	1	1	1
50	1	1	1	1	1
51	1	1	1	1	1
52	1	1	1	1	1
53	1	1	1	1	1
54	1	1	1	1	1
55	1	1	1	1	1
56	1	1	1	1	1
57	1	1	1	1	1
58	1	1	1	1	1
59	1	1	1	1	1
60	1	1	1	1	1

61	0	0	0	0	1
62	1	1	1	1	1
63	1	1	1	1	1
64	1	1	1	1	1
65	1	1	1	1	1
66	1	1	1	1	1
67	1	1	1	1	1
68	1	1	1	1	1
69	1	1	1	1	1
70	1	1	1	1	1
71	1	1	1	1	1
72	1	1	1	1	1
73	1	1	1	1	1
74	1	1	1	1	1
75	1	1	1	1	1

Table 4.6: An example indicator matrix for C2 constraint

Item	F_1	F_2	F_3	F_4	F_5
1	1	0	0	0	0
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1
7	1	1	1	1	1
8	1	1	1	1	1

9	1	1	1	1	1
10	1	1	1	1	1
11	1	1	1	1	1
12	1	1	1	1	1
13	1	1	1	1	1
14	1	1	1	1	1
15	1	1	1	1	1
16	1	1	0	0	0
17	1	1	1	1	1
18	1	1	1	1	1
19	1	1	1	1	1
20	1	1	1	1	1
21	1	1	1	1	1
22	1	1	1	1	1
23	1	1	1	1	1
24	1	1	1	1	1
25	1	1	1	1	1
26	1	1	1	1	1
27	1	1	1	1	1
28	1	1	1	1	1
29	1	1	1	1	1
30	1	1	1	1	1
31	1	1	1	0	0
32	1	1	1	1	1
33	1	1	1	1	1
34	1	1	1	1	1
35	1	1	1	1	1

36	1	1	1	1	1
37	1	1	1	1	1
38	1	1	1	1	1
39	1	1	1	1	1
40	1	1	1	1	1
41	1	1	1	1	1
42	1	1	1	1	1
43	1	1	1	1	1
44	1	1	1	1	1
45	1	1	1	1	1
46	1	1	1	1	0
47	1	1	1	1	1
48	1	1	1	1	1
49	1	1	1	1	1
50	1	1	1	1	1
51	1	1	1	1	1
52	1	1	1	1	1
53	1	1	1	1	1
54	1	1	1	1	1
55	1	1	1	1	1
56	1	1	1	1	1
57	1	1	1	1	1
58	1	1	1	1	1
59	1	1	1	1	1
60	1	1	1	1	1
61	1	1	1	1	1
62	1	1	1	1	1

63	1	1	1	1	1
64	1	1	1	1	1
65	1	1	1	1	1
66	1	1	1	1	1
67	1	1	1	1	1
68	1	1	1	1	1
69	1	1	1	1	1
70	1	1	1	1	1
71	1	1	1	1	1
72	1	1	1	1	1
73	1	1	1	1	1
74	1	1	1	1	1
75	1	1	1	1	1

The execution time for each EFA method is shown below. Based on the current example dataset, `gvem_2PLEFA_rot` appears to be the most computationally efficient, whereas `gvem_2PLEFA_lasso` with "C2" constraint is the most time-consuming. However, it is important to note that this pattern can vary for different datasets.

Table 4.7: The execution time for each EFA method

Function	Rotation/Constraint	Execution time (min)
`gvem_2PLEFA_rot`	Promax	0.336
`gvem_2PLEFA_lasso`	C1	12.877
`gvem_2PLEFA_lasso`	C2	205.386
`gvem_2PLEFA_adaptlasso`	C1	20.150
`gvem_2PLEFA_rot`	C2	19.271

4.3.3 Stochastic GVEM Methods for the M3PL Model

There are four functions to implement stochastic GVEM methods to conduct EFA and CFA for the M3PL model. : `sgvem_3PLCFA` for confirmatory analysis, `sgvem_3PLEFA_rot` for exploratory analysis using post-hoc rotation, `sgvem_3PLEFA_lasso` for exploratory analysis using Lasso, and `sgvem_3PLEFA_adaptlasso` for exploratory analysis using adaptive Lasso. Here, we use another simulated dataset **exampleData_3pl** to evaluate their performance.

Different from previous GVEM functions, stochastic GVEM methods need to identify the subsample for each iteration and the forget rate for the stochastic algorithm. By default, these values are set as 50 and 0.51, respectively. Additionally, these functions have some parameters to specify priors for item difficulty parameters and guessing parameters to improve the stability and robustness of the parameter estimation process. For instance, we can assume prior distributions of $b_j \sim N(0, 4), c_j \sim beta(10, 40)$, and then conduct the following commands:

```
R> sgvem_3PLCFA(u=exampleData_3pl, indic=exampleIndic_cfa3pl,samp=50,
   forgetrate=0.51,mu_b=0,sigma2_b=4,Alpha=10,Beta=40)
R> sgvem_3PLEFA_rot(u=exampleData_3pl,domain=3,samp=50,forgetrate=0.51,mu_b=0,
   sigma2_b=4,Alpha=10,Beta=40,rot="Promax")
R> sgvem_3PLEFA_lasso(u=exampleData_3pl, indic=exampleIndic_efa3pl_c1,samp=50,
   forgetrate=0.51,mu_b=0,sigma2_b=4,Alpha=10,Beta=40,constrain="C1",non_pen
   =NULL)
R> sgvem_3PLEFA_lasso(u=exampleData_3pl, indic=exampleIndic_efa3pl_c2,samp=50,
   forgetrate=0.51,mu_b=0,sigma2_b=4,Alpha=10,Beta=40,constrain="C2",non_pen
   =19)
R> sgvem_3PLEFA_adaptlasso(u=exampleData_3pl, indic=exampleIndic_efa3pl_c1,
   samp=50,forgetrate=0.51,mu_b=0,sigma2_b=4,Alpha=10,Beta=40,constrain="C1"
   ,non_pen=NULL,gamma=2)
R> sgvem_3PLEFA_adaptlasso(u=exampleData_3pl, indic=exampleIndic_efa3pl_c2,
   samp=50,forgetrate=0.51,mu_b=0,sigma2_b=4,Alpha=10,Beta=40,constrain="C2"
   ,non_pen=19,gamma=2)
```

Table 4.8 presents the execution time for each stochastic GVEM method. In this example, all commands are completed within 11 minutes.

Table 4.8: The execution time for each stochastic GVEM method

Function	Rotation/Constraint	Execution time (min)
`sgvem_3PLCFA`		0.304
`sgvem_3PLEFA_lasso`	Promax	0.706
`sgvem_3PLEFA_lasso`	C1	7.272
`sgvem_3PLEFA_lasso`	C2	10.932
`sgvem_3PLEFA_adaptlasso`	C1	2.316
`sgvem_3PLEFA_adaptlasso`	C2	2.259

4.3.4 Parallel Analysis

To apply the GVEM functions for EFA, the number of factors is required. When the number of factors is unknown, the parallel analysis can be conducted using the command below:

```
R> pa_poly(data=exampleData_3pl, n.iter=5, figure=TRUE)
Parallel analysis suggests that the number of factors =  3
   Actual Data Simulated Data
1   10.0501804       1.7694793
2    3.5085385       1.6775089
3    3.2534002       1.6131262
4    1.3540815       1.5456064
5    1.2194996       1.5058899
6    1.1790974       1.4672360
7    1.1172935       1.4216975
8    1.0983607       1.3962212
9    1.0545311       1.3594145
10   1.0093053       1.3276505
```

The function returns a data frame with the eigenvalues for the real data and the simulated data based on 5 simulated analysis (`n.iter=5`) and here we only present the first 10 rows. By default, `pa_poly` draws an eigenvalue plot, as shown below. If `figure=FALSE`, the graphic output would be suppressed.

4.4 Discussion

As it stands, the `VEMIRT` package emerges as a valuable tool for researchers and practitioners interested in conducting exploratory and confirmatory analysis for high-dimensional MIRT models. It significantly enhances computation efficiency and item parameter recovery accuracy, offering a diverse range of GVEM-based methods such as bootstrap sampling, importance sampling, post-hoc rotation, Lasso penalization, adaptive Lasso penalization, and stochastic optimization. Additionally, the package provides functionality for two different types of standard error methods and incorporates parallel analysis to determine the number of factors for exploratory analysis. Moreover, its capability to handle missing values further extends its utility.

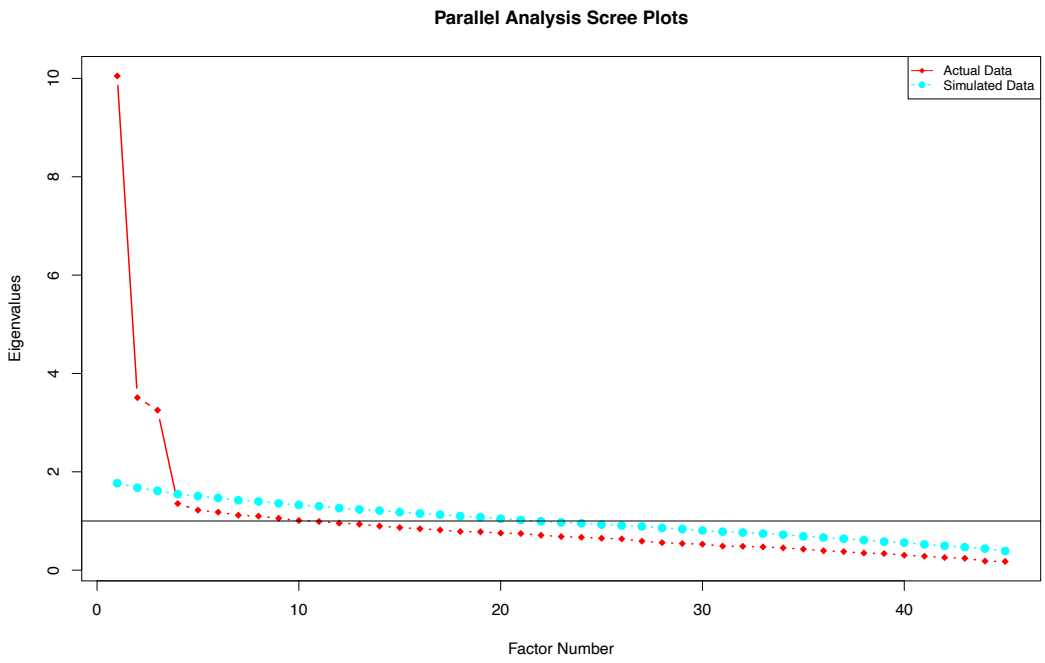

Figure 4.1: Parallel Analysis Scree Plots

The `VEMIRT` package bridges theoretical developments in GVEM with practical implementation in the field of MIRT modeling. By integrating a wide array of GVEM-based methods, it offers flexibility in model specification and estimation, allowing researchers to tailor their analyses to the specific requirements of their research questions. By providing efficient and accurate estimation methods, along with comprehensive results for item analysis and model evaluation, the `VEMIRT` package empowers researchers to conduct sophisticated MIRT analyses with ease and confidence.

Looking ahead, the `VEMIRT` package is poised for further development and enhancement. Some future directions may include: (1) providing standard error estimates for exploratory analysis and M3PL models; (2) enabling users to modify tuning parameters for Lasso penalty

methods; (3) extending the GVEM-based methods to accommodate polytomous data.

Chapter 5

SUMMARY AND DISCUSSION

Large-scale assessments (LSAs), such as PISA and PIAAC, are instrumental in shaping education policies and informing pedagogical practices (Heyneman & Lee, 2014). Multidimensional Item Response Theory (MIRT), with its ability to model multiple latent constructs simultaneously, plays a crucial role in extracting meaningful insights from LSAs. It also offers a robust framework for analyzing complex data in LSAs. Despite its potential, challenges in parameter estimation have hindered the widespread adoption of MIRT, particularly in high-dimensional settings (Andersson & Xin, 2021). To address these challenges, various algorithms, including the Gaussian variational Expectation Maximization (GVEM) algorithm, have been proposed (Cho et al., 2021; Cho et al., 2022). This dissertation aimed to enhance the applicability of the GVEM algorithm within the MIRT framework and investigate its performance across diverse scenarios, thereby advancing parameter estimation methods and facilitating deeper insights into educational and psychological assessments.

The dissertation comprised three main chapters, each addressing specific aspects of the GVEM algorithm and its applications in MIRT modeling. Chapter 2 introduced GVEM-BS, a modified version of GVEM tailored for handling missing data scenarios commonly encountered in LSAs. Through two simulation studies and real data analyses, GVEM-BS showcased its effectiveness in accurately estimating model parameters and robustness under various missing data conditions. Chapter 3 delved into the critical aspect of standard error (SE) estimation within the GVEM framework. By comparing three GVEM-based SE estimation methods (GVEM-BSP, GVEM-BS, GVEM-USEM), the study highlighted GVEM-BSP as the most promising method for SE estimation, offering researchers a reliable tool for assessing the precision of parameter estimates in LSAs. Chapter 4 introduced the `VEMIRT` R

package, designed to facilitate the application of GVEM algorithms for high-dimensional data typically encountered in LSAs. The package equips researchers with efficient computational tools and a diverse array of GVEM-based methods, empowering them to derive actionable insights from complex assessment data.

The implications and significance of this dissertation are manifold, encompassing both methodological advancements and practical applications in the field of LSAs, with a particular focus on MIRT. One of the key highlights is the refinement and adaptation of GVEM-based methods, which offers more accurate and efficient methods for estimating parameters in MIRT models. This advancement addresses the computational challenges associated with high-dimensional data, allowing researchers to derive more precise estimates of latent constructs in LSAs. Moreover, it paves the way for deeper insights into student learning and educational outcomes.

In addition to methodological advancements, the dissertation also offers practical applications through the development of a user-friendly software package, the `VEMIRT` R package. This package lowers the barrier to entry for conducting advanced analyses of LSA data, making sophisticated modeling techniques accessible to a broader audience of researchers and practitioners. These results eventually support evidence-based decision-making and drive improvements in educational policies, interventions, and practices.

Future studies on this topic could be explored in the following directions. First, the current GVEM algorithms and the associated R package can only deal with dichotomous items. However, polytomous items, which have more than two response categories, are commonly used in standardized assessments and cognitive surveys, such as Likert-scale items or constructed response items (Bolt & Adams, 2017). Polytomous items allow researchers and practitioners to assign partial scores for intermediate steps toward correctness, providing more nuanced information (Cui et al., 2024). Incorporating polytomous items into the GVEM framework would significantly broaden its applicability, allowing for a more comprehensive analysis of data that captures finer details from a wider range of assessments. This extension is critical because polytomous items often reflect more complex constructs

and offer a richer understanding of examinee behavior. By enabling the GVEM algorithm to work with polytomous data, researchers could conduct analyses across a broader spectrum of assessments, leading to more accurate estimations and deeper insights into the latent constructs underlying large-scale assessments. This development would ultimately enhance the robustness of the GVEM algorithm and expand its utility for a broader range of educational and psychological studies.

Second, exploring the performance of the GVEM algorithm and its variants under a more complex missing data mechanism, specifically missing not at random (MNAR), could yield valuable insights into their robustness and reliability in real-world assessment contexts. MNAR instances are common in large-scale assessments (LSAs), occurring when the probability of missingness depends on unobserved data or underlying factors, which pose unique challenges to scaling and item calibration (Ulitzsch, 2020). Assessing the effectiveness of the GVEM algorithm in such conditions is crucial because MNAR can lead to biased estimates and unreliable results. Further exploration in this area would contribute to a more comprehensive understanding of how the GVEM framework performs in the presence of complex missing data patterns, thereby informing best practices for handling MNAR scenarios.

Third, further investigation into the GVEM-based standard error (SE) estimation methods is warranted. The current study examined two SE estimation procedures, but additional methods could be explored to enhance the precision and accuracy of SE estimates. For example, future studies can incorporate the XPD method into the GVEM framework and evaluate its estimation performance. This exploration could also include the development of new SE estimation procedures specifically designed to address challenges posed by missing data. By improving SE estimation accuracy, researchers and practitioners could make more reliable inferences from their analyses, ultimately leading to more robust conclusions.

Lastly, future research could focus on evaluating the performance of GVEM-based methods across a range of real-world assessment designs and datasets, including those with complex multidimensional structures and large item pools. This comprehensive evaluation would involve varying factors such as the number of dimensions, item types, and test structures to

assess the adaptability of the GVEM algorithm. By conducting such analyses, researchers can better understand the algorithm's strengths and limitations, identify the optimal conditions for its application, and determine where further refinements or modifications might be necessary. These efforts would contribute to a more nuanced understanding of the GVEM framework and its broader application in educational and psychological assessments.

In summary, this dissertation advances the field of LSAs by enhancing the applicability of MIRT through methodological refinements and practical software development. By addressing gaps in the existing literature and extending the applicability of the GVEM algorithm, this research lays the groundwork for future advancements in the field of MIRT modeling. The findings have implications for both research and practice, with potential to inform policy decisions and improve assessment practices in education and psychology. As the field continues to evolve, further research is needed to enhance the robustness, efficiency, and applicability of GVEM-based methods in diverse contexts. Through continued research and innovation, the potential for leveraging MIRT in LSAs to drive positive educational outcomes remains promising.

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*(3), 251–269.

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The 1996 NAEP technical report*. ERIC.

Andersson, B., Jin, S., & Zhang, M. (2023). Fast estimation of multiple group generalized linear latent variable models for categorical observed variables. *Computational Statistics & Data Analysis*, *182*, 107710. https://doi.org/10.1016/j.csda.2023.107710

Andersson, B., & Xin, T. (2021). Estimation of latent regression item response theory models using a second-order Laplace approximation. *Journal of Educational and Behavioral Statistics*, *46*(2), 244–265. https://doi.org/10.3102/1076998620945199

Asparouhov, T., & Muthén, B. (2012). Comparison of computational methods for high dimensional item factor analysis. *Unpublished manuscript retrieved from www. statmodel. com*.

Asparouhov, T., & Muthén, B. (2016). IRT in Mplus. *Technical Appendix. Los Angeles, CA: Muthén & Muthén*.

Asparouhov, T., & Muthén, B. (2023). Penalized structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–26.

Bashkov, B. M., & DeMars, C. E. (2017). Examining the performance of the Metropolis–Hastings Robbins–Monro algorithm in the estimation of multilevel multidimensional IRT models. *Applied Psychological Measurement*, *41*(5), 323–337. https://doi.org/10.1177/0146621616688923

Bentler, P. M. (1995). *EQS structural equations program manual* (Vol. 6). Multivariate Software Encino, CA.

Bianconcini, S., & Cagnone, S. (2012). Estimation of generalized linear latent variable models via fully exponential Laplace approximation. *Journal of Multivariate Analysis*, *112*, 183–193. https://doi.org/10.1016/j.jmva.2012.06.005

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

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

Bolt, D. M., & Adams, D. J. (2017). Exploring rubric-related multidimensionality in polytomously scored test items. *Applied Psychological Measurement*, *41*(3), 163–177. https://doi.org/10.1177/0146621616677715

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Bürkner, P.-C. (2019). Bayesian item response modeling in R with brms and stan. *arXiv preprint arXiv:1905.09501*.

Cagnone, S., & Monari, P. (2013). Latent variable models for ordinal data by using the adaptive quadrature approximation. *Computational Statistics*, *28*(2), 597–619. https://doi.org/10.1007/s00180-012-0319-z

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 309–329. https://doi.org/10.1348/000711007X249603

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33–57. https://doi.org/10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335. https://doi.org/10.3102/1076998609353115

Cai, L. (2010c). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612. https://doi.org/10.1007/s11336-010-9178-0

Cai, L. (2013). Flexmirt version 2: Flexible multilevel multidimensional item analysis and test scoring [computer software]. *Chapel Hill, NC: Vector Psychometric Group*.

Cai, L., Du Toit, S., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [computer software]. *Chicago, IL: Scientific Software International*.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 245–276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2p tables. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 173–194. https://doi.org/10.1348/000711005X66419

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221. https://doi.org/10.1037/a0023350

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*. https://doi.org/10.18637/jss.v076.i01

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(1), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P., Pek, J., & Liu, Y. (2017). Profile-likelihood confidence intervals in item response theory models. *Multivariate Behavioral Research*, *52*(5), 533–550. https://doi.org/10.1080/00273171.2017.1329082

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*(1), 1–20. https://doi.org/10.1007/s11336-014-9401-5

Chen, P. (2017). A comparative study of online item calibration methods in multidimensional computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *42*(5), 559–590. https://doi.org/10.3102/1076998617695098

Chen, P., & Wang, C. (2021). Using EM algorithm for finite mixtures and reformed supplemented EM for MIRT calibration. *Psychometrika*, *86*(1), 299–326. https://doi.org/10.1007/s11336-021-09745-6

Chen, Y., Li, X., & Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, *84*, 124–146. https://doi.org/10.1007/s11336-018-9646-5

Cho, A. E., Wang, C., Zhang, X., & Xu, G. (2021). Gaussian variational estimation for multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *74*, 52–85. https://doi.org/10.1111/bmsp.12219

Cho, A. E., Xiao, J., Wang, C., & Xu, G. (2022). Regularized variational estimation for exploratory item factor analysis. *Psychometrika*, 1–29. https://doi.org/10.1007/s11336-022-09874-6

Converse, G., Curi, M., Oliveira, S., & Templin, J. (2021). Estimation of multidimensional item response theory models with correlated latent variables using variational autoencoders. *Machine Learning*, *110*(6), 1463–1480.

Coskun, A., Ceyhan, E., Inal, T. C., Serteser, M., & Unsal, I. (2013). The comparison of parametric and nonparametric bootstrap methods for reference interval computation in small sample size groups. *Accreditation and Quality Assurance*, *18*, 51–60.

Cui, C., Wang, C., & Xu, G. (2024). Variational estimation for multidimensional generalized partial credit model. *Psychometrika*, 1–29.

Curi, M., Converse, G. A., Hajewski, J., & Oliveira, S. (2019). Interpretable variational autoencoders for cognitive models. *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

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

measurement, *38*(3), 213–234. https://doi.org/10.1111/j.1745-3984.2001.tb01124.x

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*(3), 295–311. https://doi.org/10.3102/10769986030003295

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 94–128.

Diebolt, J., & Ip, E. H. (1995). Stochastic EM: Method and application. *Markov chain Monte Carlo in practice*, 259.

Edwards, J. M., & Finch, W. H. (2018). Recursive partitioning methods for data imputation in the context of item response theory: A Monte Carlo simulation. *Psicológica*, *39*(1), 88–117. https://doi.org/10.2478/psicolj-2018-0005

Edwards, M. C. (2010). A markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*(3), 474–497. https://doi.org/10.1007/s11336-010-9161-9

Eekhout, I., Enders, C. K., Twisk, J. W., de Boer, M. R., de Vet, H. C., & Heymans, M. W. (2015). Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 588–602. https://doi.org/10.1080/10705511.2014.937670

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54–75. https://doi.org/10.1214/ss/1177013815

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*(3), 430–457. https://doi.org/10.1207/S15328007SEM0803_5

Falk, C. F., & Monroe, S. (2018). On lagrange multiplier tests in multidimensional item response theory: Information matrices and model misspecification. *Educational and Psychological Measurement*, *78*(4), 653–678. https://doi.org/10.1177/0013164417714506

Fan, Y., & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *75*(3), 531–552.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225–245. https://doi.org/10.1111/j.1745-3984.2008.00062.x

Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education*, *14*(1), 31–57. https://doi.org/10.1207/S15324818AME1401_04

Garnier-Villarreal, M., Merkle, E. C., & Magnus, B. E. (2021). Between-item multidimensional IRT: How far can the estimation methods go? *Psych*, *3*(3), 404–421. https://doi.org/10.3390/psych3030029

Gonçalves, S., & White, H. (2005). Bootstrap standard error estimates for linear regression. *Journal of the American Statistical Association*, *100*(471), 970–979. https://doi.org/10.1198/016214504000002087

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323. https://doi.org/10.1037/1082-989X.11.4.323

Gurkan, G. (2021). *From OLS to multilevel multidimensional mixture IRT: A model refinement approach to investigating patterns of relationships in PISA 2012 data* (Doctoral dissertation). Boston College.

Han, K. (T., & Paek, I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement*, *38*(6), 486–498. https://doi.org/10.1177/0146621614536770

Hasan, M., Deng, L. Y., Sabatini, J., Bowman, D., Yang, C.-C., & Hollander, J. (2022). Effect of Q-matrix misspecification on variational autoencoders (VAE) for multidimensional item response theory (MIRT) models estimation. *Proceedings of the 15th International Conference on Educational Data Mining*, 811.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205.

Hendrickson, A. E., & White, P. O. (1966). A method for the rotation of higher-order factors. *British Journal of Mathematical and Statistical Psychology*, *19*(1), 97–103.

Heyneman, S. P., & Lee, B. (2014). The impact of international studies of academic achievement on policy and research. *Handbook of International Large-scale Assessment: Background, Technical Issues and Methods of Data Analysis*, 37–72.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1–17. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Houts, C. R., & Cai, L. (2013). Flexmirt: Flexible multilevel multidimensional item analysis and test scoring user's manual version 2.0.

Huang, B. H., & Flores, B. B. (2018). The English language proficiency assessment for the 21st century (ELPA21). *Language Assessment Quarterly*, *15*(4), 433–442. https://doi.org/10.1080/15434303.2018.1549241

Imbriano, P. (2018). *Methods for improving efficiency of planned missing data designs* (Doctoral dissertation).

Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(2), 257–270.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2017). A variational maximization–maximization algorithm for generalized linear mixed models with crossed random effects. *Psychometrika*, *82*, 693–716. https://doi.org/10.1007/s11336-017-9555-z

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, *52*(12), 5066–5074. https://doi.org/10.1016/j.csda.2008.05.002

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, *37*(2), 183–233. https://doi.org/10.1023/A:1007665907178

Junker, B. W. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment. *Prepared for the National Research Council Committee on the Foundations of Assessment. Retrieved April*, *2*, 2001.

Kalkan, Ö. K., Yusuf, K., & Kelecioğlu, H. (2018). Evaluating performance of missing data imputation methods in IRT analyses. *International Journal of Assessment Tools in Education*, *5*(3), 403–416. https://doi.org/10.21449/ijate.430720

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory

relationships. *Journal of Educational Measurement*, *54*(4), 397–419. https://doi.org/10.1111/jedm.12154

Lee, S.-Y., & Song, X.-Y. (2012). *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. John Wiley & Sons.

Lin, Z. (2018). *The comparison of standard error methods in the marginal maximum likelihood estimation of the two-parameter logistic item response model when the distribution of the latent trait is nonnormal* (Doctoral dissertation). The Florida State University.

Lindstrom, M. J., & Bates, D. M. (1988). Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, *83*(404), 1014–1022. https://doi.org/10.1080/01621459.1988.10478693

Liou, M., & Yu, L.-C. (1991). Assessing statistical accuracy in ability estimation: A bootstrap approach. *Psychometrika*, *56*(1), 55–67. https://doi.org/10.1007/BF02294585

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, *39*(2), 151–162. https://doi.org/10.1093/jpepsy/jst048

Liu, T., Wang, C., & Xu, G. (2022). Estimating three-and four-parameter MIRT models with importance-weighted sampling enhanced variational auto-encoder. *Frontiers in Psychology*, *13*, 935419. https://doi.org/10.3389/fpsyg.2022.935419

Liu, Y., Schulz, E. M., & Yu, L. (2008). Standard error estimation of 3PL IRT true score equating with an MCMC method. *Journal of Educational and Behavioral Statistics*, *33*(3), 257–278. https://doi.org/10.3102/1076998607306076

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, *28*(4), 989–1020.

Luo, X., & Luo, M. X. (2017). Package'xxirt'.

Ma, C., Ouyang, J., Wang, C., & Xu, G. (2023). A note on improving variational estimation for multidimensional item response theory. *Psychometrika*, 1–33. https://doi.org/10.1007/s11336-023-09939-0

Magis, D., Yan, D., & Von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer.

Mair, P., & Hatzinger, R. (2016). CRAN task view: Psychometric models and methods.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). Mcmcpack: Markov chain monte carlo in R.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, *92*(437), 162–170.

McLachlan, G. J., & Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons.

Meng, X.-L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*(416), 899–909. https://doi.org/10.1080/01621459.1991.10475130

Mislevy, R. J., & Wu, P.-K. (1988). Inferring examinee ability when some item responses are missing. *ETS Research Report Series*, *1988*(2), i–75.

Mislevy, R. J., & Wu, P.-K. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series*, *1996*(2), i–36. https://doi.org/10.1002/j.2333-8504.1996.tb01708.x

Muraki, E., & Muraki, M. (2016). Generalized partial credit model. In *Handbook of item response theory* (pp. 127–137). Chapman; Hall/CRC.

Muthén, B., & Muthén, B. O. (2009). *Statistical analysis with latent variables* (Vol. 123). Wiley New York.

Muthén, B., & Muthén, L. (2017). Mplus. In *Handbook of item response theory* (pp. 507–518). Chapman; Hall/CRC.

Naylor, J. C., & Smith, A. F. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *31*(3), 214–225.

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Volume 1 theory of statistics* (pp. 697–716). University of California Press.

Paek, I., & Cai, L. (2014). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional item response theory modeling. *Educational and Psychological Measurement*, *74*(1), 58–76. https://doi.org/10.1177/0013164413500277

Paek, I., Cui, M., Öztürk Gübeş, N., & Yang, Y. (2018). Estimation of an IRT model by Mplus for dichotomously scored responses under different estimation methods. *Educational and Psychological Measurement*, *78*(4), 569–588. https://doi.org/10.1177/0013164417715738

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. https://doi.org/10.1177/0013164413511083

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146–178.

Pfeffermann, D., & Correa, S. (2012). Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. *Biometrika*, *99*(2), 457–472. https://doi.org/10.1093/biomet/ass010

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*(1), 12–35.

R Core Team, R., et al. (2013). R: A language and environment for statistical computing.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., Solenberger, P., et al. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–96.

Ray, A., Margaret, W., et al. (2003). *PISA 2000 technical report*. OECD Publishing.

Reckase, M. D. (2009). Historical background for multidimensional item response theory (MIRT). *Multidimensional Item Response Theory*, 57–77.

Rijmen, F., & Jeon, M. (2013). Fitting an item response theory model with random item effects across groups by a variational approximation method. *Annals of Operations Research*, *206*, 647–662. https://doi.org/10.1007/s10479-012-1181-7

Rose, N., Von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series*, *2010*(1), i–53.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, *47*(3), 537–560.

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

Samejima, F. (1976). Graded response model of the latent trait theory and tailored testing. *Proceedings of the First Conference on Computerized Adaptive Testing*, 5–17.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555. https://doi.org/10.1007/s11336-003-1141-x

Sims, T. (2017). *Comparison of IRTPRO 3 and Mplus 7 for multidimensional item response item parameter and examinee ability estimation* (Doctoral dissertation).

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*(4), 375–394. https://doi.org/10.1111/j.1745-3984.2005.00021.x

Sinharay, S. (2021). Score reporting for examinees with incomplete data on large-scale educational assessments. *Educational Measurement: Issues and Practice*, *40*(1), 79–91. https://doi.org/10.1111/emip.12396

Thissen, D., Cai, L., & Bock, R. D. (2011). The nominal categories item response model. In *Handbook of polytomous item response theory models* (pp. 43–75). Routledge.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397–412.

Tian, W., Cai, L., Thissen, D., & Xin, T. (2013). Numerical differentiation methods for computing error covariance matrices in item response theory modeling: An evaluation and a new proposal. *Educational and Psychological Measurement*, *73*(3), 412–439. https://doi.org/10.1177/0013164412465875

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82–86. https://doi.org/10.1080/01621459.1986.10478240

Tsai, T.-H., Hanson, B. A., Kolen, M. J., Forsyth, & A, R. (2001). A comparison of bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design. *Applied Measurement in Education*, *14*(1), 17–30. https://doi.org/10.1207/S15324818AME1401_03

Ulitzsch, E. (2020). *Using response times for modeling missing responses in large-scale assessments* (Doctoral dissertation). Freie Universitaet Berlin (Germany).

Ulitzsch, E., & Nestler, S. (2022). Evaluating Stan's variational Bayes algorithm for estimating multidimensional IRT models. *Psych*, *4*(1), 73–88. https://doi.org/10.3390/psych4010007

Urban, C. J., & Bauer, D. J. (2021). A deep learning algorithm for high-dimensional exploratory item factor analysis. *Psychometrika*, *86*(1), 1–29. https://doi.org/10.1007/s11336-021-09748-3

Van Buuren, S., & Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. Leiden: TNO.

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

van der Linden, W. J., & Ren, H. (2015). Optimal Bayesian adaptive design for test-item calibration. *Psychometrika*, *80*(2), 263–288. https://doi.org/10.1007/s11336-013-9391-8

von Davier, M. (2009). Is there need for the 3PL model? guess what?

Wald, A. (1941). Asymptotically most powerful tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *12*(1), 1–19.

Wang, C., Chen, P., & Jiang, S. (2020). Item calibration methods with multiple subscale multistage testing. *Journal of Educational Measurement*, *57*(1), 3–28. https://doi.org/10.1111/jedm.12241

Wang, C., & Zhang, X. (2019). A note on the conversion of item parameters standard errors. *Multivariate Behavioral Research*, *54*(2), 307–321. https://doi.org/10.1080/00273171.2018.1513829

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*(1), 116. https://doi.org/10.1037/1082-989X.9.1.116

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 1–25.

Woods, C. M., Cai, L., & Wang, M. (2013). The langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532–547. https://doi.org/10.1177/0013164412464875

Xiao, J., & Bulut, O. (2020). Evaluating the performances of missing data handling methods in ability estimation from sparse data. *Educational and Psychological Measurement*, *80*(5), 932–954. https://doi.org/10.1177/0013164420911136

Yamamoto, K., Khorramdel, L., & Shin, H. J. (2018). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling*, *60*(3), 347–368.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika*, *77*(3), 495–523. https://doi.org/10.1007/s11336-012-9265-5

Yuan, K.-H., Cheng, Y., & Patton, J. (2014). Information matrices and standard errors for mles of item parameters in irt. *Psychometrika*, *79*, 232–254. https://doi.org/10.1007/s11336-013-9334-4

Zhang, H., Chen, Y., & Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, *85*, 358–372. https://doi.org/10.1007/s11336-020-09704-7

Zhang, X., Wang, C., Weiss, D. J., & Tao, J. (2021). Bayesian inference for IRT models with non-normal latent trait distributions. *Multivariate Behavioral Research*, *56*(5), 703–723. https://doi.org/10.1080/00273171.2020.1776096

Zhang, Z., & Zhao, M. (2019). Standard errors of IRT parameter scale transformation coefficients: Comparison of bootstrap method, delta method, and multiple imputation method. *Journal of Educational Measurement*, *56*(2), 302–330. https://doi.org/10.1111/jedm.12210

Zhu, J., Ge, Z., Song, Z., & Gao, F. (2018). Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. *Annual Reviews in Control*, *46*, 107–133. https://doi.org/10.1016/j.arcontrol.2018.09.003