PSYCHOMETRIKA—VOL. 89, NO. 3, 929–957
SEPTEMBER 2024
https://doi.org/10.1007/s11336-024-09955-8

# VARIATIONAL ESTIMATION FOR MULTIDIMENSIONAL GENERALIZED PARTIAL CREDIT MODEL

CHENGYU CUI

UNIVERSITY OF MICHIGAN

CHUN WANG

UNIVERSITY OF WASHINGTON

GONGJUN XU

UNIVERSITY OF MICHIGAN

Multidimensional item response theory (MIRT) models have generated increasing interest in the psychometrics literature. Efficient approaches for estimating MIRT models with dichotomous responses have been developed, but constructing an equally efficient and robust algorithm for polytomous models has received limited attention. To address this gap, this paper presents a novel Gaussian variational estimation algorithm for the multidimensional generalized partial credit model. The proposed algorithm demonstrates both fast and accurate performance, as illustrated through a series of simulation studies and two real data analyses.

Key words: marginal maximum likelihood estimation, variational method, multidimensional item response theory, expectation-maximization algorithm.

There are a wide range of psychometric models for analyzing data in educational and psychological surveys. Models including discrete and continuous latent factors have received great attention due to repeated empirical evidence of adequate model fit and success of interpretation that aligns with substantive theory. Among them, a variety of multidimensional item response theory (MIRT) models (Reckase, 2009) have been proposed to account for various multidimensional structures of the latent constructs. Within the family of MIRT models, one of the most studied models is the multidimensional two-parameter logistic model (M2PL) (Reckase, 2009) for dichotomous response, as well as the multidimensional three-parameter logistic model (M3PL) and Multidimensional Four-parameter Logistic Model (M4PL), which are often used when students can guess the answer to the test item correctly (resulting in a lower asymptote in educational measurement) or when the chance of answering an item correctly does not approach 1 (resulting in a higher asymptote in psychopathology measurement). Moreover, the multidimensional graded response model (Cai, 2010) and multidimensional (generalized) partial credit model (Yao & Schwarz, 2006) have also been proposed to handle polytomous items. These models are regarded as extensions of IRT models for dichotomous response in various ways to characterize latent cognitive structures from more complicated datasets.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11336-024-09955-8.

Correspondence should be made to Chun Wang, College of Education, University of Washington, 312 E Miller Hall, 2012 Skagit Lane, Seattle, WA98105, USA. Email: wang4066@uw.edu

Correspondence should be made to Gongjun Xu, Department of Statistics, University of Michigan, 456 West Hall, 1085 South University, Ann Arbor, MI48109, USA. Email: gongjun@umich.edu

In the IRT literature, the marginal maximum likelihood estimator (MMLE) is considered to be a consistent and efficient approach for parameter estimation by maximizing the marginal log-likelihood of item parameters (Bock & Aitkin, 1981; Bock et al., 1988). In particular, this estimator is efficient in the asymptotic regime where the test length is small with a rather large number of examinees, which is often the case in applications. However, integrating out the latent ability for marginal likelihood evaluation is notoriously time-consuming, involving multidimensional integrals whose computational complexity grows exponentially with the number of latent dimensions. This results in great difficulties especially when the latent structure is complex or within a domain of large dimension. Several methods have been proposed previously to deal with this computational challenge, such as Gaussian quadrature methods (Bock & Aitkin, 1981; Schilling & Bock, 2005), Laplace approximation methods (Lindstrom & Bates, 1988; Wolfinger & O'connell, 1993; Andersson & Xin, 2021), Metropolis-Hastings Robbins-Monro algorithms (Cai, 2010), stochastic expectation maximization (EM) algorithms (von Davier & Sinharay, 2010; Zhang et al., 2020), and variational approximation methods (Rijmen & Jeon, 2013; Cho et al., 2021; 2022). Among these, Gauss-Hermite quadrature approximation does not scale well to high-dimensional scenarios, and Laplace approximation, which is closely related to variational method (Opper & Archambeau, 2009), may suffer from numerical inaccuracies when dimensions get high or when the likelihood function is in a skewed shape. Additionally, many variants of Laplace approximations, though overcoming some deficiencies, suffer from inflexibility or may be hard to implement (Ormerod & Wand, 2010). The methods based on Monte Carlo simulations such as Metropolis-Hastings Robbins-Monro and stochastic EM algorithms, on the other hand, are more robust but may be computationally inefficient. Notably, in addition to being numerically accurate and computationally efficient, the variational method also provides good interpretability and the variational distributions contain additional useful information (Blei et al., 2017).

Despite the extensive literature on estimation methods for models of dichotomous response (Cai, 2010; Chen et al., 2019; Cho et al., 2021; Feuerstahler & Waller, 2014; Meng et al., 2020), little attention has been given to the efficient and robust estimation of MIRT models for polytomous responses (Bock et al., 1988; Kim & Wilson, 2020). In the paper, we propose a Gaussian variational expectation-maximization algorithm for the multidimensional generalized partial credit model (MGPCM), which is both computationally efficient and numerically stable. The variational method, first proposed in computer science and statistical learning, has since become an efficient approach for large-scale computation in fields such as pattern recognition and document retrieval (Titterington, 2004; Blei & Jordan, 2006). The application of the variational method in analyzing statistical models has also received wide attention (Blei et al., 2017; Ormerod & Wand, 2010). For instance, Ormerod and Wand (2012) adopted a Gaussian variational approximation approach for the estimation of generalized linear mixed effects models. In the field of psychometrics and educational measurement, variational methods have been used for efficient estimation of multidimensional 1PL models (Rijmen & Jeon, 2013) and multidimensional 2/3PL models (Cho et al., 2021), as well as analysis in cognitive diagnostic models (Yamaguchi & Okada, 2020; Oka & Okada 2023). Motivated by Cho et al. (2021), in this paper, we generalize their method to the estimation of MGPCM. This generalization is nontrivial because the MGPCM uses a generalized logit link function that requires a completely new derivation of the variational lower bound to fully enable closed-form parameter update in the EM algorithm.

The rest of the paper is organized as follows. Section 1 provides a brief introduction to the model. In Sect. 2, we introduce the polytomous Gaussian variational expectation-maximization algorithm (pGVEM) and give its derivation. Section 3 evaluates the performance of the pGVEM algorithm compared to the traditional EM algorithm and other existing methods through a comprehensive simulation study. In Sect. 4, we apply the pGVEM algorithm to analyze data from an international educational assessment database and a Big-Five personality assessment. Finally, in Sect. 5, we conclude the paper and suggest potential future research directions.

## 1. Model Setting

Generalized partial credit model (GPCM) (Muraki, 1992; Embretson & Reise, 2013), also named as a compensatory multidimensional two-parameter partial credit model (M-2PPC) (Yao & Schwarz, 2006), is a popular IRT model for polytomous responses. It allows for the assessment of partial scores for constructed response items and intermediate steps that students have accomplished on the path toward solving the items completely. Suppose we have $N$ examinees and $J$ test items, with the random variable $Y_{ij}$ denoting the partial credit of person $i$'s response to item $j$. For each item $j$, $a_j$ is a discrimination parameter that implies the strength of association between latent trait and item responses. $\beta_{jk}$ is a threshold parameter that separates two adjacent response categories. The partial credit model (Masters, 1982) is a special case of GPCM where the discrimination parameter is fixed to be the same across different items. The item response function of GPCM that characterizes the probability of a specific response is given by

$$\Pr(Y_{ij} = k|\theta_i, a_j, \beta_{jk}) = \frac{\exp\left[\sum_{r=0}^{k} a_j(\theta_i - \beta_{jr})\right]}{\sum_{v=0}^{K_j} \exp\left[\sum_{r=0}^{v} a_j(\theta_i - \beta_{jr})\right]}, \tag{1.1}$$

where $k = 0, 1, \ldots, K_j - 1$ and $K_j$ is the number of differential partial credit scores for the $j^{th}$ item.

The multidimensional generalized partial credit model (MGPCM) is a natural multidimensional extension of GPCM. The main idea is replacing the uni-dimensional latent ability with a $D$-dimensional vector, and each dimension represents a facet of the multidimensional construct (i.e., science and literacy). Similarly, the discrimination parameters $\boldsymbol{a}_j$ also become $D$-dimensional vectors to reflect the discrimination power of item $j$ with respect to each facet (i.e., dimension) of the multidimensional construct $\boldsymbol{\theta}$. The threshold parameters stay the same as in uni-dimensional models. Equation (1.1) therefore is updated as follows:

$$\Pr(Y_{ij} = k|\boldsymbol{\theta}_i, \boldsymbol{a}_j, \beta_{jk}) = \frac{\exp\{\sum_{r=0}^{k}(\boldsymbol{a}_j'\boldsymbol{\theta}_i - \beta_{jr})\}}{\sum_{v=0}^{K_j-1} \exp\left\{\sum_{r=0}^{v}(\boldsymbol{a}_j'\boldsymbol{\theta}_i - \beta_{jr})\right\}}. \tag{1.2}$$

Here $\boldsymbol{a}_j'\boldsymbol{\theta}_i$ indicates the inner product of $\boldsymbol{a}_j$ and $\boldsymbol{\theta}_i$ as $\boldsymbol{a}_j'\boldsymbol{\theta}_i = \sum_{d=1}^{D} a_{jd}\theta_{id}$ where $a_{jd}$ and $\theta_{id}$ are the $d$th component of $\boldsymbol{a}_j$ and $\boldsymbol{\theta}_i$, respectively. With a slight re-parameterization, we have the following item response function for MGPCM which we will use throughout the paper:

$$\Pr(Y_{ij} = k|\boldsymbol{\theta}_i, \boldsymbol{a}_j, b_{jk}) = \frac{\exp(k\boldsymbol{a}_j'\boldsymbol{\theta}_i - b_{jk})}{\sum_{v=0}^{K_j-1} \exp(v\boldsymbol{a}_j'\boldsymbol{\theta}_i - b_{jv})}. \tag{1.3}$$

In Equation (1.3), $b_{jk}$ replaces $\sum_{r=0}^{k} \beta_{jr}$ in Equation (1.2) for each $k = 0, 1, \ldots, K_j - 1$. Note that for model identification, we can only have $K_j - 1$ estimable threshold parameters, and hence we fix $b_{j0} = 0$.

## 2. Gaussian Variational Approximation

### 2.1. Derivation of Algorithm

In this section, we describe the derivation of the GVEM algorithm. In the following, we denote the collection of item parameters by $M_p$ being the total number of parameters to be estimated,

i.e., $M_p = \{\boldsymbol{a}_j \in \mathbb{R}^D, b_{jk} \in \mathbb{R} : j = 1, \ldots, J, k = 1, \ldots, K_j - 1\}$ for MGPCM. As we discussed above, the parameter $b_{j0}$ is fixed as 0 for all $j$. To be consistent with the common convention, we assume the latent vector $\boldsymbol{\theta}$ follows a multivariate normal distribution of $\mathbf{0}$ mean and covariance $\boldsymbol{\Sigma}_\theta$ with density function denoted by $\phi(\cdot)$. The marginal probability of response vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$ for person $i$ is defined as follows

$$\begin{aligned}\Pr(\boldsymbol{Y}_i|M_p) &= \int_{\boldsymbol{\theta}_i} \prod_{j=1}^{J} \prod_{k=0}^{K_j-1} I_{(Y_{ij}=k)} \Pr(Y_{ij} = k|\boldsymbol{\theta}_i, M_p)\phi(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \\ &= \int_{\boldsymbol{\theta}_i} \prod_{j=1}^{J} \prod_{k=0}^{K_j-1} I_{(Y_{ij}=k)} \frac{\exp(k\boldsymbol{a}_j'\boldsymbol{\theta}_i - b_{jk})}{\sum_{v=0}^{K_j-1} \exp(v\boldsymbol{a}_j'\boldsymbol{\theta}_i - b_{jv})}\phi(\boldsymbol{\theta})d\boldsymbol{\theta}_i,\end{aligned}$$

where $I_{(Y_{ij}=k)}$ is an indicator function equal to 1 if $Y_{ij} = k$ and zero otherwise, and therefore the marginal log-likelihood function for all responses from the examinees is given by

$$l(M_p|\boldsymbol{Y}) = \sum_{i=1}^{N} \log \Pr(\boldsymbol{Y}_i|M_p) = \sum_{i=1}^{N} \log \int_{\boldsymbol{\theta}_i} \prod_{j=1}^{J} \Pr(Y_{ij}|\boldsymbol{\theta}_i, M_p)\phi(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i, \tag{2.1}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)'$ is the $N \times J$ matrix of realized categorical responses.

Following the variational estimation literature (Blei et al., 2017; Cho et al., 2021), we first derive a variational lower bound for MGPCM. We define $KL\{p(\cdot)\|q(\cdot)\}$ as the Kullback-Leibler divergence of probability distribution $p$ and $q$. For an arbitrary probability density function $q(\cdot)$, the marginal log-likelihood in Equation (2.1) has the following lower bound (Blei et al., 2017):

$$\begin{aligned}l(M_p|\boldsymbol{Y}) &= \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} q_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \log \Pr(\boldsymbol{Y}_i|M_p) \\ &= \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log\Big[\frac{P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i|M_p)q_i(\boldsymbol{\theta}_i)}{P(\boldsymbol{\theta}_i|\boldsymbol{Y}_i, M_p)q_i(\boldsymbol{\theta}_i)}\Big]q_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \\ &= \sum_{i=1}^{N} \Big[\int_{\boldsymbol{\theta}_i} \big[\log P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i|M_p)\big]q_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \\ &\quad - \int_{\boldsymbol{\theta}_i} \big[\log q_i(\boldsymbol{\theta}_i)\big]q_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i + KL\Big\{q_i(\boldsymbol{\theta}_i)\Big\|P(\boldsymbol{\theta}_i|\boldsymbol{Y}_i, M_p)\Big\}\Big] \\ &\geq \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \big[\log P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i|M_p)\big]q_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i - \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \big[\log q_i(\boldsymbol{\theta}_i)\big]q_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i.\end{aligned} \tag{2.2}$$

The last inequality holds if and only if the KL divergence between the variational distribution $q_i(\cdot)$ and the posterior distribution $P(\cdot|\boldsymbol{Y}_i, M_p)$ is 0, which indicates $q_i(\boldsymbol{\theta}_i) = P(\boldsymbol{\theta}_i|\boldsymbol{Y}_i, M_p)$. In the literature of variational inference, the right-hand side of Equation (2.2) is defined to be the evidence lower bound (Blei et al., 2017), which is equivalent, up to a constant with respect to $q(\cdot)$, to the KL divergence between the assumed variational distribution $q(\cdot)$ and the conditional density of the latent variables given the observations.

In the following, we construct an approximation for the marginal maximum likelihood estimator from the evidence lower bound. The primary objective is to identify a suitable distribution

that can approximate the posterior distribution $P(\boldsymbol{\theta}_i|\boldsymbol{Y}_i, M_p)$. Motivated by this insight, we propose to construct an EM-type algorithm to compute the marginal maximum likelihood estimator. In the E-step, we evaluate the expectation of the complete data log-likelihood, and the expectation is taken with respect to the latent variables $\boldsymbol{\theta}_i$ under its variational probability density function $q_i(\cdot)$:

$$\sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i|M_p) q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i,$$

Here the density function $q_i(\boldsymbol{\theta}_i)$ is chosen to minimize the KL divergence $KL\{q_i(\boldsymbol{\theta}_i \| P(\boldsymbol{\theta}_i|\boldsymbol{Y}_i, M_p)\}$ as the best approximation to the posterior distribution. The second term in the evidence lower bound is left out since it is irrelevant to item parameters. However, a problem with respect to minimizing the KL divergence is that it is hard to find an explicit formula for the posterior distribution of $\boldsymbol{\theta}_i$ with respect to the previous estimated item parameters $\hat{M}_p$, as it involves computing $D$-dimensional integrals. Numerical methods, such as the Gauss-Hermite approximation, Monte Carlo expectation-maximization, and stochastic expectation-maximization, are often used to provide fast approximation. Herein we adopt the Gaussian variational inference method. It is widely accepted that the posterior distribution of the latent ability $P(\boldsymbol{\theta}_i|Y_i, M_p)$ can be approximated by a Gaussian distribution (Chang & Stout, 1993; Wang 2015), and hence we aim to find an optimal $q_i(\boldsymbol{\theta}_i)$ in the family of Gaussian distribution while minimizing the KL divergence between $q_i(\boldsymbol{\theta}_i)$ and $P(\boldsymbol{\theta}_i|\boldsymbol{Y}_i, M_p)$. Since the posterior distribution can be expressed as

$$P(\boldsymbol{\theta}_i|\boldsymbol{Y}_i, M_p) = \frac{P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i|M_p)}{P(\boldsymbol{Y}_i|M_p)},$$

we only need to evaluate $P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i|M_p)$ to find a proper $q_i(\cdot)$ as

$$KL\{q_i(\boldsymbol{\theta}_i)||P(\boldsymbol{\theta}_i|\boldsymbol{Y}_i, M_p)\} = KL\{q_i(\boldsymbol{\theta}_i)||P(\boldsymbol{\theta}_i, \boldsymbol{Y}_i|M_p)\} + C.$$

Under the setting of MGPCM, the logarithm of joint distribution function of $\boldsymbol{\theta}_i$ and $\boldsymbol{Y}_i$ is

$$\begin{aligned}\log P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i|M_p) &= \log P(\boldsymbol{Y}_i|\boldsymbol{\theta}_i, M_p) + \log \phi(\boldsymbol{\theta}_i) \\ &= \sum_{j=1}^{J} \left\{ \sum_{k=0}^{K_j-1} I_{(Y_{ij}=k)} \left[ k\boldsymbol{a}_j'\boldsymbol{\theta}_i - b_{jk} - \log\left( \sum_{v=0}^{K_j-1} \exp(v\boldsymbol{a}_j'\boldsymbol{\theta}_i - b_{jv}) \right) \right] \right\} + \log \phi(\boldsymbol{\theta}_i).\end{aligned} \tag{2.3}$$

The nonlinear softmax function, defined by $f_v(\boldsymbol{x}) = \exp(x_v)/[\sum_{k=1}^{n} \exp(x_k)]$ for an $n$-dimensional vector $\boldsymbol{x}$, is the main cause of the intractability of integral. To overcome this problem, a variational lower bound based on the approximation to the softmax function is proposed and by augmenting Equation (2.3) with variational parameters, the evidence lower bound can be computed explicitly without resorting to numeric integration. Among the many approximations for the softmax function, we adopt a One-Versus-Each bound (Tisais, 2016), which well approximates the softmax function and captures the model features. We start with the following inequality:

$$f_v(\boldsymbol{x}) = \frac{e^{x_v}}{\sum_{k=1}^{n} e^{x_k}} \geq \prod_{k=1,k\neq v}^{n} \frac{e^{x_v}}{e^{x_v}+e^{x_k}} = 2\prod_{k=1}^{n} \frac{e^{x_v}}{e^{x_v}+e^{x_k}}. \tag{2.4}$$

Denote by $k_{ij}$ the realized partial credit score for the response of the $i$th person for the $j$th item. Then by applying bound (2.4) to (2.3), we have

$$\begin{aligned}
\log P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i | M_p) &= \log \phi(\boldsymbol{\theta}_i) + \sum_{j=1}^{J} \left[ \sum_{k=0}^{K_j-1} I_{(Y_{ij}=k)} \log \frac{\exp(x_{ijk})}{\sum_{v=0}^{K_j-1} \exp(x_{ijv})} \right] \\
&\geq \log \phi(\boldsymbol{\theta}_i) + \sum_{j=1}^{J} \left\{ \sum_{k=0}^{K_j-1} I_{(Y_{ij}=k)} \left[ \log 2 + \sum_{v=0}^{K_j-1} \log \frac{\exp(x_{ijk})}{\exp(x_{ijv})+\exp(x_{ijk})} \right] \right\} \\
&= \log \phi(\boldsymbol{\theta}_i) + \sum_{j=1}^{J} \left\{ \log 2 - \sum_{k=0}^{K_j-1} \log\left[1+\exp(x_{ijk}-x_{ijk_{ij}})\right] \right\}.
\end{aligned}$$

Here we denote $x_{ijk} = k\boldsymbol{a}_j'\boldsymbol{\theta}_i - b_{jk}$ for short. We wish to draw attention to our selection of the "One-Versus-Each bound." It can be established for (2.4) that a strict inequivalence holds true in all cases, except for the exceptional circumstance $x_v/x_k \to \infty$ for all $k \neq v$ with at most one exception. Additionally, the approximation is closest when $x_v$ is among the largest of all $x_k$. The idea of maximum likelihood estimation indicates that, when partial credit score $Y_{ij}$ is recorded as $k_{ij}$, $k_{ij}\boldsymbol{a}_j'\boldsymbol{\theta}_i - b_{jk_{ij}}$ is the most likely to be the largest among all $v\boldsymbol{a}_j'\boldsymbol{\theta}_i - b_{jv}$ for $v = 0, 1, \ldots, K_j - 1$. Therefore the feature of the One-Versus-Each bound does fit well as an approximation to the marginal maximum likelihood estimator.

Logistic sigmoid function (2.4) can be further approximated by a local variational approach:

$$\begin{aligned}
\log P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i | M_p) \geq \log \phi(\boldsymbol{\theta}_i) + \sum_{j=1}^{J} \Bigg\{ &\log 2 - \sum_{k=0}^{K_j-1} \eta(\xi_{ijk}) \left[ (x_{ijk}-x_{ijk_{ij}})^2 - \xi_{ijk}^2 \right] \\
&- \sum_{k=0}^{K_j-1} \frac{1}{2}(x_{ijk}-x_{ijk_{ij}}-\xi_{ijk}) - \sum_{k=0}^{K_j-1} \log(1+e^{\xi_{ijk}}) \Bigg\},
\end{aligned}$$

where $\boldsymbol{\xi} = \{\xi_{ijk}\}_{i,j,k}$ are called variational parameters, which will be iteratively updated together with item parameters in the M-step. Here the function $\eta(x)$ is defined as $(e^x - 1)/[4x(e^x + 1)]$ (Jaakkola & Jordan, 2000). We use this local variational approximation for a suitable expectation of $\log P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i | M_p)$ (i.e., given below in Equation (3.5)) that can be written as a quadratic form with respect to $\boldsymbol{\theta}_i$. This will facilitate the selection of $q_i(\cdot)$ in the family of Gaussian distributions.

Next we substitute $x_{ijk}$ by $k\boldsymbol{a}_j'\boldsymbol{\theta}_i - b_{jk}$ and write the above lower bound of joint distribution function as

$$\begin{aligned}
\log P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i | M_p) \geq \sum_{j=1}^{J} \Bigg\{ &- \sum_{k=0}^{K_j-1} \Big[ \eta(\xi_{ijk})(k-k_{ij})^2 \boldsymbol{\theta}_i'\boldsymbol{a}_j\boldsymbol{a}_j'\boldsymbol{\theta}_i - 2(k-k_{ij})\eta(\xi_{ijk})(b_{jk}-b_{jk_{ij}})\boldsymbol{a}_j'\boldsymbol{\theta}_i \\
&+ \frac{1}{2}(k-k_{ij})\boldsymbol{a}_j'\boldsymbol{\theta}_i + \eta(\xi_{ijk})(b_{jk}-b_{jk_{ij}})^2 - \frac{1}{2}(b_{jk}-b_{jk_{ij}}) \\
&- \eta(\xi_{ijk})\xi_{ijk}^2 - \frac{1}{2}\xi_{ijk} + \log(1+e^{\xi_{ijk}}) \Big] + \log 2 \Bigg\} + \log \phi(\boldsymbol{\theta}_i),
\end{aligned}$$

and therefore the expectation of the log-likelihood, which needs computing in the E-step, takes the following form:

$$\begin{aligned} E(M_p, \boldsymbol{\xi}) &:= \int_{\boldsymbol{\theta}_i} \log P(\boldsymbol{Y}_i|\boldsymbol{\theta}_i, M_p) q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i + \int_{\boldsymbol{\theta}_i} \log \phi(\boldsymbol{\theta}_i) q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\ &\geq \int_{\boldsymbol{\theta}_i} \sum_{j=1}^{J} \Bigg\{ \log 2 - \sum_{k=0}^{K_j-1} \Big[ \eta(\xi_{ijk})(k-k_{ij})^2 \boldsymbol{\theta}_i' \boldsymbol{a}_j \boldsymbol{a}_j' \boldsymbol{\theta}_i - 2(k-k_{ij}) \eta(\xi_{ijk})(b_{jk} - b_{jk_{ij}}) \boldsymbol{a}_j' \boldsymbol{\theta}_i \\ &\quad + \frac{1}{2}(k-k_{ij}) \boldsymbol{a}_j' \boldsymbol{\theta}_i + \eta(\xi_{ijk})(b_{jk} - b_{jk_{ij}})^2 - \frac{1}{2}(b_{jk} - b_{jk_{ij}}) \\ &\quad - \eta(\xi_{ijk}) \xi_{ijk}^2 - \frac{1}{2}\xi_{ijk} + \log(1 + e^{\xi_{ijk}}) \Big] \Bigg\} q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i + \int_{\boldsymbol{\theta}_i} \log \phi(\boldsymbol{\theta}_i) q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \end{aligned} \tag{2.5}$$

For a minimized KL divergence, $q_i(\theta_i)$ is selected as follows:

$$\log q_i(\boldsymbol{\theta}_i) \propto \sum_{j=1}^{J} \sum_{k=0}^{K_j-1} \Big\{ (k-k_{ij}) \big[ 2\eta(\xi_{ijk})(b_{jk} - b_{jk_{ij}}) - 0.5 \big] \boldsymbol{a}_j' \boldsymbol{\theta}_i - \eta(\xi_{ijk})(k-k_{ij})^2 \boldsymbol{\theta}_i' \boldsymbol{a}_j \boldsymbol{a}_j' \boldsymbol{\theta}_i \Big\} - \frac{\boldsymbol{\theta}_i' \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\theta}_i}{2}.$$

As the choice of $q_i(\cdot)$ has been confined in the Gaussian family, it suffices to give the update for the mean and covariance matrix:

$$\boldsymbol{\mu}_i = \boldsymbol{\Sigma}_i \times \sum_{j=1}^{J} \sum_{k=0}^{K_j-1} (k - k_{ij}) \Big[ 2\eta(\xi_{ijk})(b_{jk} - b_{jk_{ij}}) - \frac{1}{2} \Big] \boldsymbol{a}_j; \tag{2.6}$$

$$\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{\Sigma}_\theta^{-1} + 2 \sum_{j=1}^{J} \sum_{k=0}^{K_j-1} \eta(\xi_{ijk})(k - k_{ij})^2 \boldsymbol{a}_j \boldsymbol{a}_j'. \tag{2.7}$$

In each iteration, the item parameters and variational parameters $\xi_{ijk}, \boldsymbol{a}_j, b_{jk}$ are obtained from the previous M-step and taken as the initial value if it is the first iteration.

For the M-step, the item parameters are chosen to maximize the above expectation of the lower bound obtained by plugging Equation (2.6) and (2.7) into Equation (2.5):

$$\begin{aligned} E(M_p, \boldsymbol{\xi}) \geq & \sum_{i=1}^{N} \sum_{j=1}^{J} \log 2 + \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{k=0}^{K_j-1} \Bigg\{ - \eta(\xi_{ijk})(k-k_{ij})^2 \boldsymbol{a}_j' \Big[ \boldsymbol{\Sigma}_i^{(t)} + (\boldsymbol{\mu}_i^{(t)})(\boldsymbol{\mu}_i^{(t)})' \Big] \boldsymbol{a}_j \\ & + (k-k_{ij}) [2\eta(\xi_{ijk})(b_{jk} - b_{jk_{ij}}) - \frac{1}{2}] \boldsymbol{a}_j' \boldsymbol{\mu}_i^{(t)} - \eta(\xi_{ijk})(b_{jk} - b_{jk_{ij}})^2 + \frac{1}{2}(b_{jk} - b_{jk_{ij}}) \\ & + \eta(\xi_{ijk}) \xi_{ijk}^2 + \frac{1}{2}\xi_{ijk} - \log(1 + e^{\xi_{ijk}}) \Bigg\} - \frac{N}{2} \log |\boldsymbol{\Sigma}_\theta^{(t)}| \\ & - \sum_{i=1}^{N} \frac{1}{2} Tr\{ (\boldsymbol{\Sigma}_\theta^{(t)})^{-1} [\boldsymbol{\Sigma}_i^{(t)} + (\boldsymbol{\mu}_i^{(t)})(\boldsymbol{\mu}_i^{(t)})'] \\ := & \underline{E}(M_p, \boldsymbol{\xi}). \end{aligned} \tag{2.8}$$

(2.9)

Through maximizing the lower bound $\underline{E}(M_p, \boldsymbol{\xi})$, we can derive a new set of item parameters that could potentially maximize the left-hand side. Updating the variational parameters helps to prevent the iteration from leading to a smaller value of the target expectation by shrinking the inequality too much when the right-hand side is maximized. The efficiency of this majorization-maximization approach depends on the goodness of fit of the adopted softmax bound.

To maximize the lower bound on the expectation concerning the item parameters $M_p$ and variational parameters $\boldsymbol{\xi}$, we employ a Gauss-Seidel scheme to handle the nonlinear terms regarding the parameters. Each iterative update uses the most recently updated copies of the parameters. The update is given as follows.

For each $j = 1, \ldots, J$,

$$\boldsymbol{a}_j = \frac{1}{2}\left[\sum_{i=1}^{N}\sum_{k=0}^{K_j-1}\eta(\xi_{ijk})(k-k_{ij})^2(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i')\right]^{-1} \left\{\sum_{i=1}^{N}\sum_{k=0}^{K_j-1}(k-k_{ij})\left[2\eta(\xi_{ijk})(b_{jk}-b_{jk_{ij}}) - \frac{1}{2}\right]\boldsymbol{\mu}_i\right\}, \tag{2.10}$$

with $\xi_{ijk}, b_{jk}$ from the last iteration or initialization and $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ from E-step. Next for the threshold parameters, for each $j = 1, \ldots, J, k = 1, \cdots K_j - 1$,

$$b_{jk} = \frac{\sum_{i=1}^{N}[\mathrm{B}_1(i,j,k)I_{(k\neq k_{ij})} + I_{(k=k_{ij})}\sum_{v=0,v\neq k}^{K_j-1}\mathrm{B}_2(i,j,v,k)]}{2\sum_{i=1}^{N}(\eta(\xi_{ijk})I_{(k\neq k_{ij})} + I_{(k=k_{ij})}\sum_{v=0,v\neq k}^{K_j-1}\eta(\xi_{ijv}))}, \tag{2.11}$$

where

$$\mathrm{B}_1(i,j,k) = 2\eta(\xi_{ijk})(k-k_{ij})\boldsymbol{a}_j'\boldsymbol{\mu}_i + 0.5 + 2\eta(\xi_{ijk})b_{jk_{ij}};$$
$$\mathrm{B}_2(i,j,v,k) = -2\eta(\xi_{ijk})(v-k)\boldsymbol{a}_j'\boldsymbol{\mu}_i - 0.5 + 2\eta(\xi_{ijv})b_{jv}.$$

Here $\boldsymbol{a}_j$ are from the previous step and $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ are from the previous E-step. Finally for the variational parameters $\xi_{ijk}$, for each $i = 1, \ldots, I, j = 1, \ldots, J, k = 0, \cdots K_j - 1$, we have update as

$$\xi_{ijk}^2 = [(k-k_{ij})\boldsymbol{a}_j'\boldsymbol{\mu}_i - (b_{jk}-b_{jk_{ij}})]^2 + (k-k_{ij})^2\boldsymbol{a}_j'\boldsymbol{\Sigma}_i\boldsymbol{a}_j. \tag{2.12}$$

with all other parameters obtained from the latest updates.

In the exploratory analysis where we do not have any prior information on the item factor loadings, so the assumed covariance $\Sigma_\theta$ is fixed as $I_D$ and later proper rotations (Browne, 2001) are imposed to allow the factors to be correlated and thus allow for analysis of latent structures. But for confirmatory factor analysis, we update the covariance as

$$\boldsymbol{\Sigma}_\theta = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i'). \tag{2.13}$$

and scale its diagonal entries to be 1.

### 2.2. Standard Error Estimation

Computing standard errors (SEs) of the item parameter estimates is crucial for various applications, such as multidimensional computerized adaptive testing, item parameter calibration as well as differential item functioning. Challenges arise in estimating SEs when dealing with a high-dimensional latent domain and polytomous responses as in MGPCM. The commonly used method for estimating SEs is based on the approximated Fisher's information matrix. However, taking the inverse of a prohibitively large information matrix (due to high dimensions and long test length) can be unstable when the sample size of the examinees is not large enough. An alternative numerical approximation using Gaussian quadrature in EM estimation has been proposed (Cagnone & Monari, 2013), but it is computationally expensive and sensitive to dimensionality. The supplemented expectation-maximization (SEM) algorithm has also been developed in the IRT literature (e.g., Tian et al., 2013). However, in pilot simulations we found that none of these methods are capable of providing stable estimations, especially when the dimension $D$ and the number of categories $K$ are large. Therefore, to estimate the standard errors of item parameters under the pGVEM framework for MGPCM, we adopt a bootstrap approach that uses a resampling procedure. Bootstrap is an efficient alternative when the standard SEs estimation is mathematically intractable (Efron & Tibshirani, 1986). The resampling procedure avoids the direct computation of SEs.

The bootstrap procedure in the pGVEM framework is implemented as follows. First we simulated $B$ bootstrap datasets based on $\hat{M}_p = \{\hat{\boldsymbol{a}}_j, \hat{\boldsymbol{b}}_j\}_j$ estimated from the pGVEM scheme. Then we apply the pGVEM method to estimate the item parameters for each of the bootstrap datasets, denoted by $\hat{M}_p^{(1)}, \ldots, \hat{M}_p^{(B)}$. The standard errors are estimated by

$$\widehat{SE}_v = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{v}^{(i)} - \hat{v})^2},$$

where $v$ denotes item parameter $a_{jr}$ or $b_{jk}$ and $\hat{v}^{(i)}$ is its $i$th bootstrap estimate. Given that our objective is to estimate SEs rather than the distributions of the estimators, in our study, we take the number of bootstrap samples to be 50, which generates stable results numerically.

### 2.3. Determining Latent Dimension

In this section, we discuss how to select the appropriate number of latent dimensions. We propose to use the information criterion such as AIC or BIC to compare the model fit with different dimensions. In the MGPCM, direct computation of the residual sum of squares is costly. So we adopt the modified version of the information criterion where the expectation is replaced by its lower bound given by (2.9):

$$AIC^* = 2(\|\hat{\boldsymbol{A}}\|_0 + \|\hat{\boldsymbol{B}}\|_0 + \|\text{nondiag}(\hat{\boldsymbol{\Sigma}}_\theta)\|_0/2) - 2\underline{E}(\hat{M}_p, \hat{\boldsymbol{\xi}}), \tag{2.14}$$

$$BIC^* = \log(N)(\|\hat{\boldsymbol{A}}\|_0 + \|\hat{\boldsymbol{B}}\|_0 + \|\text{nondiag}(\hat{\boldsymbol{\Sigma}}_\theta)\|_0/2) - 2\underline{E}(\hat{M}_p, \hat{\boldsymbol{\xi}}), \tag{2.15}$$

where $\hat{\boldsymbol{A}} = (\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_J)$ and $\hat{\boldsymbol{B}} = (\hat{\boldsymbol{b}}_1, \ldots, \hat{\boldsymbol{b}}_J)$ are assembled matrices of the discrimination and threshold parameters, respectively, $\text{nondiag}(\hat{\boldsymbol{\Sigma}}_\theta)$ denotes the nondiagonal entries of $\hat{\boldsymbol{\Sigma}}_\theta$, and the zero norm $||\cdot||_0$ counts the number of nonzero entries of the assembled matrix. Here note that the term $\|\hat{\boldsymbol{B}}\|_0$ does not increase with dimension $D$ and it denotes the number of all effective $b$ parameters. In addition, since the covariance matrix $\hat{\boldsymbol{\Sigma}}_\theta$ is symmetric with unit diagonal entries, we count the effective number of parameters in $\hat{\boldsymbol{\Sigma}}_\theta$ as $\|\text{nondiag}(\hat{\boldsymbol{\Sigma}}_\theta)\|_0/2$. The major advantage of the proposed criteria is that the lower bound of expectation is readily obtained with the updated item parameters and variation parameters, with no extra computation cost.

## 3. Simulation Studies

### *3.1. Study I*

We conducted simulation studies to compare the empirical performance of the proposed pGVEM algorithm with the EM algorithm with fixed quadrature, Metropolis-Hastings Robbins-Monro (MHRM) algorithm (Cai, 2010), and the stochastic EM (StEM) for MGPCM, which are implemented in the R package 'mirt' (Chalmers, 2012), in terms of mean squared error and bias of the estimation together with their computational time. The simulations were conducted in the exploratory factor analysis (EFA) scenario, where no constraints on the item factor loading structure were imposed during the analysis. EFA is generally more computationally challenging than confirmatory factor analysis. In our study, we assumed that the latent covariance matrix was $I_D$ to remove scale and rotational indeterminacy, and no further assumptions on the structure of the loading matrix $A$ were made during the analysis.

After estimating the parameters by our pGVEM algorithm, we performed proper oblique rotation to allow the factors to be correlated. Many methods, including varimax, direct oblimin, quartimax, equamax, and promax (Browne, 2001; Hendrickson & White, 1964), are available for factor rotation in the literature. In our simulation study, we applied promax rotation as it is one of the most computationally efficient oblique rotation methods in large-scale factor analysis. For the estimation implemented in the 'mirt' package, we use the built-in promax rotation to obtain the estimation, and for the pGVEM estimation, we use the function promax in R, with default $m = 4$, to perform the rotation after the iteration ends.

The manipulated conditions include: (1) sample size $N = 200, 500$; (2) test length $J = 10, 20$; (3) number of categories $K = 3, 6$; (5) low and high correlation among the latent traits; (6) small- and large-scale loadings. For each condition, a total number of 100 replicated cases were simulated. In the context of partial credit models, it is noted that the scaling of loadings plays a pivotal role in shaping the likelihood function. Specifically, when loadings are high and there exist multiple categories for partial credit scoring, the following case usually occurs: The probability of attaining the highest or lowest scores becomes disproportionately large. Consequently, this dominance of extreme scores may result in insufficient records of intermediate scores, thereby making the estimation of threshold parameters problematic. Therefore in the simulation studies, we considered two cases: (1) low scale loading: Parameter $a_{jr}$ was simulated from $Unif(0.5, 1)$ for all $j = 1, \ldots, J,\ r = 1, \ldots, D$; (2) high scale loading: Parameter $a_{jr}$ was simulated from $Unif(1, 2)$ for all $j = 1, \ldots, J,\ r = 1, \ldots, D$. The threshold parameters $b_{jk}$ are simulated from $N(0, 1)$ for all $j = 1, \ldots, J$ and $k = 1, \ldots, K-1$. For the latent variables, they were simulated from a multivariate normal distribution with 0 mean and covariance matrix $\mathbf{\Sigma}_\theta$. The diagonal entries were fixed as 1, and off-diagonal entries were generated from a uniform distribution $Unif(0.1, 0.3)$ in the low correlation case and $Unif(0.5, 0.7)$ in the high correlation case. For the responses generated from the simulated model parameters, we perform simulation only on cases where for all $j = 1, \ldots, J$ and $k = 0, \cdots K-1$,

$$\#\{i \mid Y_{ij} = k, i = 1, \ldots, N\} > 0.$$

Here # denotes the set counting operator. Skipping any item with all responses being 0 is necessary since the threshold parameter linked to each category of this item cannot be identified within the finite sample context. For the convergence criterion, the algorithm was terminated when the change of all item parameters between two iterations dropped below a pre-specified threshold, i.e.,

$$\frac{1}{J \times D + J \times K} \sum_{j=1}^{J} \left[ \sum_{r=1}^{D} \left( a_{jr}^{(t)} - a_{jr}^{(t-1)} \right)^2 + \sum_{k=1}^{K-1} \left( b_{jk}^{(t)} - b_{jk}^{(t-1)} \right)^2 \right] < 10^{-5}. \tag{3.1}$$

The estimation errors are presented separately in the form of mean squared error and bias for the discrimination and threshold parameters and the covariance matrix, averaged across the test items:

$$Bias_a = \frac{1}{JD}\sum_{j=1}^{J}\sum_{r=1}^{D}\hat{a}_{jr} - a_{jr}, \quad MSE_a = \frac{1}{JD}\sum_{j=1}^{J}\|\boldsymbol{a}_j - \hat{\boldsymbol{a}}_j\|_2^2;$$

$$Bias_b = \frac{1}{J(K-1)}\sum_{j=1}^{J}\sum_{k=1}^{K-1}\hat{b}_{jr} - b_{jr}, \quad MSE_b = \frac{1}{J(K-1)}\sum_{j=1}^{J}\sum_{k=1}^{K-1}(b_{jk} - \hat{b}_{jk})^2;$$

$$Bias_\Sigma = \frac{2}{D(D-1)}\sum_{l<h}\hat{\Sigma}_{hl} - \Sigma_{hl}, \quad MSE_\Sigma = \frac{2}{D(D-1)}\|\boldsymbol{\Sigma}_\theta - \hat{\boldsymbol{\Sigma}}_\theta\|_F^2.$$

Note that here both the true and estimated discrimination parameters have been rotated by the promax rotations. The number of Markov chain samples drawn in the MHRM algorithm was by default 5, 000 in the R package "mirt." The convergence criterion and optimizer were all set as default in the 'mirt' package.

Under the setting of small-scale loadings (i.e., $a_{jr} \sim$ Unif(0.5, 1)), Figs. 1 and 2 show the MSE of the four estimation methods for 3-category and 6-category cases, respectively. Each box represents the distribution of errors from 100 replications. We truncated the scale of the $y$-axis of the plot to make it easier to compare the estimation precision across different scenarios. We present the full version of the boxplots in Appendix B. Overall, our method provides more stable and accurate estimates of the MGPCM, as seen from both figures. The observed results indicate a reduced variability in estimation errors for the model parameters when comparing pGVEM to the other methods. Both MHRM and StEM exhibit better stability and accuracy compared to the standard EM algorithm. Notably, pGVEM demonstrates good stability, particularly evident when the sample size $N$ is 200. Additionally, it is noteworthy that as the number of categories increases, the estimation of the threshold parameters becomes more challenging, an anticipated result given that the model becomes more complicated for multicategory cases. Interestingly, despite a decrease in accuracy, the proposed pGVEM method exhibits a comparatively modest increase in variability in many cases compared with alternative methods, indicating the capability of the pGVEM method to handle more complex scenarios. Furthermore, we present the bias of the estimation using the four different methods in Figs. 3 and 4 for the considered cases. In general, the bias observed in pGVEM estimation tends to be more moderate across various cases, particularly with regard to the threshold parameter.

Under the setting of large-scale loadings (i.e., $a_{jr} \sim$ Unif(1,2)), the MSE and bias results are presented in Figs. 5, 6, 7, and 8. We can see from the simulation results that in this setting the error of estimation gets larger in most cases. The reason is that high level of loading increases the frequency of extreme scores, thus making the estimation more challenging. Yet pGVEM still outperforms the other methods with the regime of small sample size or low correlation. Furthermore, in terms of the recovery of the threshold parameter, pGVEM provides estimates with less bias and error. It is noteworthy that in this scenario, it occurs more frequently that for some item, there is no record of certain category from any individual, leading to us discarding such cases from the analysis. This suggests that when dealing with multiple categories of responses, the discrimination parameters may tend to be small for model interpretability.

In Fig. 9 we present the averaged computation time of the four methods under different settings of sample size $N$ and test length $J$. The results exhibit similarity across the cases, and for brevity, we displayed the case where the discrimination parameters are simulated from $Unif(0.5, 1)$ and correlation coefficient from $Unif(0.1, 0.3)$. We take a total number of 6 categories. It is

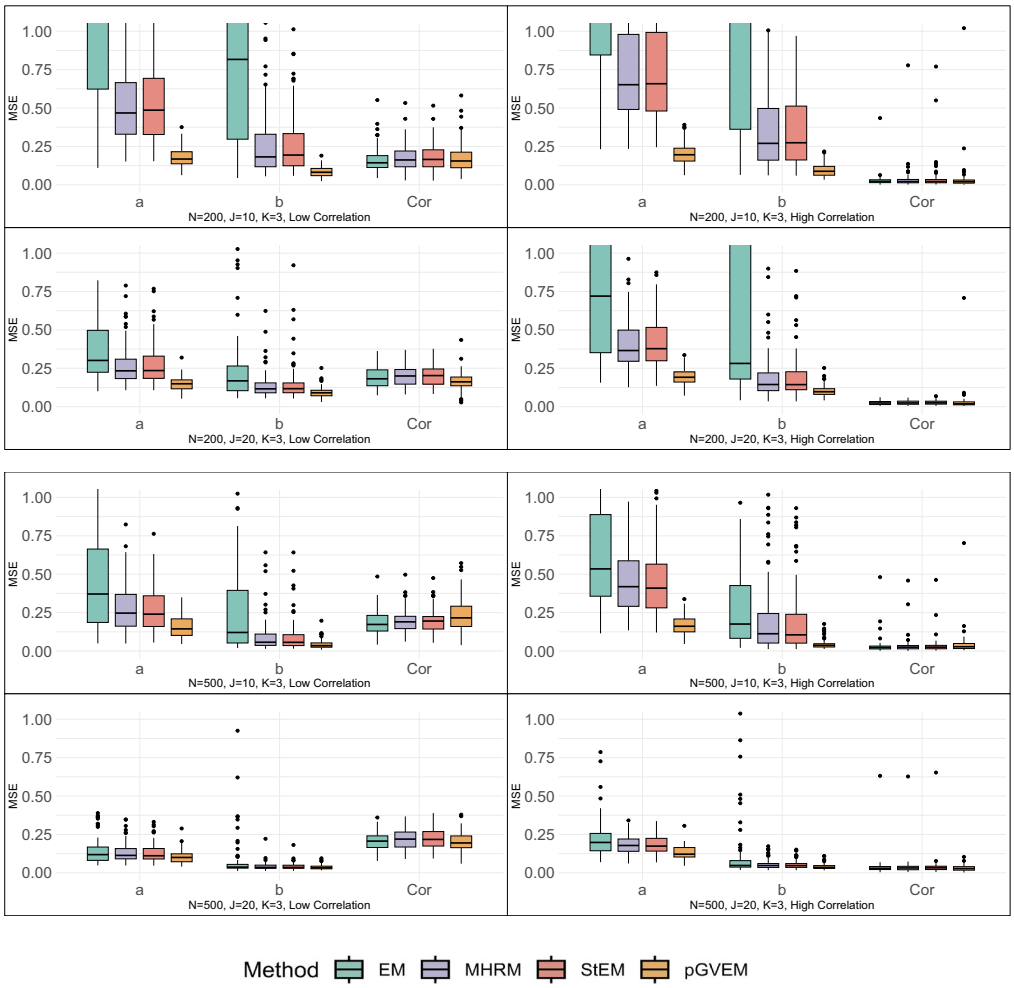



FIGURE 1.
Mean squared error of the estimation for the multidimensional generalized partial credit model of 3 categories from exploratory factor analysis with small-scale loadings using different methods.

obvious that, compared with pGVEM, the traditional EM algorithm is inefficient especially when the sample size is large. Also, pGVEM is slightly faster than the stochastic EM algorithm and achieves similar computation efficiency compared with MH-RM algorithm in the displayed case. The computational efficiency of our pGVEM algorithm makes it possible to provide fast estimation on large datasets.

The results of our study demonstrate the superiority of the pGVEM algorithm over the traditional EM algorithm along with MH-RM and StEM in terms of parameter recovery and computational efficiency. Specifically, the pGVEM algorithm achieves comparable, and in some cases, superior parameter recovery compared to the other algorithms. Moreover, pGVEM generates fewer estimation outliers, particularly in situations where the sample size and test length are large.

As a side check, we compared the pGVEM algorithm with the GVEM algorithm proposed by Cho et al. (2021) for the special case of M2PL. The comparison was made under identical

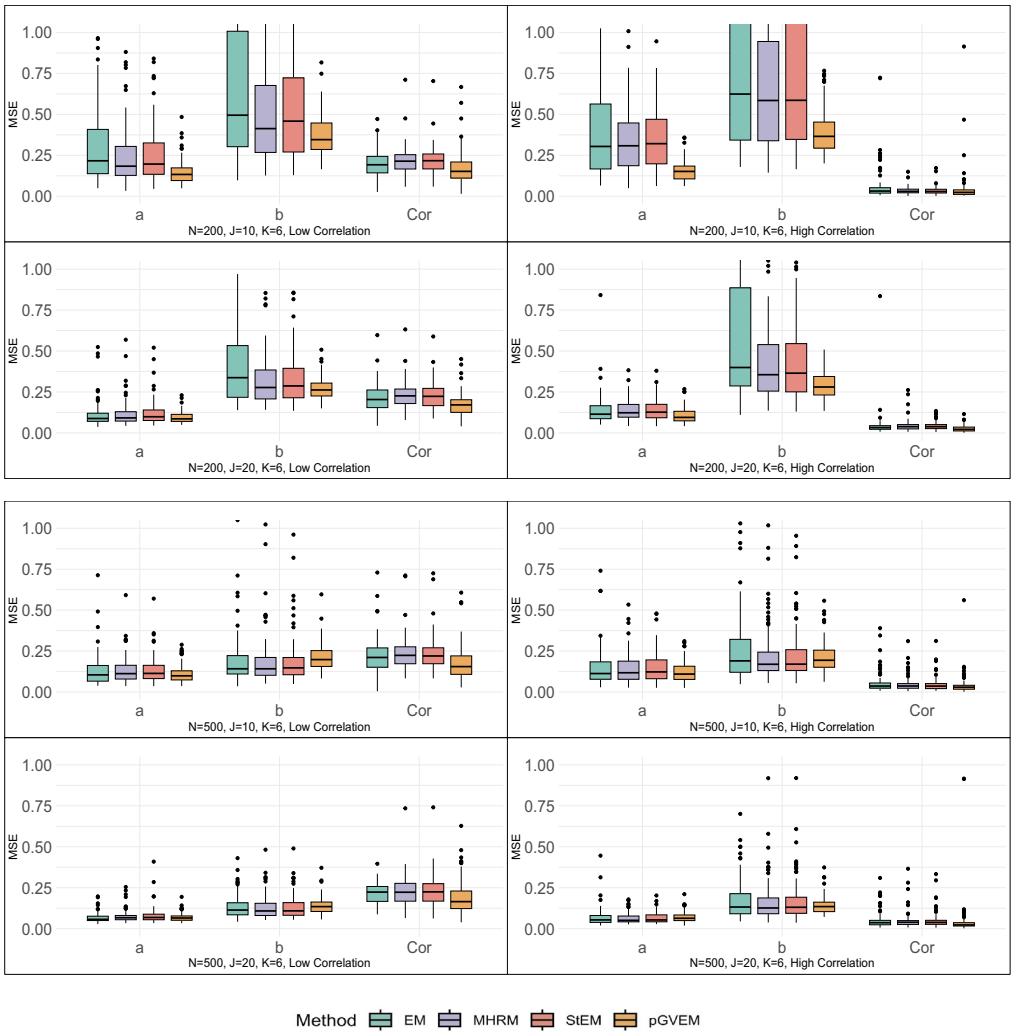



FIGURE 2.
Mean squared error of the estimation for multidimensional generalized partial credit model of 6 categories from exploratory factor analysis with small-scale loadings using different methods.

setting in this section, except for that we take $K=2$. The results, presented in Appendix B, indicate that our algorithm performs similarly to Cho et al. (2021) with binary data. Overall, our findings demonstrate that the pGVEM algorithm is a robust and efficient method for parameter recovery for the MGPCM. It can also be applied to other models, including M2PL model, with similar success. The results suggest that the pGVEM algorithm may be a valuable tool for researchers seeking to analyze complex data structures efficiently and accurately.

### *3.2. Study II*

We conducted simulation study II to assess the performance of the proposed bootstrap standard error (SE) estimation procedure. We explore the bootstrap estimation in the multidimensional case with multiple categories. In the multidimensional case, as clarified in Sect. 2.2, we found that the traditional methods (EM and MH-RM) exhibited instability and produced infeasible results

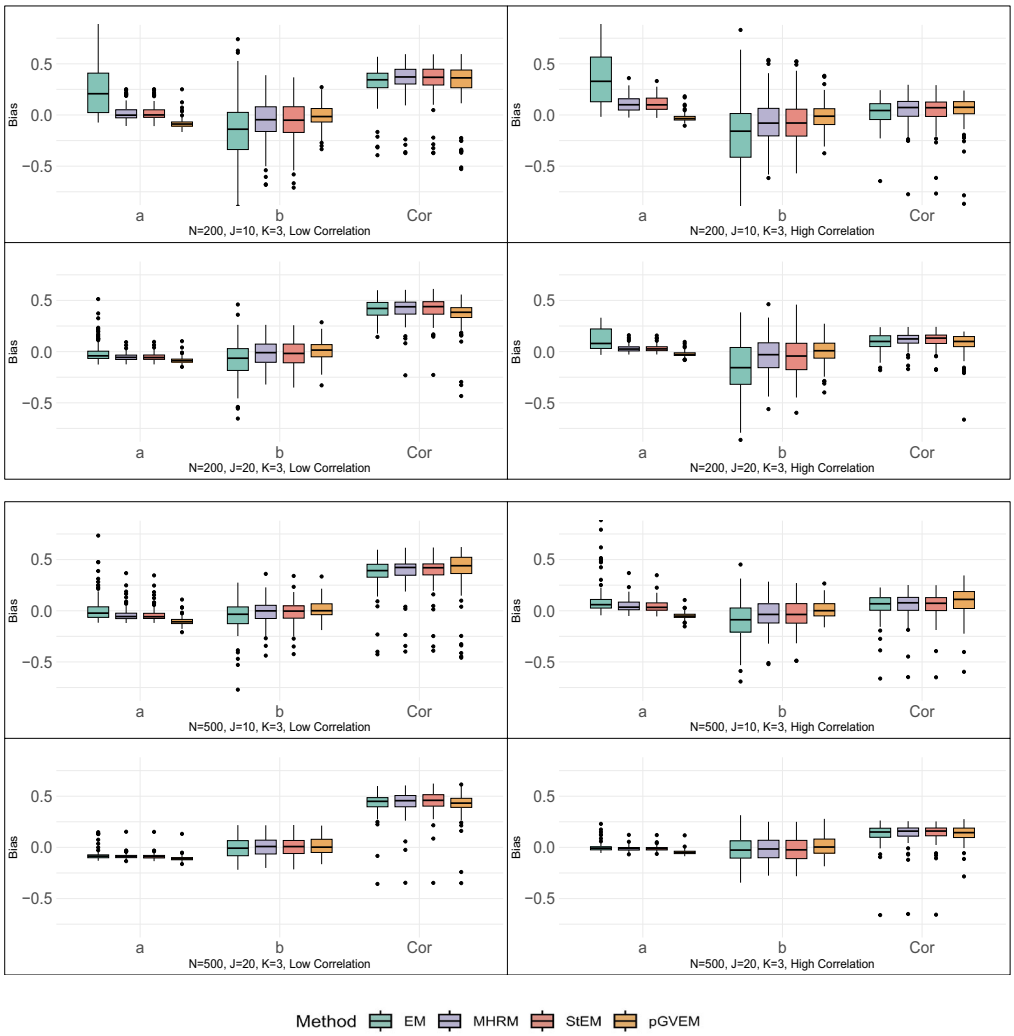



FIGURE 3.
Bias of the estimation for the multidimensional generalized partial credit model of 3 categories from exploratory factor analysis with small-scale loadings using different methods.

across numerous settings. Consequently, in this section, we focus on the bootstrap-based SE estimates under the EM algorithm, MH-RM, and the proposed pGVEM algorithm.

The comparisons were conducted under the simulation setting similar to Simulation Study I and the manipulated factors include sample size, test length, factor correlations, and the number of categories. The empirical standard deviations of the estimated item parameters as

$$SE_v = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{v}^{(r)} - v)^2$$

across replications per condition are considered as the approximations of true SEs for each method. Here $v$ stands for item parameters to represent $a_{jd}$ or $b_{jk}$, and $\hat{v}^{(r)}$ is the estimated parameter in

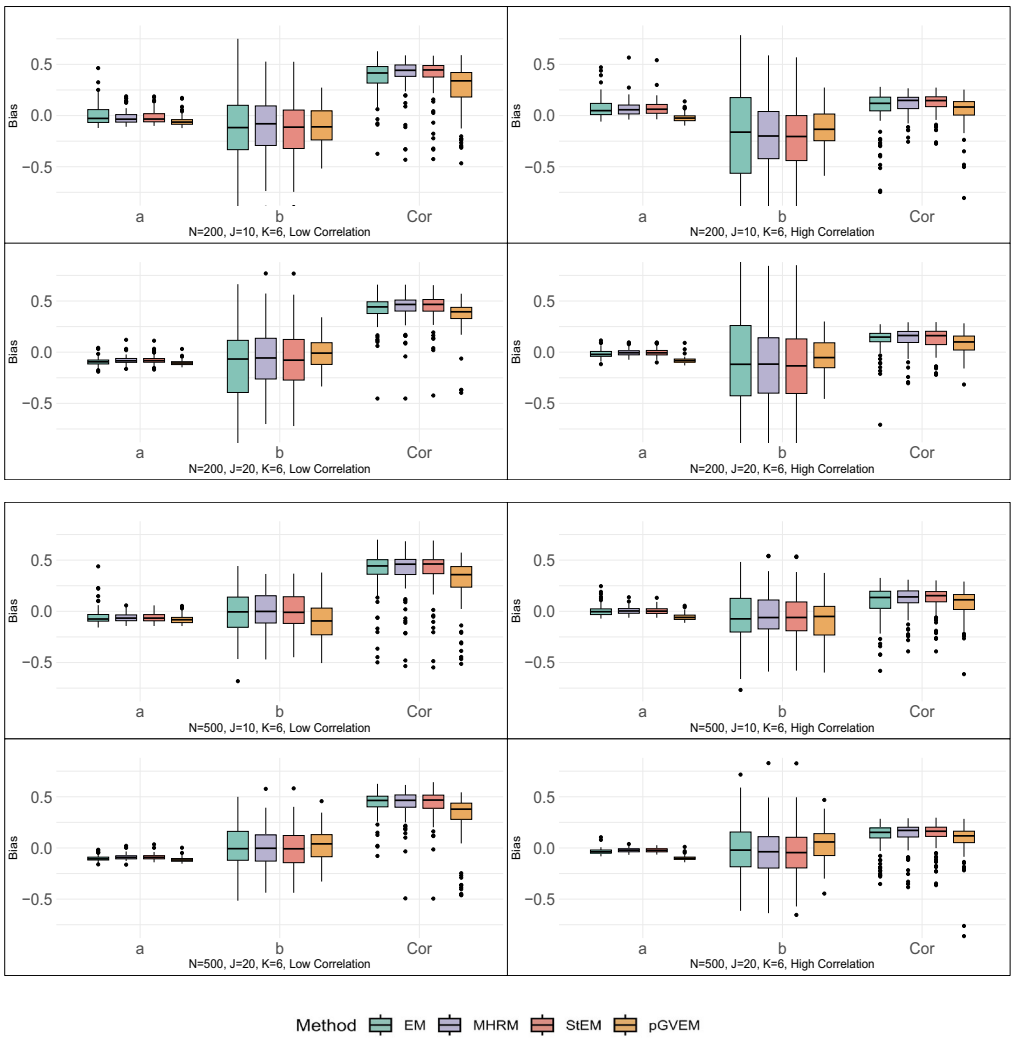



FIGURE 4.
Bias of the estimation for the multidimensional generalized partial credit model of 6 categories from exploratory factor analysis with small-scale loadings using different methods.

the $r$th replicate. To assess the performance of the proposed method, we computed SE estimations along with their bias and relative bias as follows:

$$Average\ SE = \frac{1}{J(D+K)} \sum_{j=1}^{J} \left[ \sum_{r=1}^{D} \widehat{SE}_{a_{jr}} + \sum_{k=1}^{K-1} \widehat{SE}_{b_{jk}} \right]$$

$$Bias = \frac{1}{J(D+K)} \sum_{j=1}^{J} \left[ \sum_{r=1}^{D} \widehat{SE}_{a_{jr}} - SE_{a_{jr}} + \sum_{k=1}^{K-1} \widehat{SE}_{b_{jk}} - SE_{b_{jk}} \right]$$

$$Relative\ Bias = \frac{1}{J(D+K)} \sum_{j=1}^{J} \left[ \sum_{r=1}^{D} (\widehat{SE}_{a_{jr}} - SE_{a_{jr}})/SE_{a_{jr}} + \sum_{k=1}^{K-1} (\widehat{SE}_{b_{jk}} - SE_{b_{jk}})/SE_{b_{jk}} \right],$$

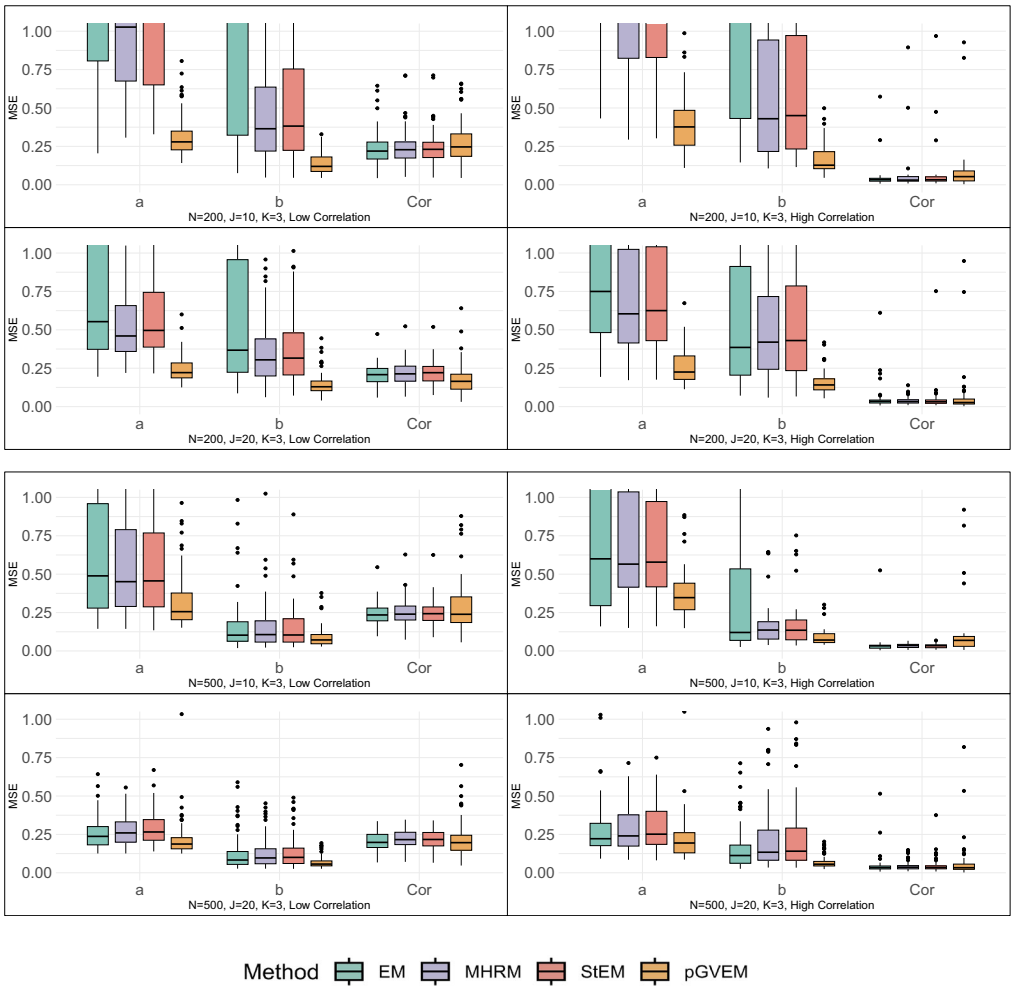



FIGURE 5.
Mean squared error of the estimation for the multidimensional generalized partial credit model of 3 categories from exploratory factor analysis with large-scale loadings using different methods.

providing a comprehensive assessment of the reliability and accuracy of the proposed bootstrap methods. Here we present the SEs of $\boldsymbol{a}$ and $\boldsymbol{b}$ pooled together with the aim of showing the effectiveness of our method compared with its alternatives. The results for $\boldsymbol{a}$ and $\boldsymbol{b}$ are similar to the pooled ones. The variability of discrimination and threshold parameters has been well illustrated in the previous study.

The results are shown in Fig. 10 for the low correlation setting and 11 for the high correlation setting. In the multidimensional case, where most of the methods fail to estimate the SEs numerically or from Fisher's information matrix, the bootstrap method still generates stable results. In comparison with bootstrap methods from alternative estimations, the pGVEM-based bootstrap exhibits a lower bias. It is observed that when the sample size is 200, the pGVEM-based bootstrap may slightly underestimate standard errors. Nevertheless, its overall performance remains relatively strong across diverse settings.

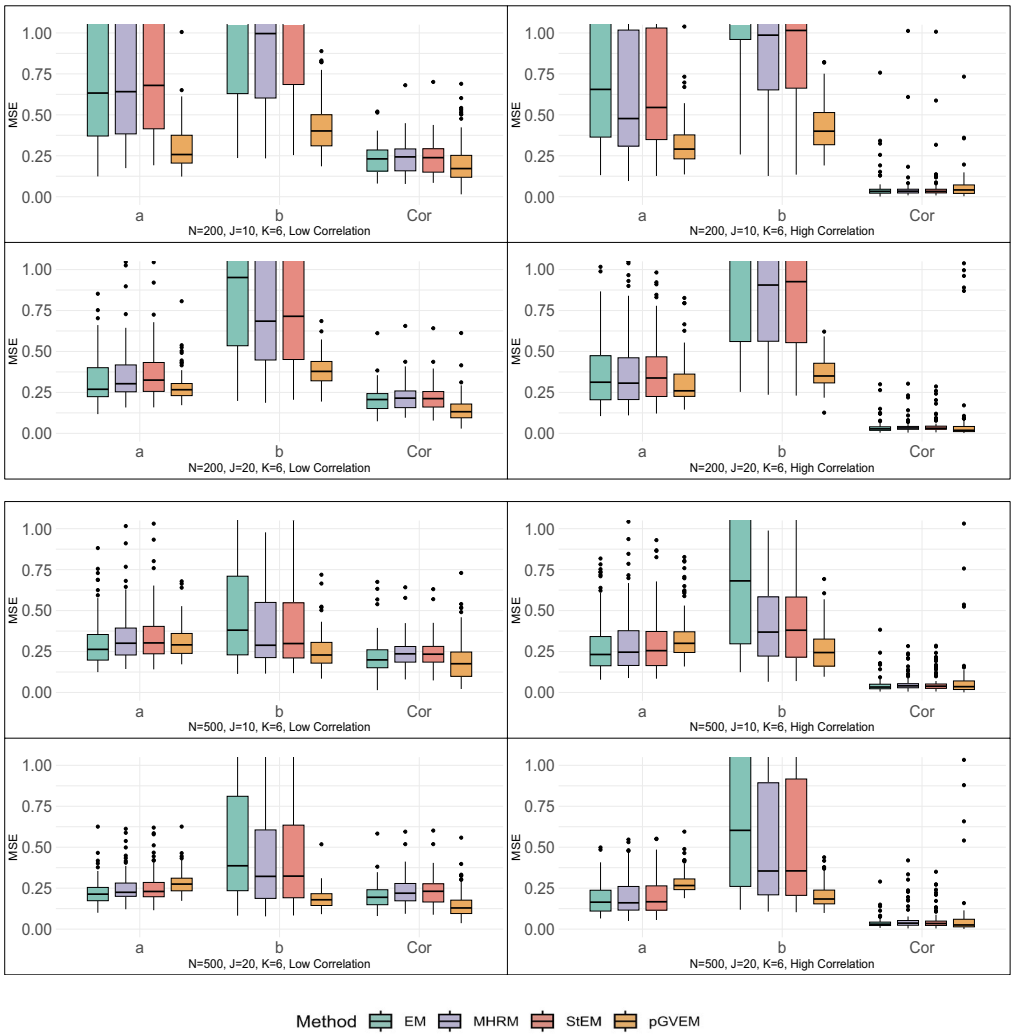



FIGURE 6.
Mean squared error of the estimation for multidimensional generalized partial credit model of 6 categories from exploratory factor analysis with large-scale loadings using different methods.

### *3.3. Study III*

In this section we conduct a simulation study to examine the performance of the proposed AIC and BIC in determining the latent dimension. Considering the complexity of the model setting of MGPCM, we explore the accuracy of factor identification in sample size $N = 500, 800$ and test length $J = 20, 40$. We consider both low correlation settings (simulated from $Unif(0.1, 0.3)$) and high correlation settings (simulated from $Unif(0.5, 0.7)$) for dimensions $D = 2, 3, 4$. In each configuration, a total number of 100 independent samples were generated and we recorded the number of cases where the number of factors was correctly identified. Discrimination parameters are generated from $Unif(1, 2)$.

Tables 1 and 2 present the correct estimation rates for the number of dimensions of 3 and 6 categories, respectively. The results indicate that an increase in sample size generally contributes to higher correct estimation rates. We can also observe that a lower correlation generally leads to

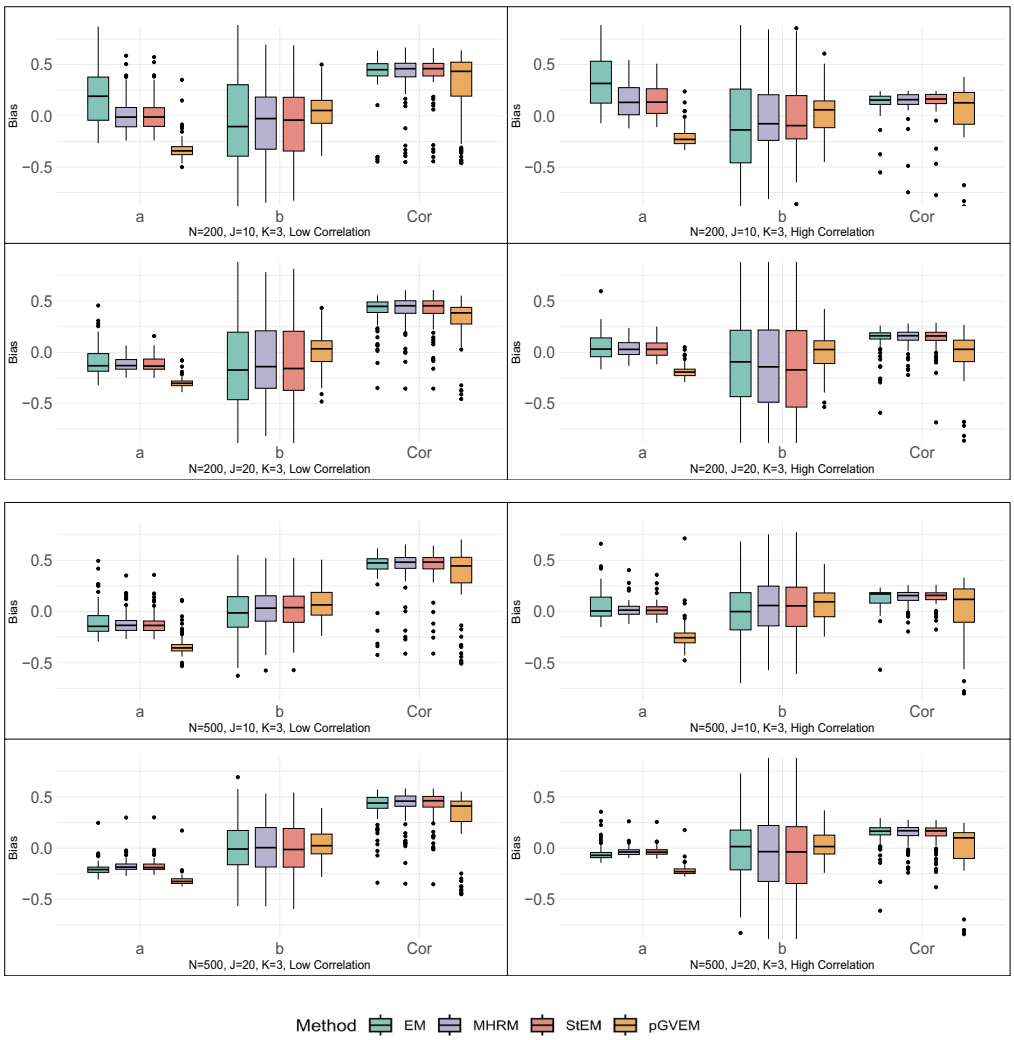



FIGURE 7.
Bias of the estimation for the multidimensional generalized partial credit model of 3 categories from exploratory factor analysis with large-scale loadings using different methods.

higher correct estimation rates, especially when there are fewer categories and shorter test lengths, as the latent structure may be more challenging to identify in such settings. Overall, we observe that AIC performs slightly better in the case of $K = 3$ and BIC has an advantage in the regime of $K = 6$. In conclusion, the proposed criterion is efficient in general, and with a larger sample size, the model is more likely to be correctly identified. Our findings are also consistent with existing studies under the MIRT model (Cho et al., 2021).

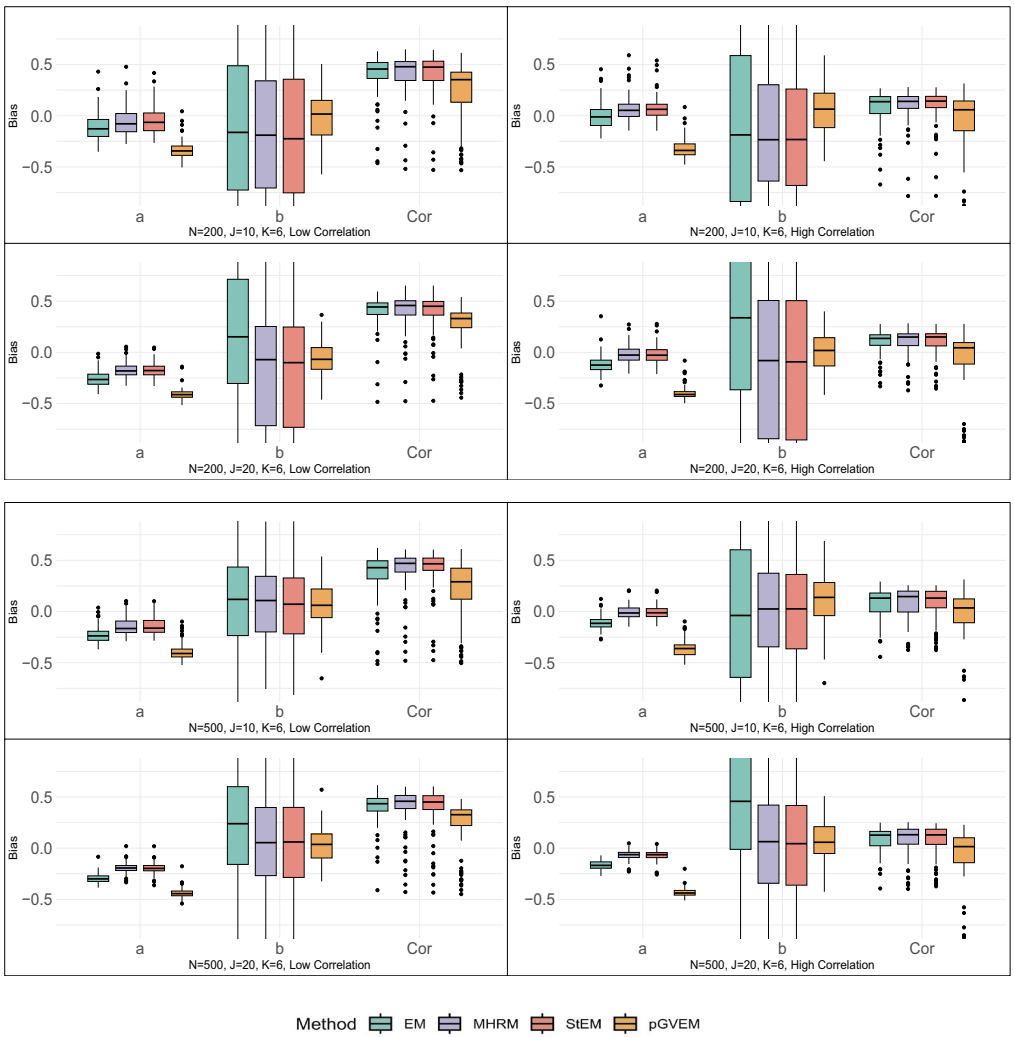



FIGURE 8.
Bias of the estimation for the multidimensional generalized partial credit model of 6 categories from exploratory factor analysis with large-scale loadings using different methods.

## 4. Real Data Analysis

### *4.1. Trend in International Mathematics and Science Study Dataset*

In this section, we demonstrate the application of the pGVEM algorithm by analyzing a dataset from the Trend in International Mathematics and Science Study (TIMSS) (Mullis & Martin, 2017; Martin & Mullis, 2019; Fishbein et al., 2018; Martin et al., 2020). TIMSS provides reliable and timely trend data on the mathematics and science achievement of US students compared to that of students in other countries. The assessment consists of a large pool of mathematics and science questions, which are divided into different blocks using the item matrix sampling design to relieve response burden. Specifically, 14 booklets were assembled, and each student was required to complete one of them. Different booklets may include the same items for linking. TIMSS 2019

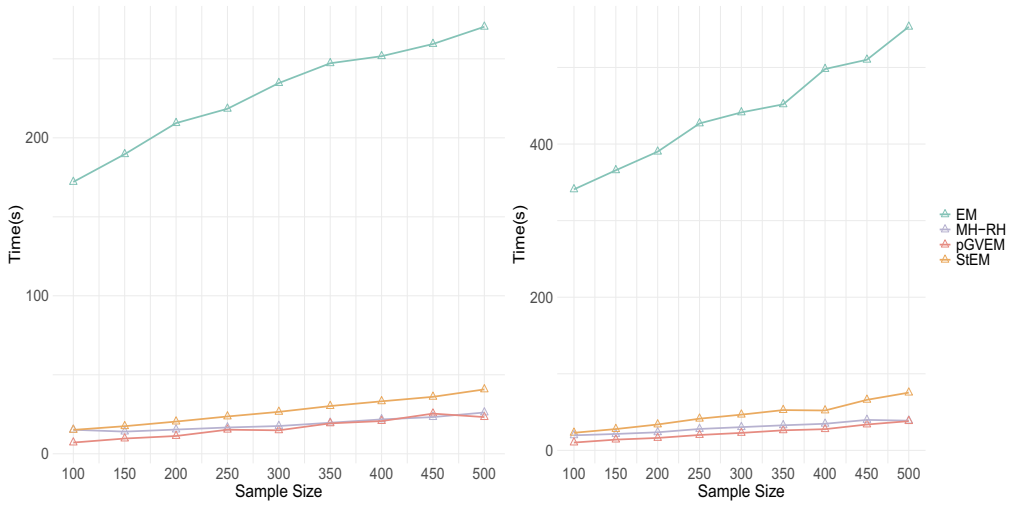


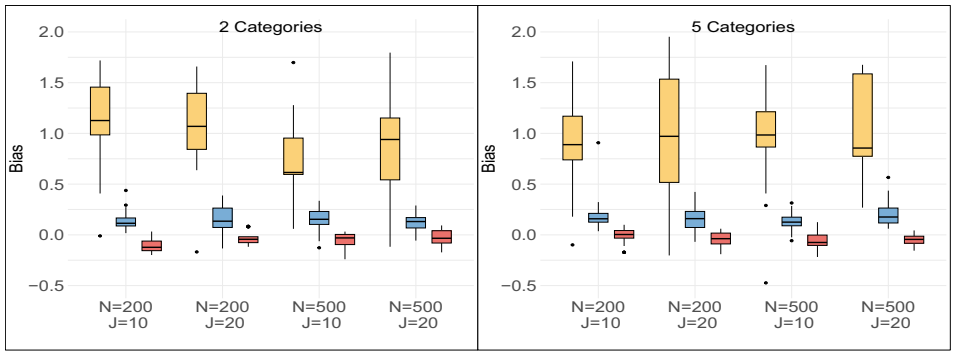

FIGURE 9.
Computational time comparison for four methods with $K = 6$: left side depicts $J = 10$, while the right side corresponds to $J = 20$.

divided the test items into 28 blocks for each grade, with 13 mathematics items and 15 science items. In the data of students in grade eight, each block contained 12 to 18 items. For our analysis, we selected a mathematics and a science block that appeared in booklet 5 of grade 8. In the mathematics block there were 2 polytomous items, and in the science there were 3. Of the 1,252 students who responded to these booklets, 918 students' responses were completely recorded. In this study, we only estimated the parameters using the data from these 918 students. Appendix C provides details of the item code and corresponding test content. The IRT parameters provided in the TIMSS assessment document (Martin & Mullis, 2019) were used as the true parameters, to which our estimated parameters were compared.

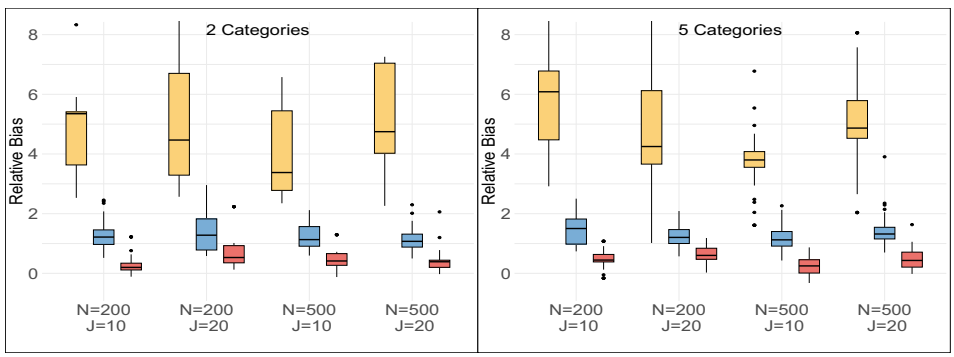
It should be noted that in operational analysis of TIMSS, uni-dimensional IRT models were used for math and science domains separately. When analyzing the items from math and science domain separately, the modified information criterion in Equation (2.14) and (2.15) computed under the EFA framework both attain the smallest value when the latent dimension is 1 (i.e., $D = 1$), which implies that both domains are essentially uni-dimensional under the generalized partial credit model setting. However, according to the information criterion provided in the 'mirt' package (i.e., using its default EM algorithm), the latent dimension cannot be clearly decided. In the following we display the parameters estimated by EM with fixed quadrature and pGVEM. The results are as follows.

To show the results visually, we plot the estimated parameters, by EM and pGVEM, in Figs. 12 and 13. Based on our analysis, it can be inferred that the $b$-parameters obtained from both methods exhibit a significant level of proximity. Similarly, the $a$-parameters demonstrate close values. However, it is worth noting that the estimates derived from the EM algorithm tend to be slightly larger for the majority of the items.

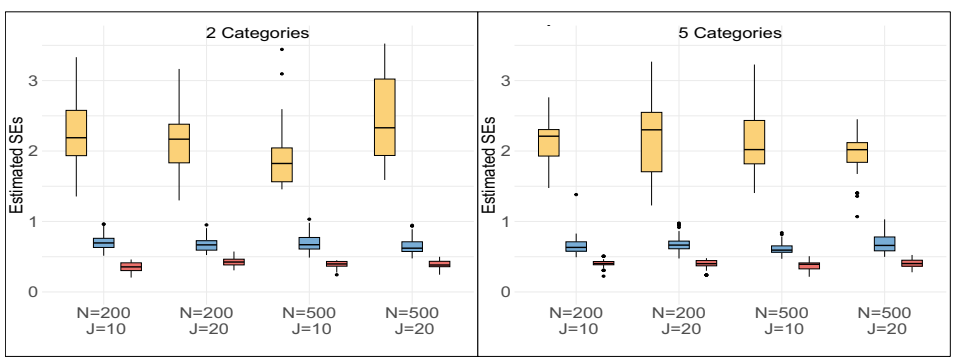
To reveal the relation of latent ability across different domains, we estimate all 28 items jointly. First, compute the modified information criterion in Equation (2.14) and (2.15) for the setting of joint estimating the data from mathematics and science block, with results presented in Table 3 and 4. From the results the optimal number of latent dimensions is 2. This result is reasonable as math proficiency and science proficiency should emerge as two separate factors.




(a) Bias of Standard Error

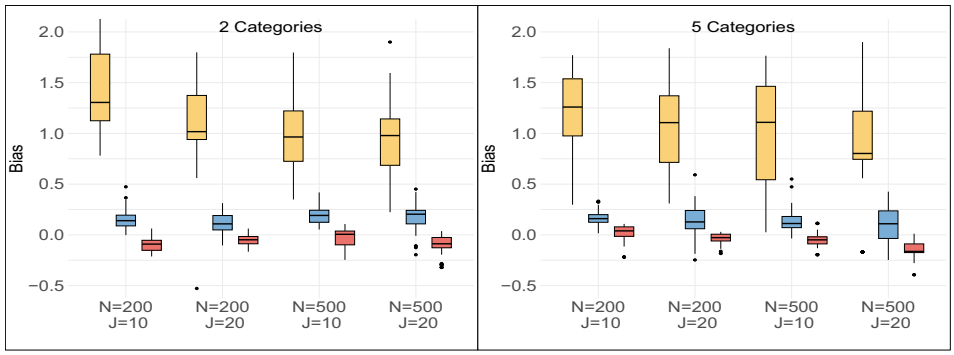



(b) Relative Bias of Standard Error

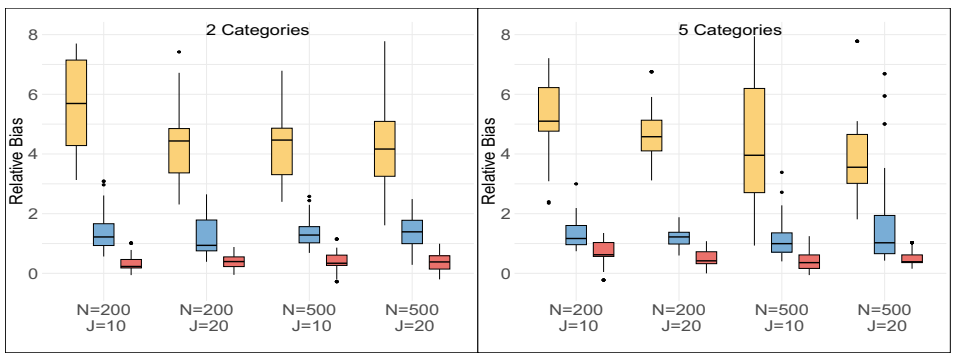



(c) Estimated Standard Error

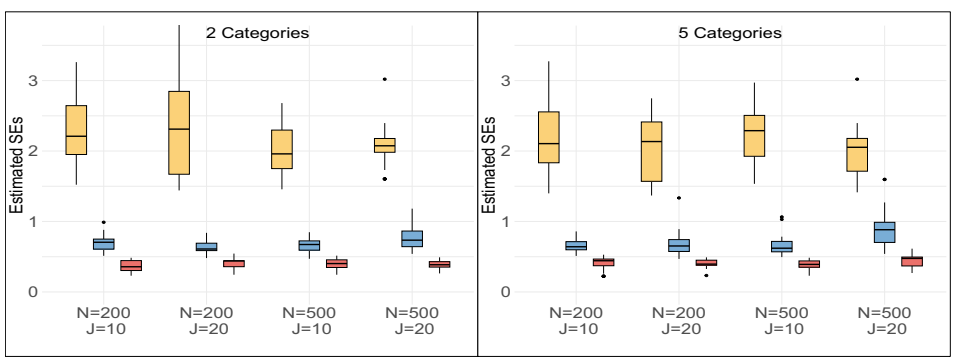

FIGURE 10.
Standard error assessment of estimation for the multidimensional generalized partial credit model from exploratory factor analysis using different methods with $D = 3$ in low factor correlation setting.




(a) Bias of Standard Error




(b) Relative Bias of Standard Error




(c) Estimated Standard Error


FIGURE 11.
SE assessment of estimation for the multidimensional generalized partial credit model from exploratory factor analysis using different methods with $D = 3$ in high factor correlation setting.

TABLE 1.
Correct number of trials in determining the latent dimension, $K = 3$

| Correctness | Low correlation | | | | | | High correlation | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | D=2 | | D=3 | | D=4 | | D=2 | | D=3 | | D=4 | |
| | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC |
| $N = 500, J = 20$ | 99 | 97 | 53 | 42 | 7 | 4 | 66 | 52 | 3 | 2 | 0 | 0 |
| $N = 500, J = 40$ | 100 | 100 | 96 | 92 | 68 | 56 | 100 | 100 | 79 | 39 | 25 | 12 |
| $N = 800, J = 20$ | 99 | 98 | 60 | 54 | 26 | 13 | 64 | 51 | 23 | 12 | 17 | 3 |
| $N = 800, J = 40$ | 100 | 100 | 99 | 97 | 80 | 70 | 100 | 100 | 83 | 67 | 44 | 32 |

TABLE 2.
Correct number of trials in determining the latent dimension, $K = 6$

| Correctness | Low correlation | | | | | | High correlation | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | D=2 | | D=3 | | D=4 | | D=2 | | D=3 | | D=4 | |
| | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC |
| $N = 500, J = 20$ | 9 | 29 | 72 | 87 | 89 | 87 | 37 | 75 | 85 | 85 | 39 | 32 |
| $N = 500, J = 40$ | 6 | 38 | 25 | 75 | 61 | 76 | 10 | 18 | 40 | 96 | 86 | 90 |
| $N = 800, J = 20$ | 28 | 48 | 90 | 94 | 91 | 85 | 77 | 91 | 92 | 88 | 41 | 33 |
| $N = 800, J = 40$ | 34 | 62 | 46 | 71 | 58 | 87 | 31 | 43 | 79 | 98 | 93 | 92 |

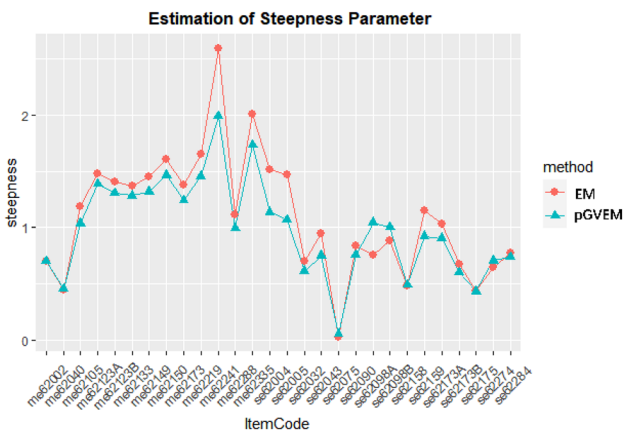



FIGURE 12.
Comparison of discrimination parameters estimated from 'mirt' and pGVEM.

Yet the information criterion output from 'mirt' package again did not provide sensible results. By assuming the latent dimension is 2, we present the results estimated by the two methods after promax rotation in Fig. 14. The analysis of estimated parameters reveals a clear two-factor loading structure as shown in Fig. 14. Notably, the second dimension emerges as primarily the students' proficiency in solving mathematical problems, whereas the first dimension tends to measure students' science proficiency. This finding suggests that the two dimensions capture distinct but interconnected aspects of overall cognitive ability. Moreover, we can infer that certain types of questions are particularly effective in assessing students' latent abilities in either mathematics or science. These questions demonstrate a higher degree of typicality in evaluating students'

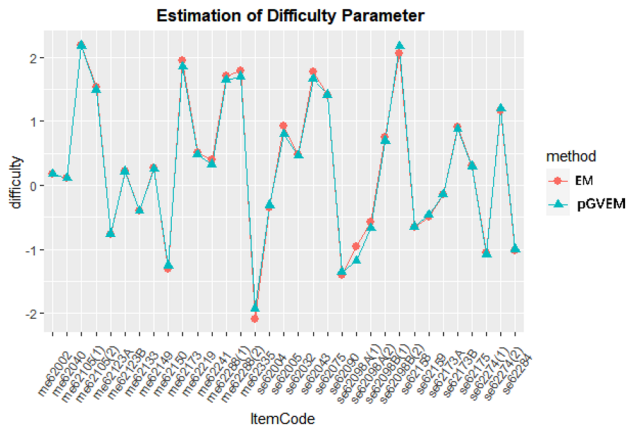



FIGURE 13.
Comparison of threshold parameters estimated from 'mirt' and pGVEM.

TABLE 3.
Information criterion from 'mirt'. The smallest value of each information criterion is given in bold

| Dimension | Math (AIC) | Math (BIC) | Sci (AIC) | Sci (BIC) | Joint (AIC) | Joint (BIC) |
|---|---|---|---|---|---|---|
| 1 | **12923.50** | 19087.96 | **13048.86** | **19256.74** | **31296.65** | **31735.79** |
| 2 | 12939.39 | **19022.30** | 13117.81 | 19263.41 | 31311.77 | 31765.61 |
| 3 | 12940.75 | 19027.24 | 13167.39 | 19335.86 | 31341.26 | 31861.50 |
| 4 | 12963.25 | 19039.05 | 13233.29 | 19410.39 | 31441.63 | 31966.93 |

Math, Sci, Joint denote, respectively, information criterion for math items, science items and the collection of all items jointly.

TABLE 4.
Information criterion from pGVEM algorithm. The smallest value of each information criterion is given in bold

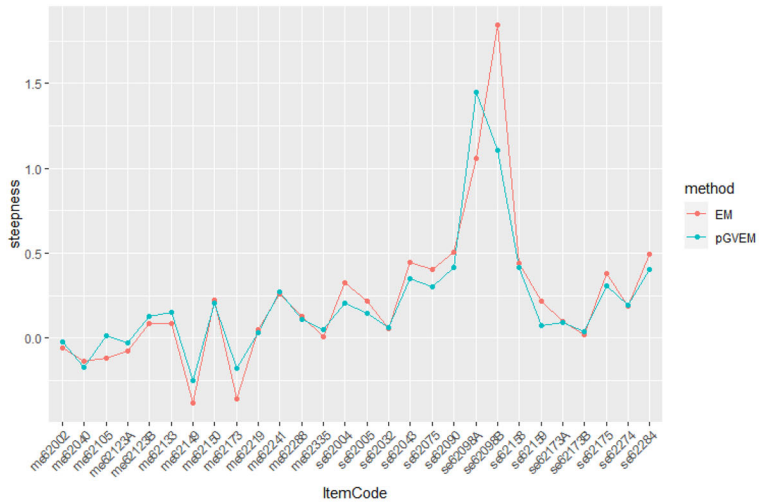
| Dimension | Math (AIC) | Math (BIC) | Sci (AIC) | Sci (BIC) | Joint (AIC) | Joint (BIC) |
|---|---|---|---|---|---|---|
| 1 | **13795.40** | **18533.71** | **13935.24** | **18697.66** | 31873.35 | 32172.33 |
| 2 | 14744.12 | 18809.71 | 14961.12 | 19060.46 | **31806.70** | **32162.33** |
| 3 | 15656.59 | 19630.67 | 15960.38 | 19977.87 | 32392.96 | 33000.55 |
| 4 | 16588.21 | 20575.46 | 16988.45 | 21028.75 | 33187.73 | 33964.10 |

Math, Sci, Joint denote, respectively, information criterion for math items, science items and the collection of all items jointly.

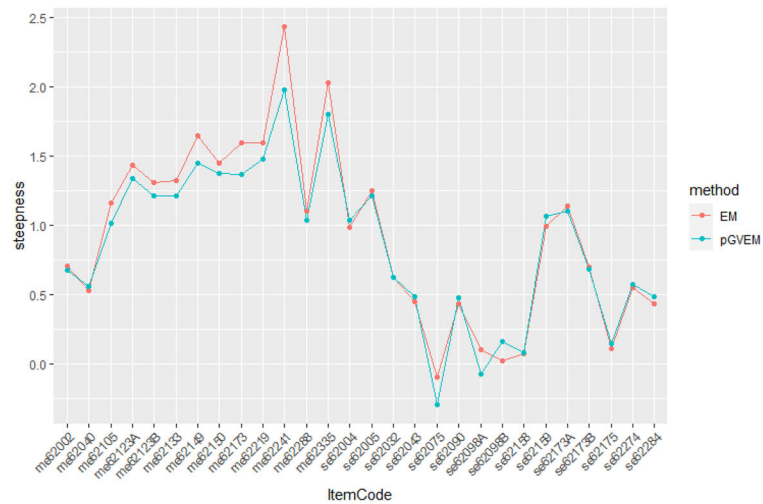
competence within their respective domains. This insight underscores the importance of carefully selecting assessment items that align with the targeted cognitive skills, as they offer a more accurate reflection of students' underlying abilities.

### *4.2. Big-Five Factor Personality Dataset*

We further demonstrate the application of the pGVEM algorithm by analyzing a dataset from the Big-Five personality assessment. The five-factor model (FFM) stands as the most widely recognized and extensively used model in psychology for understanding and measuring personality (Goldberg, 1992). This model provides a framework to capture the richness and complexity of individual differences, assuming five domains that encompass a comprehensive spectrum of personality traits. The five dimensions include Openness, Conscientiousness, Extraversion,




(a) The First Component of Estimated Parameters




(b) The Second Component of Estimated Parameters

FIGURE 14.
Jointly estimated discrimination parameters from 'mirt' and pGVEM.

Agreeableness, and Neuroticism (often referred to by the acronym OCEAN) (Costa & McCrae, 2008).

To measure the latent traits, various assessment tools have been developed to assess an individual's standing on the Big-Five dimensions (Wiggins & Trapnell, 1997; McCrae et al., 1996; Goldberg et al., 1999). In our study, we employ the Big-Five Factor Markers derived from the International Personality Item Pool (IPIP), a widely recognized instrument developed by Goldberg (1992). The dataset is publicly available at http://openpsychometrics.org/tests/IPIP-BFFM/. This dataset consists of responses from a substantial sample of 19,718 individuals, each evaluated on the fifty-item Big-Five Factor Markers. The IPIP consists of fifty items, each requiring respondents to rate the extent to which they perceive each statement as true about themselves on

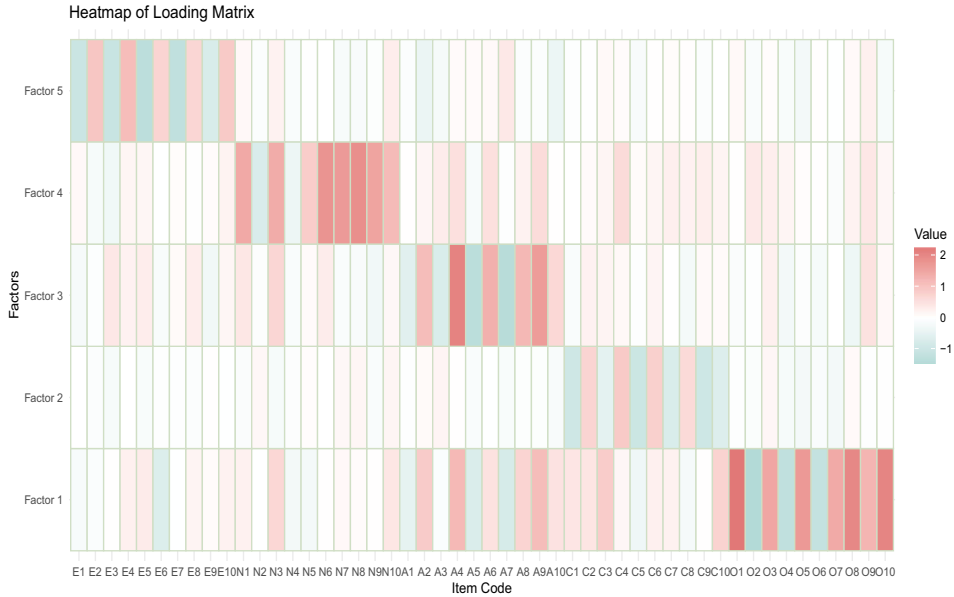



FIGURE 15.
Heatmap of loading matrix from Big-Five personality study.

TABLE 5.
Estimated correlation matrix of the 5 latent factors for the Big-Five dataset. The absolute values of correlations higher than 0.1 are given in bold

| | F1 | F2 | F3 | F4 | F5 |
|---|---|---|---|---|---|
| F1 | 1.000 | **0.237** | **0.529** | **0.459** | 0.030 |
| F2 | **0.237** | 1.000 | **0.294** | −0.098 | 0.031 |
| F3 | **0.529** | **0.294** | 1.000 | **0.305** | 0.096 |
| F4 | **0.459** | −0.098 | **0.305** | 1.000 | **−0.213** |
| F5 | 0.030 | 0.031 | 0.096 | **−0.213** | 1.000 |

a five-point scale. This scale ranges from 1 (Disagree) to 3 (Neutral) to 5 (Agree), providing a nuanced and graded assessment across the five categories.

Under the setting of exploratory factor analysis, we fit the MGPCM using the proposed pGVEM algorithm. The estimated factor loadings are represented in Fig. 15. The heatmap reveals that each factor exhibits a distinct and salient association with the grouped items. Also, we can see that not all latent factors have a positive influence on the overall rating—a phenomenon commonly observed in personality tests (McCrae et al., 1996; Goldberg et al., 1999). A significant correlation between Factor 1 and Factor 3 can be observed, as they both significantly impact responses to the A-items. To further illustrate these relationships, we present the estimated correlation matrix of the factors in Table 5. The largest correlation is between Factor 1 and Factor 3 with a value of 0.529, which is consistent with our observation from Fig. 15. Overall, the estimated factor structure aligns with existing literature, demonstrating the usefulness of the proposed method as a computationally efficient tool to analyze large scale assessment data.

## 5. Conclusion

In this paper, we proposed a new Gausssian variational estimation algorithm of polytomous responses (namely, pGVEM) for the multidimensional generalized partial credit model (MGPCM). The MGPCM is one of the most widely used models in cognitive assessment and educational measurement for items that are scored polytomously, such as assigning partial scores for intermediate correct steps. There are limited methods for an efficient estimation of MGPCM, and it is shown that our pGVEM algorithm outperforms many existing approaches in that it is relatively stable and much faster without resulting in much aberrant estimates or convergence issues. When the sample size and test length are both large, our proposed variational lower bound seems to approximate the target marginal likelihood closely. The computation efficiency is achieved by replacing the intractable high-dimensional integral with a variational lower bound that contributes to faster EM-type updates involving only small-scale linear equations. The simulation study in Sect. 3 provides simulation evidence to support pGVEM in producing accurate parameter estimates quickly, as compared to the traditional EM implementation of marginal maximum likelihood estimators. The real data analysis demonstrates that our estimation scheme is capable of extracting proper information about the latent variables.

Variational inference has emerged as a prominent and efficient methodology in the field of psychometrics, particularly due to its ability to handle large-scale datasets with both accuracy and computational efficiency. In addition to IRT models with continuous latent variables, Yamaguchi and Okada (2020) recently proposed a variational Bayesian (VB) inference algorithm for the saturated cognitive diagnosis models, which represents a notable advancement in scalable and computationally efficient Bayesian estimation for discrete latent variable models. Oka and Okada (2023) developed a scalable estimation algorithm for the DINA Q-matrix, which employs an iteration scheme utilizing stochastic optimization and variational inference. Our method, on the other hand, extends the literature on continuous latent variable estimation (Cho et al., 2021, 2022) by considering multiple response categories. It is encouraging to further explore the relationship of various variational approximation methods, which may lead to a more robust and flexible estimation framework for many psychometric models.

There are also some other potential directions to extend the current work. First, as with many nonconvex optimization problems, the efficacy of our algorithm may be influenced by the chosen initial values. Instances where initial values deviate substantially from the actual parameters may lead the algorithm to converge to local optima rather than the global one. Given the propensity for variational approximations to yield multiple local optimal values, investigating the initialization strategy for our estimation process warrants further exploration. Second, the current method of determining the number of latent dimensions may not work for certain cases, especially when the dimension is very high or there exists a high correlation between latent factors. Therefore, a more accurate and robust model selection method that will work in these challenging scenarios is needed.

## Acknowledgments

This work is partially supported by IES grant R305D200015 and NSF grants SES-1846747 and SES-2150601.

### Declarations

**Conflict of interest** The authors declare that there is no conflict of interest.

**Data Availability** The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

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.

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis, 1*(1), 121–143.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association, 112*(518), 859–877.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261–280.

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

Cagnone, S., & Monari, P. (2013). Latent variable models for ordinal data by using the adaptive quadrature approximation. *Computational Statistics, 28*, 597–619.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika, 75*(1), 33–57.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1–29.

Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58*(1), 37–52.

Chen, Y., Li, X., & Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika, 84*(1), 124–146.

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.

Cho, A. E., Xiao, J., Wang, C., & Xu, G. (2022). Regularized variational estimation for exploratory item factor analysis. *Psychometrika*.

Costa, P. T., & McCrae, R. R. (2008). The revised NEO personality inventory (NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment, 2*(2), 179–198.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, pp. 54–75.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Feuerstahler, L. M., & Waller, N. G. (2014). Estimation of the 4-parameter model with marginal maximum likelihood. *Multivariate Behavioral Research, 49*(3), 285.

Fishbein, B., Martin, M. O., Mullis, I. V., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-Scale Assessments in Education, 6*(1), 1–23.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26–42.

Goldberg, L. R., et al. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe, 7*(1), 7–28.

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology, 17*(1), 65–70.

Jaakkola, T. S., & Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing, 10*(1), 25–37.

Kim, J., & Wilson, M. (2020). Polytomous item explanatory item response theory models. *Educational and Psychological Measurement, 80*(4), 726–755.

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.

Martin, M. O., & Mullis, I. V. (2019). TIMSS 2015: Illustrating advancements in large-scale international assessments. *Journal of Educational and Behavioral Statistics, 44*(6), 752–781.

Martin, M. O., von Davier, M., & Mullis, I. V. (2020). Methods and procedures: TIMSS 2019 technical report. *International Association for the Evaluation of Educational Achievement*.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the revised NEO personality inventory: Confirmatory factor analysis versus procrustes rotation. *Journal of Personality and Social Psychology, 70*(3), 552–566.

Meng, X., Xu, G., Zhang, J., & Tao, J. (2020). Marginalized maximum a posteriori estimation for the four-parameter logistic model under a mixture modelling framework. *British Journal of Mathematical and Statistical Psychology, 73*, 51–82.

Mullis, I. V. & Martin, M. O. (2017). *TIMSS 2019 assessment frameworks*. ERIC.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Oka, M., & Okada, K. (2023). Scalable Bayesian approach for the DINA Q-matrix estimation combining stochastic optimization and variational inference. *Psychometrika, 88*(1), 302–331.

Opper, M., & Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural Computation, 21*(3), 786–792.

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician, 64*(2), 140–153.

Ormerod, J. T., & Wand, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics, 21*(1), 2–17.

Reckase, M. D. (2009). *Multidimensional item response theory models, in Multidimensional item response theory*. Springer.

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*(1), 647–662.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika, 70*(3), 533–555.

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.

Tisais, M. (2016). One-vs-each approximation to softmax for scalable estimation of probabilities. *Neural Information Processing Systems, 29*, 4161–4169.

Titterington, D. (2004). Bayesian methods for neural networks and related models. *Statistical Science, 19*(1), 128–139.

von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics, 35*(2), 174–193.

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika, 80*(2), 428–449.

Wiggins, J. S., & Trapnell, P. D. (1997). Personality structure: The return of the Big Five. In *Handbook of personality psychology*, pp. 737–765. Academic Press.

Wolfinger, R., & O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation, 48*(3–4), 233–243.

Yamaguchi, K., & Okada, K. (2020). Variational Bayes inference algorithm for the saturated diagnostic classification model. *Psychometrika, 85*(4), 973–995.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30*(6), 469–492.

Zhang, S., Chen, Y., & Liu, Y. (2020). An improved stochastic EM algorithm for large-scale full-information item factor analysis. *British Journal of Mathematical and Statistical Psychology, 73*(1), 44–71.

*Manuscript Received: 3 AUG 2023*
*Accepted: 22 JAN 2024*
*Published Online Date: 1 MAR 2024*