Statistics and Its Interface Volume 12 (2019) 35–48

Bayesian estimation of a multilevel multidimensional item response model using auxiliary variables method: an exploration of the correlation between multiple latent variables and covariates in hierarchical data

Jiwei Zhang*, Jing Lu*, and Jian Tao†

Within the framework of Bayesian analysis, we present a multilevel multidimensional item response modeling and estimation method to study the relations among multiple abilities and covariates in a hierarchical data structure. The proposed method is well suited to examining a scenario in which a test measures multidimensional latent traits (e.g., reading ability, cognitive ability, and computing ability) and in which students are nested within classes or schools. The developed Gibbs sampling algorithm based on auxiliary variables can accurately estimate the correlations among multidimensional latent traits, along with the correlation between person- and school-level covariates and latent traits. Three information criteria and the pseudo-Bayes factor approach are used to evaluate model fit and make model comparison. Simulation studies show that the proposed method works well in estimating all model parameters across a broad spectrum of scenarios. A case study on an educational assessment data is investigated to demonstrate the practical application of the proposed procedure.

Keywords and phrases: Bayesian inference, Gibbs sampling algorithm, Information criteria, Cross-validation log-likelihood, Pseudo-Bayes factor.

1. INTRODUCTION

Item response theory (IRT) is widely used in the analysis of educational and psychological tests for measuring the examinees' latent traits based on their responses to test items. Modern tests often consist of several subtests each aiming on one or more latent traits. Analyses of such tests require multidimensional IRT models [10, 11, 32, 40] to achieve more precise measurement by accurately quantifying the examinee/item interaction and utilizing the dependency among subtests. In addition, large-scale modern tests often involve hierarchical structure in the design. For example, the Program for International Student Assessment (PISA) has an important goal in better understanding cross-cultural differences both in science achievement and attitudes towards science, and hence involves students from different schools in multiple countries. To elucidate the effects at the school and country level in this hierarchical design, a standard thought is to use multilevel models [16, 33], which inspires the need of multilevel multidimensional IRT (MMIRT) models [7, 18, 19, 20, 26, 28, 37].

The MMIRT is often needed in school effectiveness research in education. The objective of a school effectiveness study is to investigate the relationship between outcome factors (student's latent traits) and student characteristics (social background), teacher characteristics and school (such as teachers' attitude, class size, and school climate). The typical nested structure on student/class/teacher requires a multilevel model. Meanwhile, most administered tests in such studies contain various subtest components measuring different latent traits, which requires a multidimensional IRT. For example, an English test usually consists of several subscales and each subscale is used to measure a subdimensional latent trait (such as vocabulary cognitive ability and the ability to diagnose grammar structure). Sometimes, one can consider a simplification by arguing a general ability (e.g. a linear combination of multiple latent traits [37]) is adequate in describing the item responses and then adopt a unidimensional IRT. However, such an approach is lack of generality as individual latent trait often possesses different between- and within-cluster (e.g. school) variations.

Compared with the existing MMIRT works, the problem to be solved and the viewpoint of modeling in this study are very different. [37] developed multidimensional IRT models with a hierarchical structural relationship between specific ability and general ability. That is, each specific ability is a linear function of the general ability or the general ability is a linear combination of all the specific abilities. For more similar modeling methods, see [19, 20]. In addition, [28] proposed an MMIRT model to analyze students' aggressive-disruptive behavior in elementary school classrooms. In the

[*] Jiwei Zhang and Jing Lu are co-first authors. They contributed equally to this work.

[†] Corresponding author.

multilevel modeling process, the ability (factor) of each dimension has between- and within-cluster variations. However, the source of the between- and within-cluster variations — whether the between-cluster (school) covariates and within-cluster individual background variables have effects on two part variations — was not considered further. For a similar modeling method, see [18, 26]. Furthermore, [7] presented an extended mixed-effect IRT model to analyze PISA data, where the individual background covariates (level-1 predictors) and level-2 school predictors are used to model the fixed effects rather than to directly explain multiple dimensional abilities. [12, 21, 25] proposed multilevel IRT models to represent the interactions between unidimensional ability and individual- and school-level covariates where the ability parameters have a hierarchical nested structure. Our study can be conceived of as a multidimensional extension of the model developed by [12, 21, 25], replacing their unidimensional IRT model with a multidimensional two-parameter logistic model. The advantage of this modeling method is that it can better reflect hierarchical structural data and provide simultaneous estimation of item parameters and person measures as well as accurate inference about higher-level measures where covariates are introduced to explain the relationship between predictors and multiple latent traits.

Due to its model complexity, estimation of an MMIRT model is often computationally costly. Finding the marginal maximum likelihood estimation (MMLE) requires numerical evaluation of the marginal likelihood, and it soon becomes infeasible as the dimensionality increases. On the other hand, Bayesian approaches either use the Metropolis-Hastings algorithm [8, 17, 27], which is prone to slow mixing or non-convergence for high-dimensional data, or require rigid prior distributions (e.g. the conjugate distribution in the normal ogive framework) to perform Gibbs sampling [2, 4, 12]. In this article, we propose a novel and effective Gibbs sampling algorithm for estimating the MMIRT model based on some cleverly designed auxiliary variables. We demonstrate the approach via the multilevel multidimensional two-parameter logistic model (MM2PLM) and expect it to work in general. Due to its Bayesian nature, our approach avoids numerical evaluation of the marginal likelihood. Meanwhile, its Gibbs sampling structure often leads to fast convergence, and more importantly, it allows the use of flexible prior distributions which can adequately quantify various prior information. Besides parameter estimation, we further develop information criteria and Bayes factors for our method, and provide tools for model assessment.

We demonstrate our approach through a case study on large-scale English achievement tests with a three-level nested structure. The merit of our approach is manifested by answering the following questions important to field researchers. (1) Conditionally on the individual-level gender (GD), school-level teacher satisfaction (ST) and school climate (CT), how will students with high socioeconomic-status (SES) scores perform compared to low SES scores, in terms of English performances as measured by four types of latent traits? (2) Are the performances between males and females identical for the different latent traits when controlling for SES, ST and CT? (3) Do the teachers' or schools' effects (covariates) affect the individuals' performances? If so, what are the effects? (4) Can a measurement tool (items of subtest) be used to test whether items' factor patterns reflect the subscales of the test battery? That is, can the four subtests of the test battery be traced in the discrimination parameters on the four dimensions? (5) According to the model selection results, which model is the best to fit the data and how can judge the individual-level regression coefficients be judged as fixed effect or random effect?

The rest of the article is organized as follows. Section 2 presents the detailed development of the proposed general MMIRT and procedure for hierarchical data. Section 3 provides a new computational strategy based on auxiliary variables to meet computational challenges for the proposed model. Bayesian model comparison criteria are discussed in Section 4. In Section 5, simulation studies are conducted to examine the performance of parameter recovery using the Gibbs sampling algorithm based on auxiliary variables and to assess model fit using the information criteria and pseudo-Bayes factor. In addition, a real data analysis of the education quality assessment is given in Section 6. We conclude this article with a brief discussion and suggestions for further research in Section 7.

2. MODEL AND MODEL IDENTIFICATION

2.1 Multilevel multidimensional IRT model

The model contains three levels. At the first level, a multidimensional two-parameter logistic model is used to model the relationship between items, persons, and responses. At the second level, person parameters are predicted by person-level covariates, such as an individual's SES. At the third level, persons are nested within schools, and school-level covariates (such as school climate) are included.

• Measurement model at level 1 (Multidimensional two-parameter logistic model):

(1)

$$p_{ijk} = p\left(Y_{ijk} = 1 \,|\boldsymbol{\theta}_{ij}, \boldsymbol{a}_k, b_k\right) = \frac{\exp\left[\sum_{q=1}^{Q} a_{kq}\theta_{ijq} - b_k\right]}{1 + \exp\left[\sum_{q=1}^{Q} a_{kq}\theta_{ijq} - b_k\right]},$$

where $j = 1, \cdots, J$ represent J schools, and within school j, there are $i = 1, \cdots, n_j$ persons. $k = 1, \cdots, K$ indicate the items. Let y_{ijk} denote the response of the ith examinee in the jth group answering the kth item. Then, the correct response probability can be expressed as p_{ijk}, and $\boldsymbol{\theta}_{ij}$ denotes a Q-dimensional vector of ability parameters for the ith person in the jth group, i.e., $\boldsymbol{\theta}_{ij} = (\theta_{ij1}, \cdots, \theta_{ijQ})'$. Let

$\boldsymbol{\xi}_k = (a_{k1}, \cdots, a_{kQ}, b_k)'$ denote the vector of item parameters for the kth item, where $\boldsymbol{a}_k = (a_{k1}, \cdots, a_{kQ})'$ is a vector of discrimination or slope parameters and b_k is the difficulty or intercept parameter.

- Multilevel structural model at level 2 (individual level):

$$\theta_{ijq} = \beta_{0jq} + x_{1ij}\beta_{1jq} + \cdots + x_{hij}\beta_{hjq} + e_{ijq}, \tag{2}$$

where $\boldsymbol{x}$'s represent student-level covariates, such as an individual's GD and SES. h denotes the number of student covariates at level 2. The residual term $\boldsymbol{e}_{ij} = (e_{ij1}, \cdots, e_{ijQ})'$ is assumed to follow a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma}_e)$, where $\boldsymbol{\Sigma}_e$ is a Q-by-Q covariance matrix. The student's latent traits are considered to be the latent outcome variables of the multilevel regression model. Differences in latent traits among individuals within the same school are modeled given student-level characteristics. Therefore, the explanatory information $\boldsymbol{x}$ at the individual level explains variability in the latent traits within school.

- Level 3 (school level):

$$\beta_{hjq} = \gamma_{h0q} + w_{1j}\gamma_{h1q} + \cdots + w_{sj}\gamma_{hsq} + u_{hjq},$$

where s denotes the number of school covariates at level 3. Each level-2 random regression coefficient parameter is β_{hjq}, which can be interpreted by school-level covariates. The residual vector $(u_{0jq}, \cdots, u_{hjq})'$ is assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{T}_q$, where $\boldsymbol{T}_q$ is an $(h+1)$-by-$(h+1)$ covariance matrix, $q = 1, \cdots, Q$. The variation across schools is modeled given background information at the school level. To control the model complexity, we assume that the level-3 residual covariance between different dimensions is 0; that is,

$$\text{Cov}\,(u_{hjq_1}, u_{hjq_2}) = 0,\ q_1 \neq q_2,\ j = 1, \cdots, J,\ h = 0, 1, 2, \cdots \tag{3}$$

2.2 Model identification

A common approach to ensure identification of the single-level two-parameter IRT model [6, 24, 39] is to set the mean and variance of the ability distribution to zero and one, respectively. Alternatively, one can impose constraints of $\prod_k a_k = 1$ and $\sum_k b_k = 1$ for model item parameters; the equivalent form is to anchor one discrimination parameter to one, and one difficulty parameter to zero. In our hierarchical model, we impose some constraints on discrimination and difficulty parameters to identify the MM2PLMs. For the discrimination and difficulty parameters, we set Q item parameters b_k equal to zero if $k = q$, imposing the restrictions $a_{kq} = 1$, in which $k = 1, \cdots, Q$, and $q = 1, \cdots, Q$. If $k \neq q$, $a_{kq} = 0$. If $k > q$, b_k and a_{kq} are free parameters to estimate (For details see [4], page 545). Another potential identification method is to rescale the latent trait estimates to make them having zero mean and unity variance, please see [31] for more details.

3. MODEL ESTIMATION USING GIBBS SAMPLING ALGORITHM BASED ON AUXILIARY VARIABLES

3.1 Computational development

Let $\boldsymbol{\Omega} = (\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \boldsymbol{T})$, where $\boldsymbol{\xi}$ represents the set of all the item parameters. The joint posterior distribution of the parameters given the data can be written as follows:

$$\begin{aligned} P(\boldsymbol{\Omega} \,|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}) \propto \prod_{i,j,q,k} & p\,(y_{ijk} \,|\theta_{ijq}, \boldsymbol{\xi}_k)\, p\left(\boldsymbol{\theta}_{ij} \,|\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_e, \boldsymbol{X}_{ij}\right) \\ & \times p\left(\boldsymbol{\beta}_j \,|\boldsymbol{\gamma}_q, \boldsymbol{T}_q, \boldsymbol{W}_j\right) p\left(\boldsymbol{\gamma}_q \,|\boldsymbol{T}_q\right) p\,(\boldsymbol{\xi}_k) \\ & \times p\,(\boldsymbol{\Sigma}_e)\, p\,(\boldsymbol{T}_q)\,. \end{aligned} \tag{4}$$

To implement the Gibbs sampling algorithm based on auxiliary variables, we introduce two mutually independent random variables λ_{ijk} and η_{ijk}. The auxiliary variables λ_{ijk} and η_{ijk} are assumed to follow the uniform distribution $U\,(0, 1)$. The motivation for the algorithm is that the inferred samples can easily be drawn from the full conditional distribution by introducing the auxiliary variables [5, 9, 13, 29]. The following two cases must be satisfied.

- Case 1: When $y_{ijk} = 1$, an equivalent condition for $y_{ijk} = 1$ is that the indicator function $I\,(0 < \lambda_{ijk} \leq p_{ijk})$ must be equal to 1, as opposed to $0 < \eta_{ijk} \leq \psi_{ijk}$ being set to 0. In addition, if the auxiliary variable λ_{ijk} is integrated out of the joint distribution of λ_{ijk} and p_{ijk}, the obtained marginal distribution is just equal to the correct response probability of the ith individual answering the jth item, p_{ijk}. Here $\psi_{ijk} = 1 - p_{ijk}$.
- Case 2: Similarly, when $y_{ijk} = 0$, an equivalent condition for $y_{ijk} = 0$ is that the indicator function $I\,(0 < \eta_{ijk} \leq \psi_{ijk})$ must be equal to 1, as opposed to $I\,(0 < \lambda_{ijk} \leq p_{ijk})$ being set to 0.

The joint posterior distribution after introducing the auxiliary variables λ_{ijk} and η_{ijk} can be written as:

$$\begin{aligned} P\,(\boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\lambda} \,|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}) \propto \prod_{i,j,q,k} & [I\,(y_{ijk} = 1)\, I\,(0 < \lambda_{ijk} \leq p_{ijk}) \\ & + I\,(y_{ijk} = 0)\, I\,(0 < \eta_{ijk} \leq \psi_{ijk})] \\ & \times p\left(\boldsymbol{\theta}_{ij} \,|\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_e, \boldsymbol{X}_{ij}\right) p\left(\boldsymbol{\beta}_j \,|\boldsymbol{\gamma}_q, \boldsymbol{T}_q, \boldsymbol{W}_j\right) \\ & \times p\left(\boldsymbol{\gamma}_q \,|\boldsymbol{T}_q\right) p\,(\boldsymbol{\xi}_k)\, p\,(\boldsymbol{\Sigma}_e)\, p\,(\boldsymbol{T}_q)\,. \end{aligned} \tag{5}$$

The Gibbs sampling algorithm based on auxiliary variables requires sampling from full conditional distributions in turn:

$$\langle 1 \rangle\ [\lambda_{ijk} \,|\boldsymbol{Y}, \boldsymbol{\Omega}]\,, [\eta_{ijk} \,|\boldsymbol{Y}, \boldsymbol{\Omega}]\,;$$

$$\langle 2 \rangle\ [b_k \,|\boldsymbol{Y}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\Omega}]\,;$$

$$\langle 3 \rangle\ \left[a_{kq} \,|\boldsymbol{Y}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{a}_{(-kq)}, \boldsymbol{b}\right]\,;$$

$$\langle 4 \rangle\ \left[\theta_{ijq} \,|\boldsymbol{Y}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\theta}_{ij(-q)}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_e\right]\,;$$

$\langle 5 \rangle\, [\boldsymbol{T}\,|\boldsymbol{\theta}_{ij}, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \boldsymbol{T}]\,;$

$\langle 6 \rangle\, [\gamma\,|\boldsymbol{\beta}_j, \boldsymbol{T}]\,;$

$\langle 7 \rangle\, [\boldsymbol{\Sigma}_e\,|\boldsymbol{\theta}, \boldsymbol{\beta}]\,;$

$\langle 8 \rangle\, [\boldsymbol{T}_q\,|\boldsymbol{\beta}_{jq}, \boldsymbol{\gamma}_q]\,.$

For $\langle 2 \rangle$, we can get a constraint interval about the difficulty parameter by solving the inequality which is constructed by introducing the auxiliary variables. The samples are thus drawn from a truncated prior distribution. Suppose that the prior of the difficulty parameters is $b_k \sim N\left(\mu_b, \sigma_b^2\right)$. According to case 1, given item k, $\forall i, j$, when $y_{ijk} = 1$, we have $0 < \lambda_{ijk} \leq p_{ijk}$, and the following inequalities are established:

$$\sum_{q=1}^{Q} a_{kq}\theta_{ijq} - b_k \geq \log\left(\frac{\lambda_{ijk}}{1-\lambda_{ijk}}\right) \text{ or equivalently}$$

$$b_k \leq \sum_{q=1}^{Q} a_{kq}\theta_{ijq} - \log\left(\frac{\lambda_{ijk}}{1-\lambda_{ijk}}\right). \tag{6}$$

In the same way, for case 2, when $y_{ijk} = 0$, we have $0 < \eta_{ijk} \leq \psi_{ijk}$, and the inequalities are established:

$$\sum_{q=1}^{Q} a_{kq}\theta_{ijq} - b_k \leq \log\left(\frac{1-\eta_{ijk}}{\eta_{ijk}}\right) \text{ or equivalently}$$

$$b_k \geq \sum_{q=1}^{Q} a_{kq}\theta_{ijq} - \log\left(\frac{1-\eta_{ijk}}{\eta_{ijk}}\right). \tag{7}$$

Let $\mathbb{D}_k = \{(i,j)|y_{ijk} = 1, \lambda_{ijk} \leq p_{ijk}\}$ and $\mathbb{E}_k = \{(i,j)|y_{ijk} = 0, \eta_{ijk} \leq \psi_{ijk}\}$. Given the response variable $\boldsymbol{Y}$, the auxiliary variables $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$, and the parameters $\boldsymbol{\Omega}$, the full conditional distribution can be written as:

$$b_k\,|\boldsymbol{Y}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\Omega} \sim N\left(\mu_b, \sigma_b^2\right) I\left(b_k^{\mathrm{L}} \leq b_k \leq b_k^{\mathrm{U}}\right). \tag{8}$$

where

$$b_k^{\mathrm{L}} = \max_{(i,j)\in\mathbb{E}_k} \left\{\sum_{q=1}^{Q} a_{kq}\theta_{ijq} - \log\left(\frac{1-\eta_{ijk}}{\eta_{ijk}}\right)\right\},$$

$$b_k^{\mathrm{U}} = \min_{(i,j)\in\mathbb{D}_k} \left\{\sum_{q=1}^{Q} a_{kq}\theta_{ijq} - \log\left(\frac{\lambda_{ijk}}{1-\lambda_{ijk}}\right)\right\}.$$

For $\langle 3 \rangle$ and $\langle 4 \rangle$, we can apply the identity technique to draw samples. The details of the above Gibbs sampling algorithm based on auxiliary variables and its corresponding Matlab code are provided as online supplementary materials (http://intlpress.com/site/pub/pages/journals/items/sii/content/vols/0012/0001/s001).

4. MODEL SELECTION

4.1 Pseudo-Bayes factor based on cross validation to assess MMIRT models

Within the Bayesian framework, the Bayes factor has played a major role in assessing the goodness of fit of competing models [22]. It is a good choice to compare two fitted models after the model parameters have been estimated. The best-fitting model is chosen based on the largest value of marginal likelihood among a set of candidate models. However, one of the obstacles to use the Bayes factors is the difficulty associated with calculating them. As is well-known, while the candidate models with high-dimensional parameters are used to fit the data, it is impossible to integrate out all the model parameters to obtain the closed-form expression of marginal distribution. In addition, it is acutely sensitive to the choice of prior distributions. The use of improper priors for the parameters in alternative models results in Bayes factors that are not well defined. Therefore, numerous approaches have been proposed to solve the above-mentioned problems ranging from the use of various pseudo-Bayes factor (PsBF) approaches [15]. In this study, the PsBF approach on the basis of the cross-validation predictive density (CVPD) is used to compare the MM2PLMs. Considering the ith individual within the jth school response to the kth item, the CVPD is defined as:

$$p\Big(y_{ijk}\,\Big|\boldsymbol{y}_{-(ijk)}\Big) = \int p\left(y_{ijk}\,\Big|\boldsymbol{y}_{-(ijk)}, \boldsymbol{\Omega}\right) p\left(\boldsymbol{\Omega}\,\Big|\boldsymbol{y}_{-(ijk)}\right) d\boldsymbol{\Omega}, \tag{9}$$

where $\boldsymbol{y}_{-(ijk)}$ denotes the observed data without the ijkth observation. $\boldsymbol{\Omega} = (\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \boldsymbol{T})$ indicates the model parameters. According to the conditional independence hypothesis, the equation $p\left(y_{ijk}\,\Big|\boldsymbol{y}_{-(ijk)}, \boldsymbol{\Omega}\right) = p\left(y_{ijk}\,|\boldsymbol{\Omega}\right)$ can be established, and the responses on the different items are independent given that the latent traits and the responses of students are independent of one another. Therefore, the CVPD can be rewritten as

$$p\left(y_{ijk}\,\Big|\boldsymbol{y}_{-(ijk)}\right) = \int p\left(y_{ijk}\,|\boldsymbol{\Omega}\right) p\left(\boldsymbol{\Omega}\,\Big|\boldsymbol{y}_{-(ijk)}\right) d\boldsymbol{\Omega}. \tag{10}$$

The PsBF for comparing two models (say, $\boldsymbol{\Xi}_1$ and $\boldsymbol{\Xi}_2$) is expressed in terms of the CVPDs, that is,

$$\text{PsBF} = \prod_{i,j,k} \frac{p\left(y_{ijk}\,\Big|\boldsymbol{y}_{-(ijk)}, \boldsymbol{\Xi}_1\right)}{p\left(y_{ijk}\,\Big|\boldsymbol{y}_{-(ijk)}, \boldsymbol{\Xi}_2\right)}. \tag{11}$$

In practice, we often calculate the logarithms of the numerator and denominator of the PsBF to compare different competing models. [14, 30] proposed an importance sampling method to evaluate the marginal likelihood of the data. For $m = 1, \cdots, M$, the samples $\boldsymbol{\Omega}^{(m)}$ from the posterior distribution $p\left(\boldsymbol{\Omega}\,\Big|\boldsymbol{y}_{-(ijk)}\right)$ are easily obtained via an MCMC sampler, where m indicates the index of the MCMC samples. The estimated CVPD result can be written as follows:

$$\hat{p}\left(y_{ijk}\,\middle|\,\boldsymbol{y}_{-(ijk)}\right)=\left[\frac{1}{M}\sum_{m=1}^{M}\frac{1}{p_{ijk}^{(m)}\left(1-p_{ijk}^{(m)}\right)}\right]^{-1}. \tag{12}$$

4.2 Information criteria to assess structural multilevel models

As is known, the natural logarithm transform (abbreviated to log) of the complete-data likelihood of the MM2PLM consists of two parts, one part from the multidimensional item response model and the other part from structural multilevel model. It can be written as follows:

$$\begin{aligned}&\log p\left(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{W},\boldsymbol{\theta},\boldsymbol{\beta}\,|\boldsymbol{\xi},\boldsymbol{\Sigma}_e,\boldsymbol{\gamma},\boldsymbol{T}\right)\\&=\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\underbrace{\sum_{k=1}^{K}\log p\left(y_{ijk}\,|\boldsymbol{\theta}_{ij},\boldsymbol{\xi}_k\right)}_{\text{Multidimensional item response part}}\right.\\&\left.\quad+\underbrace{\log p\left(\boldsymbol{\theta}_{ij}\,|\boldsymbol{\beta}_j,\boldsymbol{\Sigma}_e,\boldsymbol{X}_{ij}\right)+\log p\left(\boldsymbol{\beta}_j\,|\boldsymbol{\gamma},\boldsymbol{T},\boldsymbol{W}_j\right)}_{\text{Structural multilevel part}}\right).\end{aligned}\tag{13}$$

The aim is to make it possible for the changes of the structural multilevel part to be tested conditional on the item response part such that relatively small changes in the log-likelihood of the multilevel part can be detected. In practice, we mainly focus on comparing the MM2PLMs with different multilevel parts and equivalent multidimensional item response parts. Hence, the likelihood of a multilevel structure that has integrated out the random effect $\boldsymbol{\beta}$ is defined as

$$p(\boldsymbol{\theta}\,|\boldsymbol{\gamma},\boldsymbol{\Sigma}_e,\boldsymbol{T},\boldsymbol{X},\boldsymbol{W})=\int p(\boldsymbol{\theta}\,|\boldsymbol{\beta},\boldsymbol{\Sigma}_e,\boldsymbol{X})\,p\left(\boldsymbol{\beta}\,|\boldsymbol{\gamma},\boldsymbol{T},\boldsymbol{W}\right)d\boldsymbol{\beta}.$$

Therefore, the deviance can be defined as

$$\begin{aligned}D\left(\boldsymbol{\Omega}^*\right)=&\,N\cdot Q\log\left(2\pi\right)+N\log\left|\boldsymbol{\Sigma}_e\right|-\sum_{j=1}^{J}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{\beta}_j}\right|+J\log\left|\boldsymbol{T}\right|\\&+\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\boldsymbol{\theta}_{ij}-\boldsymbol{X}_{ij}\tilde{\boldsymbol{\beta}}_j\right)'\boldsymbol{\Sigma}_e^{-1}\left(\boldsymbol{\theta}_{ij}-\boldsymbol{X}_{ij}\tilde{\boldsymbol{\beta}}_j\right)\\&+\sum_{j=1}^{J}\left(\tilde{\boldsymbol{\beta}}_j-\boldsymbol{w}_j\boldsymbol{\gamma}\right)'\boldsymbol{A}^{-1}\left(\tilde{\boldsymbol{\beta}}_j-\boldsymbol{w}_j\boldsymbol{\gamma}\right),\end{aligned}$$

where $\boldsymbol{\Omega}^*=(\boldsymbol{\gamma},\boldsymbol{\Sigma}_e,\boldsymbol{T})$, and $\boldsymbol{A}=\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_j}+\boldsymbol{T}$ with $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_j}=\left(\sum_{i=1}^{n_j}\boldsymbol{X}_{ij}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{X}_{ij}\right)^{-1}$. The posterior mean and covariance of random regression coefficient are $\tilde{\boldsymbol{\beta}}_j=\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_j}\sum_{i=1}^{n_j}\boldsymbol{X}_{ij}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{\theta}_{ij}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_j}=\left(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_j}^{-1}+\boldsymbol{T}^{-1}\right)$. Three information criteria are widely used for model assessment, that is, the Aikaike information criteria (AIC; 'Akaike', 1973), the Bayesian information criteria (BIC; 'Schwarz', 1978) and the deviance information criteria (DIC; 'Spiegelhalter' et al., 2002), the forms can be shown as follows.

$$\begin{aligned}\text{AIC}=&\overline{D\left(\boldsymbol{\Omega}^*\right)}+2\rho,\\\text{BIC}=&\overline{D\left(\boldsymbol{\Omega}^*\right)}+\rho\log N,\\\text{DIC}=&2\overline{D\left(\boldsymbol{\Omega}^*\right)}-D\left(\widehat{\boldsymbol{\Omega}^*}\right),\end{aligned}$$

where $\overline{D\left(\boldsymbol{\Omega}^*\right)}$ is the estimated posterior mean deviance [3, 38]. $D\left(\widehat{\boldsymbol{\Omega}^*}\right)$ is the deviance for the posterior mean of the parameter values. They can be computed using the output from an MCMC sampling scheme. ρ is the total number of parameters. N is the total number of individuals.

5. SIMULATION

5.1 Simulation study 1

This simulation study is performed to validate the model specification (such as the selection of prior distributions) and evaluate the parameter recovery with Gibbs sampling algorithm based on auxiliary variables. For illustration, we only consider one explanatory variable on both levels, and the number of dimensions is fixed at 2 ($q=2$). The true model is the following structural multilevel model.

The individual-level model:

$$\theta_{ijq}=\beta_{0jq}+x_{ij}\beta_{1jq}+e_{ijq}, \tag{14}$$

where

$$\boldsymbol{e}=\begin{pmatrix}e_{ij1}\\e_{ij2}\end{pmatrix}\sim N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma_{e_1}^2&\sigma_{e_1e_2}\\\sigma_{e_2e_1}&\sigma_{e_2}^2\end{pmatrix}\right). \tag{15}$$

The school-level model:

$$\begin{aligned}\beta_{0jq}&=\gamma_{00q}+\gamma_{01q}w_j+u_{0jq},\\\beta_{1jq}&=\gamma_{10q}+\gamma_{11q}w_j+u_{1jq},\end{aligned}\tag{16}$$

where

$$\begin{pmatrix}u_{0jq}\\u_{1jq}\end{pmatrix}\sim N\left(\begin{pmatrix}0\\0\end{pmatrix},\boldsymbol{T}\right),\ \boldsymbol{T}=\begin{pmatrix}\tau_{00q}&\tau_{01q}\\\tau_{10q}&\tau_{11q}\end{pmatrix}. \tag{17}$$

The multidimensional two-parameter logistic model is used to generate responses. The test length is set to 30. The true values of the discrimination and difficulty parameters are generated from truncated normal distribution and standard normal distribution, i.e. $a_{kq}\sim N\left(1.5,1\right)I\left(a_{kq}>0\right)$, $q=1,2$, and $b_k\sim N\left(0,1\right)$, respectively. The ability parameters of 2,000 students from population $N\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}_j,\boldsymbol{\Sigma}_e\right)$ are divided into $J=10$ groups, with n_j students in each group.

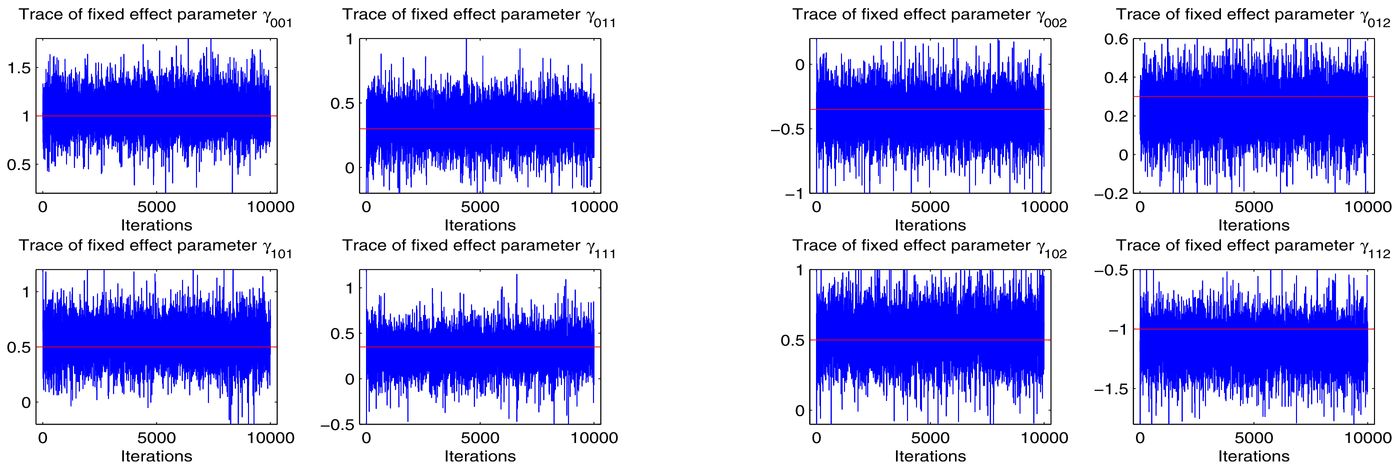

Figure 1. Trace plots of fixed effects for simulation study 1.

The fixed effect $\boldsymbol{\gamma}$ is chosen as an arbitrary value between -1 and 1. For simplicity, we suppose that at level 3, each of the dimensional covariances τ_{01q} and τ_{10q} is equal to 0 for $q = 1, 2$, which means that the level-3 residuals between random coefficients $\boldsymbol{\beta}_q = (\beta_{0jq}, \beta_{01jq})$ are independent of each other. The level-3 variances τ_{00q} and τ_{11q} are respectively set equal to 0.250 and 0.200, for $q = 1, 2$ such that they have very low stochastic volatility in the vicinity of the level-3 mean. The level-2 residual variance-covariance (VC) are set to 0.300, 0.500, and 0.075. The explanatory variables $\boldsymbol{X}$ and $\boldsymbol{W}$ are drawn from $N(0.25, 1)$ and $N(0.5, 1)$, respectively. The priors to the discrimination parameters and difficulty parameters are set as the non-informative priors $\boldsymbol{a}_k \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}\right) I\left(\boldsymbol{a}_k \left| a_{k1} > 0, a_{k2} > 0\right.\right)$ and $N(0, 100)$. The fixed effect $\boldsymbol{\gamma}$ follows a uniform distribution $U(-2, 2)$. The prior to the VC matrix of the level-2 ability dimensions is a 2-by-2 identity matrix. As used in many educational and psychological research studies (see, e.g., [12, 23, 36]), and the priors to the VC matrices of the level-3, $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$, are set to the non-informative priors based on Fox (2001)'s paper (see, [12]), where $p(\boldsymbol{T}_q) \propto 1$, $q = 1, 2$.

The convergence of Gibbs sampler based on auxiliary variables is checked by monitoring the trace plots of the parameters for consecutive sequences of 10,000 iterations. Figure 1 represents the trace plots of the fixed-effect parameters. The trace plots show that all parameter estimates stabilize after 5,000 iterations and then converge quickly. Thus, we set the first 5,000 iterations as the burn-in period. 500 replicas were generated. The true values, the averaged estimated values and the coverage probabilities (CPs) based on the 95% highest posterior density intervals (HPDIs) for item parameters are shown in Table 1. Table 2 presents the true values and the estimated values of fixed effects $\boldsymbol{\gamma}$, level-2 covariance components, and level-3 variance components $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$.

The accuracy of the parameter estimates is measured by two evaluation indexes, namely, Bias and root mean squared error (RMSE). The recovery results are based on 500 times MCMC repeated iterations. The results of the accuracy of the parameter estimates are displayed in Tables 3 and 4. From Tables 3 and 4, we see that Gibbs sampling algorithm based on auxiliary variables provides accurate estimates of the structural parameters in the sense of having small Bias and RMSE values.

5.2 Simulation study 2

The aim of this simulation is twofold. First, we test the changes of the structural multilevel part conditional on the item response part such that relatively small changes in the log-likelihood of the multilevel part can be detected. Second, we evaluate different MM2PLMs by their prediction power, specifically using the PsBF approach based on the CVPDs. We simulate data from the same multilevel IRT model used in simulation study 1. First, two competing models are estimated using the simulated data sets to investigate the performance of the information criteria of the structural multilevel model comparison, where the observed sum scores of 1,000 students on 30 items are imputed for two-dimensional latent trait. The first alternative model (model 1) is the structural multilevel model with a level-2 explanatory variable, that is,

$$\text{(18)} \qquad \textbf{Model 1.} \quad \begin{cases} \theta_{ijq} = \beta_{0jq} + x_{ij}\beta_{1jq} + e_{ijq}, \\ \qquad \beta_{0jq} = \gamma_{00q} + u_{0jq}, \\ \qquad \beta_{1jq} = \gamma_{10q} + u_{1jq}. \end{cases}$$

Model 2 is an extension of model 1. Considering one explanatory variable on both levels, that is,

$$\text{(19)} \qquad \textbf{Model 2.} \quad \begin{cases} \theta_{ijq} = \beta_{0jq} + x_{ij}\beta_{1jq} + e_{ijq}, \\ \beta_{0jq} = \gamma_{00q} + \gamma_{01q}w_j + u_{0jq}, \\ \beta_{1jq} = \gamma_{10q} + \gamma_{11q}w_j + u_{1jq}. \end{cases}$$

Table 1. Estimation of simulated item parameters using Gibbs sampling algorithm based on auxiliary variables

	a_{k1}			a_{k2}			b_k		
Item	True	Estimated	CP	True	Estimated	CP	True	Estimated	CP
1	1^*	1^*	–	0^*	0^*	–	0^*	0^*	–
2	0^*	0^*	–	1^*	1^*	–	0^*	0^*	–
3	1.203	1.149	0.980	0.974	1.001	0.962	0.855	0.879	0.964
4	0.529	0.561	0.970	0.744	0.787	0.952	−0.297	−0.316	0.950
5	1.010	0.936	0.980	0.833	0.854	0.940	1.537	1.488	0.976
6	0.981	0.954	0.958	0.682	0.665	0.940	0.021	−0.015	0.966
7	0.602	0.599	0.972	1.059	0.973	0.956	−0.392	−0.361	0.950
8	1.205	1.184	0.942	0.687	0.667	0.940	−0.644	−0.583	0.910
9	0.922	0.944	0.970	1.338	1.319	0.948	−0.523	−0.551	0.984
10	0.788	0.754	0.988	0.556	0.480	0.930	−0.079	−0.143	0.950
11	0.625	0.617	0.942	0.536	0.553	0.960	−0.122	−0.105	0.940
12	0.568	0.621	0.950	0.735	0.692	0.960	−0.215	−0.186	0.950
13	1.067	1.083	0.934	0.790	0.718	0.924	0.971	0.904	0.952
14	0.806	0.785	0.940	1.207	1.188	0.926	0.313	0.399	0.942
15	0.872	0.830	0.960	0.964	0.879	0.950	0.150	0.056	0.946
16	1.078	1.096	0.962	0.615	0.630	0.972	0.489	0.543	0.940
17	0.762	0.794	0.960	1.436	1.480	0.970	−1.089	−1.110	0.978
18	1.332	1.367	0.980	1.308	1.289	0.960	0.898	0.849	0.970
19	1.194	1.213	0.974	0.680	0.656	0.920	0.138	0.201	0.948
20	1.418	1.408	0.976	1.213	1.131	0.950	−0.383	−0.372	0.960
21	1.063	1.027	0.996	1.466	1.483	0.956	−0.619	−0.739	0.940
22	0.429	0.443	0.962	0.619	0.570	0.940	−0.728	−0.762	0.946
23	0.644	0.591	0.920	1.317	1.388	0.976	−0.792	−0.753	0.912
24	1.181	1.139	0.978	0.698	0.723	0.978	−1.982	−1.996	0.950
25	0.947	0.917	0.976	1.038	0.980	0.962	0.149	0.175	0.964
26	1.544	1.582	0.960	0.780	0.863	0.960	−1.714	−1.679	0.982
27	1.380	1.357	0.924	0.631	0.641	0.960	−1.450	−1.387	0.980
28	0.771	0.759	0.932	1.040	1.113	0.932	0.274	0.246	0.970
29	1.129	1.218	0.952	1.338	1.432	0.962	−1.084	−1.178	0.978
30	0.736	0.807	0.940	0.924	0.871	0.950	0.639	0.655	0.940

Note: Asterisks (*) indicate the constraints for model identification. CP denotes the coverage probability computed from the 500 95% highest posterior density intervals.

Table 2. Parameter estimates of the two-dimensional fixed effects using Gibbs sampling algorithm based on auxiliary variables in simulation study 1

Fixed effect	True	Estimated	CP	Fixed effect	True	Estimated	CP
γ_{001}	1.000	1.037	0.970	γ_{002}	−0.350	−0.371	0.942
γ_{011}	0.300	0.315	0.940	γ_{012}	0.300	0.282	0.960
γ_{101}	0.500	0.546	0.940	γ_{102}	0.500	0.562	0.926
γ_{111}	0.350	0.339	0.952	γ_{112}	−1.000	−1.017	0.946

Level-2 random effect	True	Estimated	CP
$\sigma^2_{e_1}$	0.300	0.304	0.964
$\sigma_{e_1e_2}$	0.075	0.056	0.970
$\sigma_{e_2e_1}$	0.075	0.056	0.970
$\sigma^2_{e_2}$	0.500	0.472	0.948

Level-3 T_1	True	Estimated	CP	Level-3 T_2	True	Estimated	CP
τ_{001}	0.250	0.274	0.964	τ_{002}	0.250	0.232	0.966
τ_{011}	0	0.039	0.942	τ_{012}	0	0.011	0.970
τ_{101}	0	0.039	0.942	τ_{102}	0	0.011	0.970
τ_{111}	0.200	0.212	0.930	τ_{112}	0.200	0.208	0.968

Table 3. Evaluating the accuracy of item parameter estimation

	a_{k1}			a_{k2}			b_k		
Item	True	Bias	RMSE	True	Bias	RMSE	True	Bias	RMSE
1	1^*	0	0	0^*	0	0	0^*	0	0
2	0^*	0	0	1^*	0	0	0^*	0	0
3	1.203	−0.054	0.115	0.974	0.027	0.114	0.855	0.024	0.125
4	0.529	0.032	0.108	0.744	0.043	0.121	−0.297	−0.019	0.118
5	1.010	−0.074	0.129	0.833	0.021	0.137	1.537	−0.049	0.024
6	0.981	−0.027	0.137	0.682	−0.017	0.052	0.021	−0.036	0.132
7	0.602	−0.003	0.109	1.059	−0.086	0.207	−0.392	0.031	0.101
8	1.205	−0.021	0.125	0.687	−0.020	0.137	−0.644	0.061	0.081
9	0.922	0.022	0.126	1.338	−0.019	0.105	−0.523	−0.028	0.065
10	0.788	−0.034	0.210	0.556	−0.076	0.149	−0.079	−0.064	0.136
11	0.625	−0.008	0.136	0.536	0.017	0.082	−0.122	0.017	0.023
12	0.568	0.053	0.143	0.735	−0.043	0.168	−0.215	0.029	0.143
13	1.067	0.016	0.142	0.790	−0.072	0.126	0.971	−0.067	0.148
14	0.806	−0.021	0.083	1.207	−0.019	0.099	0.313	0.086	0.130
15	0.872	−0.042	0.213	0.964	−0.085	0.107	0.150	−0.094	0.119
16	1.078	0.018	0.024	0.615	0.015	0.139	0.489	0.054	0.175
17	0.762	0.032	0.104	1.436	0.044	0.154	−1.089	−0.021	0.039
18	1.332	0.035	0.111	1.308	−0.019	0.048	0.898	−0.049	0.125
19	1.194	0.019	0.085	0.680	−0.024	0.059	0.138	0.063	0.124
20	1.418	−0.010	0.128	1.213	−0.082	0.174	−0.383	0.011	0.086
21	1.063	−0.036	0.087	1.466	0.017	0.123	−0.619	−0.120	0.096
22	0.429	0.014	0.128	0.619	−0.049	0.121	−0.728	−0.034	0.123
23	0.644	−0.053	0.187	1.317	0.071	0.164	−0.792	0.039	0.138
24	1.181	−0.042	0.151	0.698	0.025	0.126	−1.982	−0.014	0.104
25	0.947	−0.030	0.175	1.038	−0.058	0.072	0.149	0.026	0.137
26	1.544	0.038	0.129	0.780	0.083	0.192	−1.714	0.035	0.130
27	1.380	−0.023	0.156	0.631	0.010	0.125	−1.450	0.063	0.131
28	0.771	−0.012	0.076	1.040	0.073	0.250	0.274	−0.028	0.084
29	1.129	0.089	0.217	1.338	0.094	0.118	−1.084	−0.094	0.107
30	0.736	0.071	0.123	0.924	−0.053	0.137	0.639	0.016	0.162

Table 4. Evaluating the accuracy of the two-dimensional fixed effects and variance-covariance components

Fixed effect	True	Bias	RMSE	Fixed effect	True	Bias	RMSE
γ_{001}	1.000	0.027	0.140	γ_{002}	−0.350	−0.021	0.108
γ_{011}	0.300	0.015	0.124	γ_{012}	0.300	−0.018	0.152
γ_{101}	0.500	0.046	0.109	γ_{102}	0.500	0.062	0.131
γ_{111}	0.350	−0.011	0.101	γ_{112}	−1.000	−0.017	0.116
Level-2 random effect			True		Bias		RMSE
$\sigma^2_{e_1}$			0.300		0.004		0.043
$\sigma_{e_1 e_2}$			0.075		−0.019		0.119
$\sigma_{e_2 e_1}$			0.075		−0.019		0.119
$\sigma^2_{e_2}$			0.500		−0.028		0.161
Level-3 T_1	True	Bias	RMSE	Level-3 T_2	True	Bias	RMSE
τ_{001}	0.250	0.024	0.146	τ_{002}	0.250	−0.018	0.105
τ_{011}	0	0.039	0.179	τ_{012}	0	0.011	0.118
τ_{101}	0	0.039	0.179	τ_{102}	0	0.011	0.118
τ_{111}	0.200	0.012	0.046	τ_{112}	0.200	0.008	0.039

Table 5 presents the results of a model comparison with the averaged AIC, BIC and DIC values across 500 replications. Both AIC and DIC choose model 2 as the better-fitting model compared with model 1. The difference in the averaged AIC is 886.443 between model 1 and model 2. The difference in the averaged DIC is 964.271 between model 1 and model 2. However, BIC prefers the simpler model (model 1). The difference in the averaged BIC is 863.421 between model 2 and model 1. Relatively speaking, the complex multilevel structural model better fits the simulated data than the simple one. Moreover, the cross-validation log-likelihoods across 500 replications are used to assess overall MM2PLMs. From Table 6, we find that Model 2⊕M2PLM is better-fitting than Model 1⊕M2PLM. The difference in the averaged cross-validation log-likelihood (which is equivalent to the log-PsBF) is 4,148.132.
Note: Model 1⊕M2PLM denotes model 1 with a multidimensional two-parameter logistic model.

Table 5. Multilevel structural model comparison using information criteria for the simulated data

Model	AIC	BIC	DIC
Model 1	34,847.759	34,063.242	35,152.568
Model 2	33,961.316	34,926.663	34,188.297

Table 6. Overall evaluation of the multilevel IRT model based on the cross-validation (CV) log-likelihood method

Model	CV log-likelihood (log-PsBF)
Model 1⊕M2PLM	−66,867.386
Model 2⊕M2PLM	−62,719.254

5.3 Simulation study 3

The purpose of this simulation study is to verify whether the algorithm can guarantee the accuracy of parameter estimation for the various numbers of individuals and items. The simulation design is as follows: The number of dimensions is fixed at 4. The multidimensional two-parameter logistic model is used to generate responses. Two factors and their varied conditions are considered: (a) three different numbers of individuals $N =$ 1,000, 2,000, or 3,000; (b) number of items, $K =$ 40,100 or 200, and for per subtest number of items 10, 25 or 50. Fully crossing the different levels of these two factors yields 9 conditions. Individuals ($N = 1{,}000, 2{,}000, 3{,}000$) are equally distributed to 10 schools ($J = 10$). True values of item parameters and priors of all parameters are generated as in simulation study 1. The true values of the fixed effects are 1.000 (γ_{00q}), 0.300 (γ_{01q}), 0.500 (γ_{10q}) and 0.350 (γ_{11q}), $q = 1, \cdots, 4$, respectively, and the level-2 variances are 0.300 $\left(\sigma^2_{e_1}\right)$, 0.500 $\left(\sigma^2_{e_2}\right)$, 0.750 $\left(\sigma^2_{e_3}\right)$ and 1.000 $\left(\sigma^2_{e_4}\right)$, and the covariances are set to 0.075. The level-3 variances are respectively 0.250 and 0.200 (τ_{00q}, τ_{11q}), and the covariances are 0 (τ_{01q}, τ_{10q}). The multilevel structural model (Equation 19) in simulation study 2 are used, but the dimensions are fixed at 4.

The accuracy of the parameter estimates is measured by two evaluation indexes, namely, Bias and RMSE. The recovery results are based on the MCMC iterations repeated 500 times. The biases are −0.097∼0.103 for the fixed effect parameters, −0.032∼0.093 for the level-2 variance-covariance component parameters, and −0.064∼0.108 for the level-3 variance-covariance component parameters. The RMSEs are 0.169∼0.273 for the fixed effect parameters, 0.128∼0.267 for the level-2 variance-covariance component parameters, and 0.153∼0.231 for the level-3 variance-covariance component parameters. Furthermore, the Bias and RMSE have a smaller trend with the increase in the number of individuals and items; in other words, increasing the number of individuals and items helps to improve the estimation accuracy of the structural parameters. In summary, the sampling algorithm is effective for various numbers of individuals and items.

6. ANALYSIS OF THE EDUCATION QUALITY ASSESSMENT DATA

6.1 Purpose

To illustrate the applicability of the MM2PLM method in operational large-scale assessments, we consider a data set about students' English achievement test for junior middle schools conducted by NENU Branch, Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University. The analysis of the test data will help us to gain a better understanding of the practical situation of students' English academic latent traits and to explore the factors that affect their English academic latent traits. The results of this analysis will be potentially very valuable for development and improvement of educational quality monitoring mechanism in China.

6.2 Sampling design

The test data contain a two-stage cluster sample of 2,108 students in grade 2 of junior middle school. These students are from 16 schools, with 121 to 139 students in each school. In the first stage, the sampling population is classified according to district, and schools are selected at random. In the second stage, students are selected at random from each school. The English test battery consists of four subscales: vocabulary (40 items), grammar (24 items), comprehensive reading (40 items), and table computing (20 items). All 124 multiple-choice items were scored using a dichotomous scale. The Cronbach's alpha coefficients for vocabulary, grammar, reading comprehension and table computing items are 0.942, 0.875, 0.843, and 0.816, respectively. Level-2 and level-3 background covariates of individuals, teacher satisfaction, and school climate (teachers and schools constitute level 3) are measured. At the individual level, gender (0 = male, 1 = female) and socioeconomic status are measured; the latter is measured by the average of two indicators: the father's and mother's educational levels, which are five-point Likert items; scores range from 0 to 8. At the teacher and school levels, teacher satisfaction is measured by 20 five-point Likert items, and school environment from the principal's perspective is measured by 23 five-point Likert items. A description of the sampling procedure and the questionnaires can be found in [35]. The Gibbs sampling runs 20,000 iterations for real data, with a burn-in period of 5,000 iterations. The average over the drawn parameters is calculated after the burn-in period.

6.3 Model assessment

We consider four dimensions of latent trait: vocabulary cognitive ability, grammar structure diagnosing ability, reading comprehension ability, and table computing ability. These latent traits are affected by individual covariates such as socioeconomic status (SES) and gender (GD). The individual can be nested into higher group levels (such as schools), which are affected by group covariates such as teacher satisfaction (ST) and school climate (CT) from the teachers' perspective. According to the two model assessment methods mentioned above, three models are considered for fitting the real data. The best-fitting model is eventually used to analyze the data.

At the first level, a multidimensional two-parameter logistic model (IRT model) is used to model the relationship between items, persons, and responses. The different structural multilevel models are represented as follows:

The following model 3 (i.e., structural multilevel model) consists of two level-2 background variables SES and GD and the level-2 random intercept. The effect of level-2 background variables SES and GD are allowed to fix across school. The structural multilevel part is given by

(20)

$$\textbf{Model 3.} \begin{cases} \theta_{ijq} = \beta_{0jq} + SES_{ij}\beta_{1jq} + GD_{ij}\beta_{2jq} + e_{ijq}, \\ \beta_{0jq} = \gamma_{00q} + u_{0jq}, \\ \beta_{1jq} = \gamma_{10q}, \\ \beta_{2jq} = \gamma_{20q}. \end{cases}$$

Model 4 is an extended version by including two latent predictors at level 3: ST and CT. The structural multilevel part is given by

(21)

$$\textbf{Model 4.} \begin{cases} \theta_{ijq} = \beta_{0jq} + SES_{ij}\beta_{1jq} + GD_{ij}\beta_{2jq} + e_{ijq}, \\ \beta_{0jq} = \gamma_{00q} + ST_j\gamma_{01q} + CT_j\gamma_{02q} + u_{0jq}, \\ \beta_{1jq} = \gamma_{10q}, \\ \beta_{2jq} = \gamma_{20q}. \end{cases}$$

When the effects of level-2 background variables SES and GD are allowed to vary across school, we extend model 4 to model 5 with the following structural multilevel part:

(22)

$$\textbf{Model 5.} \begin{cases} \theta_{ijq} = \beta_{0jq} + SES_{ij}\beta_{1jq} + GD_{ij}\beta_{2jq} + e_{ijq}, \\ \beta_{0jq} = \gamma_{00q} + ST_j\gamma_{01q} + CT_j\gamma_{02q} + u_{0jq}, \\ \beta_{1jq} = \gamma_{10q} + u_{1jq}, \\ \beta_{2jq} = \gamma_{20q} + u_{2jq}. \end{cases}$$

First, we focus on which is the best structural multilevel model (model 3, 4, or 5) to fit the real data. The standardized item response total scores are imputed for four-dimensional latent trait. The information criteria can be formulated for choosing between models that differ in the fixed and/or random part of the structural multilevel model. From Table 7, the AIC, BIC, and DIC consistently choose model 3 as the worst-fitting model. The AIC (75,036.875) and DIC (79,527.306) prefer model 5. Model 4 is ranked second by the AIC (79,580.306) and DIC (83,816.179). The BIC (76,386.563) prefers the more parsimonious model 4 to model 5. In addition, the results for selecting the optimal multilevel IRT model based on the cross-validation log-likelihood are presented in Table 8. It can be found that Model 5⊕M2PLM is the best-fitting model compared to the other models, and Model 4⊕M2PLM is ranked second. The differences are 333.738 between Model 5⊕M2PLM and Model 4⊕M2PLM, and 4,216.043 between Model 4⊕2MPLM and Model 3⊕2MPLM. The reason that Model 5⊕M2PLM and Model 4⊕M2PLM are markedly better than Model 3⊕M2PLM can be attributed to the additional latent predictors at level 3, i.e., ST and CT. In summary, model 5 is preferred based on the values of both AIC and DIC for linear multilevel models given the outcome variables. In addition, the log-PsBFs of MM2PLMs show that Model 5⊕M2PLM is preferred under both model assessment methods.

Table 7. Multilevel structural model comparison using the information criteria for real data

Model	AIC	BIC	DIC
Model 1	90,155.324	91,282.027	94,541.657
Model 2	79,580.619	76,386.563	83,816.179
Model 3	75,036.875	79,843.457	79,527.306

Table 8. Selecting the optimal multilevel IRT model using the cross-validation (CV) log-likelihood for real data

Model	CV log-likelihood (log-PsBF)
Model 3⊕M2PLM	$-1{,}161,982.168$
Model 4⊕M2PLM	$-1{,}157,865.125$
Model 5⊕M2PLM	$-1{,}157,522.387$

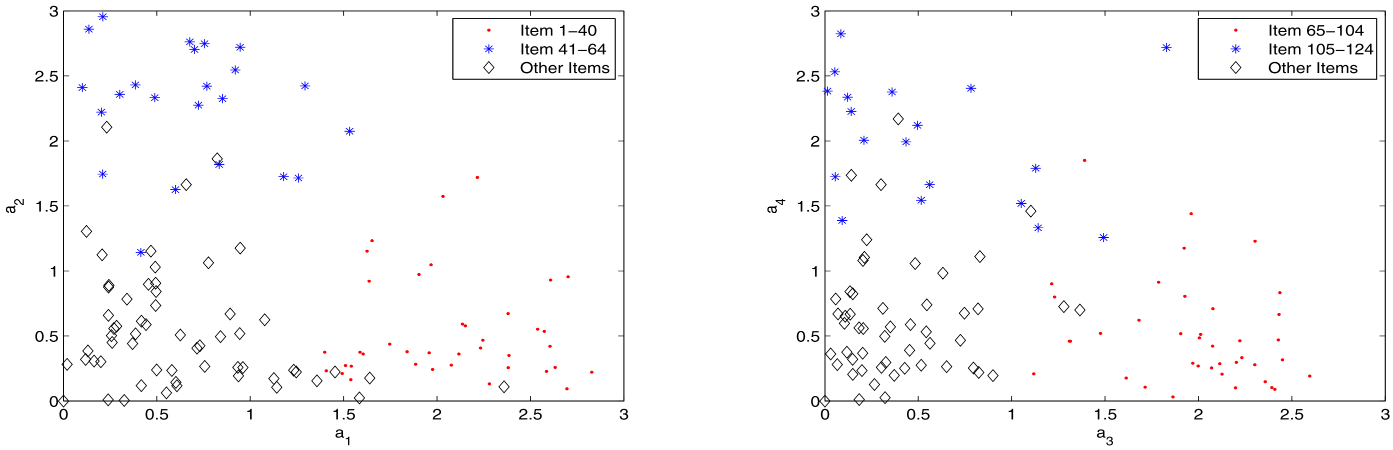

Figure 2. Parameter estimation of a_{k1}, a_{k2}, a_{k3}, and a_{k4} for subscale 1 (item 1-40), subscale 2 (item 41-64), subscale 3 (item 65-104), and subscale 4 (item 105-124).

6.4 Item test dimension evaluation

A test battery contains four subtests, which consists of items that measure four dimensions of latent traits, and the dimension of a latent trait can be measured mainly by a subtest. The expected a posteriori (EAP) estimates of the discrimination parameters are plotted to reveal whether the items' factor patterns reflect the subtest of the test battery in Figure 2. In the left panel of Figure 2, the discrimination parameters of the first two dimensions are plotted for subtest 1 (items marked by a dot) and subtest 2 (items marked by a cross), and the other items are marked by a diamond. The items of subtest 1 (item 1-40) have a high factor loading on the first dimension on average and a low factor loading on the second dimension on average, and the items of subtest 2 (item 41-64) have a high factor loading on the second dimension on average and a low factor loading on the first dimension on average. The other items do not vary appreciably between the two dimensions. The right panel of Figure 2 shows the pattern of the discrimination parameters of the third and fourth subtests on the third and fourth dimensions. By and large, the items of subtest 3 (item 65-104) have a high factor loading on the third dimension and a low factor loading on the fourth dimension, and the items of subtest 4 (item 105-124) have a high factor loading on the fourth dimension and a low factor loading on the third dimension. The overall pattern of the discrimination parameters are used to fit the test battery, demonstrating that each dimension is generally identified by the items of one subtest.

6.5 Estimation of structural parameters

From Tables 9 and 10, we can find that parents' educational level differs by one unit for the male students from the same class and school. The vocabulary cognitive ability, the ability to diagnose grammar structure, reading comprehension ability, and table computing ability have the differences of 0.661, 0.483, 0.562 and 0.393, respectively. In comparison with male students, the differences in the four dimensions of female ability are 1.034, 0.919, 0.806 and 0.106, respectively. The educational level of parents has an effect on the English learning ability of their kids. The parents with a high *SES* values may have more prospective English learning awareness based on their own learning experiences, provide more diversified learning ways, and know how to create a better English learning environment for their kids. In addition, parents with a higher educational level are able to provide learning guidance in English. In general, the higher educational level of parents, the more ability of tutoring the English learning activities of their kids.

For both male and female students from the same class and school with the same *SES* scores, the performance of female students in vocabulary cognitive ability, the ability to diagnose grammar structure, and reading comprehension ability are higher than the performance of male students by 0.373, 0.436, and 0.244, respectively. However, male students have scored higher than female students in table computing ability by 0.287. From the previous comparison, it concludes that female students have the advantage over male students at vivid memory and mechanical memory but not at logical reasoning, deductive induction, and computing ability.

For male students who have the same *SES* scores from different schools, if the difference in teacher satisfaction is taken as a baseline unit, the differences at the levels of vocabulary cognitive ability, the ability to diagnose grammar structure, and reading comprehension ability are 0.562, 0.375 and 0.332, respectively. However, the difference at computing ability's level is very small. Through further analysis, we can find out that teachers' factor has an important impact on students' cognitive ability, the ability to diagnose grammar structure, and reading ability, except for the table computing ability. From this study, we conclude the

Table 9. Parameter estimation of the MMIRT model for vocabulary cognitive ability and grammar structure diagnosing ability

	Vocabulary cognitive ability				Grammar structure diagnosing ability		
Fixed effect	Coefficient	SD	HPDI	Fixed effect	Coefficient	SD	HPDI
γ_{001}	0.775	0.199	$[0.453, 1.101]$	γ_{002}	0.654	0.155	$[0.401, 0.934]$
$\gamma_{011}\,(ST)$	0.562	0.151	$[0.326, 0.827]$	$\gamma_{012}\,(ST)$	0.375	0.127	$[0.171, 0.649]$
$\gamma_{021}\,(CT)$	0.271	0.158	$[0.028, 0.539]$	$\gamma_{022}\,(CT)$	0.104	0.136	$[-0.096, 0.343]$
$\gamma_{101}\,(SES)$	0.661	0.140	$[0.437, 0.914]$	$\gamma_{102}\,(SES)$	0.483	0.161	$[0.218, 0.751]$
$\gamma_{201}\,(GD)$	0.373	0.184	$[0.079, 0.675]$	$\gamma_{202}\,(GD)$	0.436	0.191	$[0.122, 0.765]$
Random effect	Covariance	SD	HPDI	Random effect	Covariance	SD	HPDI
τ_{001}	0.534	0.108	$[0.209, 1.102]$	τ_{002}	0.370	0.181	$[0.174, 0.704]$
τ_{011}	-0.019	0.142	$[-0.238, 0.195]$	τ_{012}	-0.024	0.121	$[-0.215, 0.161]$
τ_{021}	-0.242	0.196	$[-0.592, -0.001]$	τ_{022}	0.055	0.154	$[-0.174, 0.309]$
$\tau_{111}\,(SES)$	0.275	0.144	$[0.117, 0.534]$	$\tau_{112}\,(SES)$	0.232	0.196	$[0.021, 0.584]$
τ_{121}	-0.082	0.127	$[-0.299, 0.094]$	τ_{122}	-0.026	0.151	$[-0.266, 0.203]$
$\tau_{221}\,(GD)$	0.128	0.135	$[-0.135, 0.569]$	$\tau_{222}\,(GD)$	0.167	0.162	$[-0.116, 0.627]$

Table 10. Parameter estimation of the MMIRT model for reading comprehension ability and table computing ability

	Reading comprehension ability				Table computing ability		
Fixed effect	Coefficient	SD	HPDI	Fixed effect	Coefficient	SD	HPDI
γ_{003}	0.763	0.207	$[0.410, 1.116]$	γ_{004}	0.319	0.138	$[0.090, 0.543]$
$\gamma_{013}\,(ST)$	0.332	0.172	$[0.041, 0.628]$	$\gamma_{014}\,(ST)$	0.077	0.118	$[-0.131, 0.287]$
$\gamma_{023}\,(CT)$	0.083	0.197	$[-0.292, 0.401]$	$\gamma_{024}\,(CT)$	0.255	0.108	$[0.068, 0.491]$
$\gamma_{103}\,(SES)$	0.562	0.128	$[0.355, 0.791]$	$\gamma_{104}\,(SES)$	0.393	0.135	$[0.176, 0.627]$
$\gamma_{203}\,(GD)$	0.244	0.190	$[-0.049, 0.573]$	$\gamma_{204}\,(GD)$	-0.287	0.123	$[-0.488, -0.083]$
Random effect	Covariance	SD	HPDI	Random effect	Covariance	SD	HPDI
τ_{003}	0.529	0.160	$[0.149, 1.181]$	τ_{004}	0.294	0.146	$[0.138, 0.562]$
τ_{013}	-0.024	0.135	$[-0.227, 0.186]$	τ_{014}	0.100	0.098	$[-0.026, 0.269]$
τ_{023}	0.014	0.212	$[-0.308, 0.345]$	τ_{024}	-0.065	0.094	$[-0.244, 0.063]$
$\tau_{113}\,(SES)$	0.261	0.115	$[0.131, 0.475]$	$\tau_{114}\,(SES)$	-0.025	0.131	$[0.144, 0.533]$
τ_{123}	-0.040	0.118	$[-0.233, 0.135]$	τ_{124}	-0.026	0.081	$[-0.156, 0.093]$
$\tau_{223}\,(GD)$	0.173	0.156	$[-0.121, 0.652]$	$\tau_{224}\,(GD)$	0.128	0.105	$[0.014, 0.325]$

existence of strong relation between teacher satisfaction factor and sense of responsibility factor at junior middle school, and it can be explained by the work environment that maintains enthusiasm of education and teaching, and inspire students' learning motivation. This has a great improvement at level of the students' vocabulary cognitive ability, the ability to grammatical structure analysis, and reading comprehension ability owing to teachers' teaching attitude and responsibility. However, the improvement for the table computing ability is small. It is possible to play a decisive role in the students' internal factors as compared with the teachers' external factors.

Tables 9 and 10 present the estimated results of the correlations between ability type for the latent dimensions and the covariates of different levels. The estimated values for school climate effects γ_{02q} are 0.271, 0.104, 0.083, and 0.255 for $q = 1, \cdots, 4$, respectively. The performances associated with vocabulary cognitive ability, the ability to diagnose grammar structure and table computing ability are markedly affected by the level-3 school climate covariates, whereas reading comprehension ability is not markedly affected when controlling for the level-2 *SES* and *GD* individual covariates and the level-3 (school-level) teacher satisfaction covariates. Analysis of the level-3 variance components reveals that the values of $\tau_{11q}(SES)$ are markedly different from 0, and their estimates are 0.275, 0.232, 0.261 and 0.289 for $q = 1, \cdots, 4$, respectively. This result illustrates that the effect of *SES* varies from school to school. In addition, the $\tau_{22q}(GD)$ values are also markedly different from 0. According to the information criteria and PsBF model selection results, Model 5⊕M2PLM shows the best fit with the real data when β_{1jq} and β_{2jq} are included as random effects. The estimation results show that the proportion of females to males varies among schools. None of the estimated covariances between the random effects τ_{01q}, τ_{02q}, and τ_{12q} are markedly different from 0. It can be concluded that the random effects are independent of each other for each type of ability.

7. CONCLUDING REMARKS

To explore the relations between multiple latent traits and covariates in a hierarchical data structure, this study presented a Bayesian MMIRT modeling and estimation procedure. An improved Gibbs sampling algorithm based on auxiliary variables for estimating MMIRT models is developed. The new algorithm overcomes the traditional Gibbs sampling algorithm's dependence on the conjugate prior for complex IRT model, and avoids some shortcomings of the Metropolis algorithm (such as sensitivity to step size, severe dependency on the candidate function or tuning parameter). Based on the simulation results, we see that the new algorithm provides accurate estimates of the structural parameters in the sense of having small Bias and RMSE values, and the coverage probability of the 95% highest posterior density interval is around 0.950 for each structural parameter. Therefore, the algorithm is effective and can be used to analyze the real data.

However, the computational burden of the new algorithm becomes intensive especially when a large number of examinees or the items is considered, or a large number of the MCMC sample size is used. Therefore, it is desirable to develop a standing-alone `R` package associated with `C++` or Fortran software for more extensive large-scale assessment program.

In addition, the new algorithm based on auxiliary variables can be extended to estimate some more complex item response and response time models, e.g., graded response model, Weibull response time model and so on.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (grant number 11571069).

Received 18 January 2017

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *2nd International Symposium on Information Theory.* Budapest: Akademiai Kiado. MR0483125

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

[3] Azevedo, C. L. N., Fox, J. P., and Andrade, D. F. (2016). Bayesian longitudinal item response modeling with restricted covariance pattern structures. *Statistics and Computing* **26**. 443–460. MR3439384

[4] Béguin, A. A. and Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika* **66**. 541–561. MR1961913

[5] Bishop, C. (2006). Slice sampling. *Pattern Recognition and Machine Learning.* Springer. MR2247587

[6] Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46**. 443–469. MR0668311

[7] Chalmers, R. P. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *Journal of Educational Measurement* **52**. 200–222.

[8] Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation.* New York: Springer. MR1742311

[9] Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by auxiliary variables. *Journal of the Royal Statistical Society, Series B* **61**. 331–344. MR1680334

[10] De la Torre, J. and 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**. 295–331.

[11] Embretson, S. E. and Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum.

[12] Fox, J. P. and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* **66**. 271–288. MR1836937

[13] Fu, Z. H., Tao, J., and Shi, N.-Z. (2009). Bayesian estimation in the multidimensional three-parameter logistic model. *Journal of Statistical Computation and Simulation* **46**. 669–690. MR2523023

[14] Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**. 501–514. MR1278223

[15] Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. E. M. Smith (Eds.), *Bayesian Statistics 4* (pp. 147–167). Oxford University Press. MR1380275

[16] Goldstein, H. (2003). *Multilevel Statistical Models,* 3rd edn, Edward Arnold, London.

[17] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**. 97–109. MR3363437

[18] Höhler, J., Harting, J., and Goldhammer, F. (2010). Modeling the multidimensional structure of students' foreign language competence within and between classrooms. *Psychological Test and Assessment Modeling* **52**. 323–340.

[19] Huang, H. Y., Wang, W. C., Chen, P. H., and Su, C. M. (2013). Higher-order item response theory models for hierarchical latent traits. *Applied Psychological Measurement* **37**. 619–637.

[20] Huang, H. Y. and Wang, W. C. (2014). Multilevel higher-order item response theory models. *Educational and Psychological Measurement* **73**. 495–515.

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

[22] Kass, R. E. and Wasserman, L. (2001). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90**. 928–934. MR1354008

[23] Kim, S. (2001). An evaluation of the Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement* **25**. 163–176. MR1824529

[24] Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

[25] Lu, J., Zhang, J., and Tao, J. (2018). Slice-Gibbs sampling algorithm for estimating the parameters of a multilevel item response model. *Journal of Mathematical Psychology* **82**. 12–25. MR3773680

[26] Lu, Y. (2012). *A multilevel multidimensional item response theory model to address the role of response style on measurement of attitudes in PISA 2006.* University of Wisconsin, Madison: Doctoral dissertation. MR3054976

[27] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**. 1087–1092.

[28] Muthén, B. and Asparouhov, T. (2013). *Item response modeling in Mplus: A multi-dimensional, multi-level, and multi-time point example.* Handbook of Item Response Theory: Models, Statistical Tools, and Applications.

[29] Neal, R. (2003). Slice sampling. *The Annals of Statistics* **31**. 705–767. MR1994729

[30] Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B* **56**. 3–48. MR1257793

[31] Padilla, J. L., Azevedo, C. L. N., and Lachos, V. H. (2017). Multidimensional multiple group IRT models with skew normal latent trait distributions. https://www.ime.unicamp.br/sites/default/files/pesquisa/relatorios/rp-2017-08.pdf.

[32] Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer Science Business Media, LLC.

[33] Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods,* second ed. Thousand Oaks, CA: Sage.

[34] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**. 461–464. MR0468014

[35] Shalabi, F. (2002). *Effective Schooling in the West Bank.* Doctoral dissertation, Twente University, Enschede, Netherlands.

[36] Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: effects of prior specifications on parameter estimates. *Behaviormetrika* **37**. 87–110.

[37] Sheng, Y. and Wikle, C. K. (2007). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement* **68**. 413–430. MR2432233

[38] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64**. 583–639. MR1979380

[39] Van der linden, W. J. and Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory.* New York: Springer-Verlag. MR1601043

[40] Yao, L. and Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement* **30**. 469–492. MR2252383

Jiwei Zhang
School of Mathematics and Statistics
Northeast Normal University
Changchun, 130024, Jilin
China
E-mail address: zhangjw713@nenu.edu.cn

Jing Lu
School of Mathematics and Statistics
Northeast Normal University
Changchun, 130024, Jilin
China
E-mail address: luj282@nenu.edu.cn

Jian Tao
School of Mathematics and Statistics
Northeast Normal University
Changchun, 130024, Jilin
China
E-mail address: taoj@nenu.edu.cn
url: http://js.nenu.edu.cn/teacher/index.php?zgh=1993900033