Bayesian longitudinal multilevel item response modeling approach for studying individual growth differences

Shuang Qu*, Jiwei Zhang and Jian Tao

Abstract

A longitudinal multilevel item response model is proposed for measuring changes in individual growth over time. To estimate the model parameters, a combined Bayesian procedure is developed. The deviance information criterion (DIC) and the widely applicable information criterion (WAIC) are used to assess the competing models. The simulation results show that the combined Bayesian estimation method performs perfectly in terms of recovering model parameters under various design conditions. Finally, a longitudinal dataset about the development of achievement in mathematics illustrates the significance and implementation of the proposed procedure.

Keywords: Item response theory, Longitudinal multilevel model, Markov chain Monte Carlo, Metropolis-Hastings within Gibbs algorithm.

*Corresponding author.

1 Introduction

Longitudinal studies have attracted interest in many fields, such as the health, social and behavioral sciences (Harville, 1977; Laird & Ware, 1982; Muthén, 2002; Bollen & Curran, 2006; Bacci, 2012). Specifically, in educational and psychological research, changes over time are often investigated through longitudinal analysis of observations collected at several time points. The purpose of such investigation is not only to study the achievement of individuals over time, but also to explore differences in individual growth trajectories among individuals of varying genders, family socioeconomic statuses, etc. There is a rich literature on the longitudinal studies in educational and psychological research, including Laird and Ware, 1982; Andersen, 1985; Embretson, 1991; Bollen and Curran, 2006; Muthén, 2002; Kim and Camilli, 2014; Andrade and Tavares, 2005; Azevedo et al., 2016; von Davier, Xu, and Carstensen, 2011.

Although the longitudinal studies in educational and psychological research have been deeply studied, there are still some deficiencies in the existing literature. Next, we compare the existing longitudinal models with our new model and analyze the advantages of our new model from multiple aspects. (1) A hierarchical modeling approach for measuring growth change provides a way to account efficiently for dependence resulting from the fact that the same individuals are assessed repeatedly, as in the case for random-effect and growth curve models (Laird & Ware, 1982; Bollen & Curran, 2006; Muthén, 2002). The two approaches dealing with latent traits are based on linear models for continuous responses that can be approximately normally distributed, where responses are typically obtained as simple or weighted sums across items through a particular assessment instrument. However, in many studies of educational psychology, responses are often discrete. Linear models are no longer appropriate for relating changes in mean responses to covariates. Instead, we construct a time-specific item response theory (IRT; Lord, 1980; van der Linden & Hambleton, 1997) model to describe the relathionship between individual and item at different time points through the binary responses. The time-specific IRT(TS-

IRT) model overcomes a number of potential problems that linear mixed models bring about by using a simple aggregate score for investigating change (such as paradoxical reliability of change scores, spurious negative correlations of change with initial status) (Kim & Camilli, 2014). Moreover, the TS-IRT model also solves inconsistent scale units for change encountered in linear mixed models, so that the latent traits of different time points are transformed into a single scale (Kim & Camilli, 2014). (2) Numerous studies on longitudinal IRT models have been conducted to measure individual growth. For example, Andersen (1985) proposed an extended Rasch model for the repeated administration of the same items over time points where item responses given on each occasion are modeled with a unidimensional IRT model and where the latent traits of each occasion are correlated. However, statistical inference results can present serious deviations due to strict assumptions of constant item difficulty parameters, and thus we cannot distinguish latent trait enhancement levels from later learning or the predisclosure of items (practical effects). Our TS-IRT model overcomes the deviations of statistical inference results caused by this strict assumptions, and evaluates the latent trait development by adopting the method that difficulty parameters are different at each time points and the different anchor items are employed to link multiple time points. (3) Andrade and Tavares (2005) extended Andersen's Rasch model to a three-parameter logistic model, from which they allowed latent traits for different occasions to follow a multivariate normal distribution so that serial correlations among latent traits are captured by a covariance matrix. Although the critical assumptions of strong factorial invariance over time can be satisfied by constraining all item parameters for known fixed values, the test cost will increase to precalibarate all of the test items at different time points. However, in our model, all items except anchor items do not need to be calibrated in advance as known values, and the unknown item parameters are estimated simultaneously by Bayesian sampling algorithm. Therefore, it avoids the huge expense in test items precalibration. (4) The model proposed by Azevedo et al. (2016) can be viewed as an extension of Andrade and Tavares (2005) where several restricted covariance pattern structures are considered to capture time-specific between-student variability and time heterogeneous longitudinal dependencies among latent traits. At the individual level, the time-specific latent traits are assumed to be multivariate normally distributed, and the within-individual correlation structure is modeled using a covariance pattern model. However, in our paper, each individual's time-specific latent traits is represented by an individual growth trajectory that is dependent on a unique set of parameters at the individual level rather than to assume to follow a multivariate normally distribution. In addition, the main purpose of our paper is to explore differences in individual growth trajectories between individuals of varying genders and family socioeconomic statuses rather than to analyze the correlation between the latent traits of multiple dimensions. (5) To relax the assumption of setting item difficulty parameters as constants, Embretson (1991) developed a multidimensional Rasch model for learning and change (MRMLC) to provide parameters for individual differences in change where the model assumed that on the first occasion ($t = 1$), only an initial latent trait is involved in item responses while for later occasions, latent traits θ_t ($t > 1$) quantified by $t - 1$ additional latent traits are involved in performance. Thus, the increment of the latent trait between successive occasions can be quantified directly. Embretson (1991) described the growth of the individual's latent trait through the increment of latent trait, which was obviously quite different from that by the growth curve as shown in our study. (6) von Davier, Xu, and Carstensen (2011) developed a mixture longitudinal multidimensional IRT model to explore whether multidimensional academic growth is homogeneous across different types of schools. However, the abovementioned models only consider latent traits as special values to compare them with other latent traits for different time points. In this paper, we are more concerned with the nature of latent trait growth trajectory (linear or quadratic growth) and with whether growth patterns are identical for different individual background variables (e.g., genders and socioeconomic statuses).

In this paper, we propose a longitudinal multilevel TS-IRT(LMTS-IRT) model that measures changes in individual growth over time. We use a combined Bayesian algorithm that combines the Metropolis-within-Gibbs algorithm (Metropolis et al., 1953; Hastings, 1970; Tierney, 1994) with the Gibbs algorithm (Geman & Geman, 1984; Gelfand & Smith, 1990) to simultaneously estimate parameters, and a combined Bayesian procedure is developed. Specifically, the Metropolis-within-Gibbs algorithm is used to estimate parameters without conjugate priors so that the full conditional distributions are not available (Hahn, 2014) while the Gibbs algorithm is used to estimate other parameters with conjugate priors. Additionally, the DIC and WAIC were used to assess model fit in the simulation study. Finally, a longitudinal dataset about the development of achievement in mathematics illustrates the significance and implementation of the proposed procedure.

The remainder of this paper is organized as follows. In Section 2, the LMTS-IRT model and its identifiability are described. This is followed by a description of our combined Bayesian sampling procedure and a discussion of model selection criteria in Section 3. In Section 4, simulation studies are conducted to evaluate the performance of our Bayesian sampling algorithm and of the model assessment method. In addition, an analysis of the longitudinal education quality assessment data is given in Section 5. Finally, some concluding remarks are presented.

2 Model and Its Identification

A longitudinal multilevel item response model is proposed that consists of three levels. At level 1, a TS-IRT model is considered for the measurement of the time-specific latent traits. At level 2, within-individual dependence is described by a polynomial growth trajectory model. That is, latent trait parameters are predicted from an individual growth curve, which is a polynomial of degree H ($H = 1$, linear growth model; $H = 2$, quadratic growth model). At level 3, between-individual dependence is explained based on individual's background covariates under the framework of the multilevel model.

2.1 TS-IRT model (Level 1)

Assume that there are K items and T measurement occasions for a longitudinal assessment. For level 1, the correct response probability is expressed as

$$p_{tik} = P\left(Y_{tik} = 1 \left| \theta_{ti}, \boldsymbol{\xi}_{tk} \right.\right) = \frac{\exp\left(a_{tk}\theta_{ti} - b_{tk}\right)}{1 + \exp\left(a_{tk}\theta_{ti} - b_{tk}\right)}. \tag{1}$$

In Equation (1), Y_{tik} denotes the response of the ith examinee at the tth measurement occasion on the kth item, and the correct response probability is expressed p_{tik}; θ_{ti} is the latent trait of examinee i $(i = 1, ..., n)$ at measurement occasion t $(t = 1, ..., T)$; and $\boldsymbol{\xi}_{tk} = \left(a_{tk}, b_{tk}\right)'$ denotes the vector of item parameters, whereby a_{tk} and b_{tk} $(k = 1, ..., K)$ are respectively the discrimination (slope) parameter and difficulty (intercept) parameter for the kth item at the tth measurement occasion.

2.2 Longitudinal individual growth model (Level 2)

Many phenomena related to individual ability changes can be represented through a two-level model. At level 2, each individual's latent trait development is represented by an individual growth trajectory, that is dependent on a unique set of parameters. These individual growth parameters become outcome variables in the level-3 model, wherein they can depend on individual background characteristics (Raudenbush & Bryk, 2002). Measurements made at different time points are regarded as "nested" within individuals. Therefore, the individual growth trajectory model can be described as follows:

$$\theta_{ti} = \pi_{0i} + \pi_{1i}d_{ti} + \pi_{2i}d_{ti}^2 + ... + \pi_{Hi}d_{ti}^H + e_{ti}. \tag{2}$$

In Equation (2), the latent trait growth level over time is represented as a polynomial of degree H. The variable d_{ti} is the test time parameter at occasion t for examinee i, and πs denote coefficients of the polynomial function. Random error terms, $e_{ti}s$, are assumed to follow a common normal distribution with mean 0 and variance σ^2. Note that Bryk and Raudenbush (1992) argued that it is defensible to assume a simple error variance structure (the error terms are equal and uncorrelated between the time points), wherein there are a limited number of time points. In such cases with short time series, this assumption is very practical and analysis results are robust.

2.3 Multilevel model (Level 3)

Assume that the growth parameters vary across individuals, thus individual growth trajectory parameters can be represented by person-level background covariates such as an individual's socioeconomic status (SES) and gender. We formulate the person-level model to explain this variation as follows:

$$\pi_{hi} = \beta_{h0} + \beta_{h1}x_{1i} + \beta_{h2}x_{2i} + ... + \beta_{hS}x_{Si} + u_{hi}. \tag{3}$$

In Equation (3), x_{si} is the sth$(s = 1, ..., S)$ person-level background covariate for examinee i, and β_{hs} is the effect of x_{si} on the hth growth parameter. u_{hi} $(h = 0, ..., H)$ is the level-3 random residual effect for examinee i, and the vector $\boldsymbol{u} = (u_{0i}, u_{1i}, u_{2i}, ..., u_{Hi})$ is assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}_{(H+1)\times(H+1)}$.

2.4 Model identification

To ensure the identification of the single-level two-parameter IRT model, either the scale of latent traits or the scale of item parameters have to be restricted (van der Linden & Hambleton, 1997; Lord, 1980). One can set the mean and variance of the latent traits to zero and one, respectively (Bock & Aitkin, 1981). Alternatively, one way to restrict the scale of item parameters is to impose constraints of $\prod_k a_k = 1$ and $\sum_k b_k = 0$ on model item parameters; the equivalent form anchors one discrimination parameter to 1 and one difficulty parameter to 0 (Fox & Glas, 2001). On the other hand, as there is overlap between items anchored at different times (i.e., anchor items) in longitudinal analysis, in this article we restrict the anchor item parameters at different time points as known and pre-linked to identify the LMTS-IRT model(Wang et al., 2016).

3 Bayesian Estimation and Model Selection

3.1 Bayesian estimation

A combined Bayesian algorithm is used to estimate parameters of interest. Let $\mathbf{\Psi} = (\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\pi}, \sigma^2, \boldsymbol{\beta}, \mathbf{\Omega})$ represent the set of all item parameters at different time points. Let denote the time-based loading matrix. The joint posterior distribution of the parameters given the data can be written as follows:

$$p\left(\mathbf{\Psi}\left|\boldsymbol{Y}, D, \boldsymbol{X}\right.\right) \propto \prod_{t=1}^{T}\prod_{i=1}^{n}\prod_{k=1}^{K} p\left(Y_{tik}\left|\theta_{ti}, \boldsymbol{\xi}_{tk}\right.\right) p\left(\theta_{ti}\left|\boldsymbol{\pi}_i, \sigma^2, d_{ti}\right.\right) p\left(\boldsymbol{\pi}_i\left|\boldsymbol{\beta}, \mathbf{\Omega}, \boldsymbol{X}_i\right.\right) \\ \times p\left(\boldsymbol{\beta}\right) p\left(\boldsymbol{\xi}_{tk}\right) p\left(\sigma^2\right) p\left(\mathbf{\Omega}\right). \tag{4}$$

Our combined algorithm requires sampling from the following posterior distributions in turn:

- **Step 1**: Sample the ability parameter θ_{ti} for the ith individual for the measurement occasion t from the full conditional distribution $[\theta_{ti}\left|\boldsymbol{a}_{t\cdot}, \boldsymbol{b}_{t\cdot}, d_{ti}, \boldsymbol{\pi}_i, \sigma^2, \boldsymbol{Y}_{ti}\right.]$. Here, $\boldsymbol{a}_{t\cdot} = (a_{t1}, a_{t2}, \cdots, a_{tK})$, $\boldsymbol{b}_{t\cdot} = (b_{t1}, b_{t2}, \cdots, b_{tK})$ and $\boldsymbol{Y}_{ti} = (Y_{ti1}, Y_{ti2}, \cdots, Y_{tiK})$.

- **Step 2**: Sample the difficulty parameter b_{tk} for the measurement occasion t from the full conditional distribution $[b_{tk}\left|a_{tk}, \boldsymbol{\theta}_{t\cdot}, \boldsymbol{Y}_{tk}\right.]$. Here, $\boldsymbol{\theta}_{t\cdot} = (\theta_{t1}, \theta_{t2}, \cdots, \theta_{tn})$ and $\boldsymbol{Y}_{tk} = (Y_{t1k}, Y_{t2k}, \cdots, Y_{tnk})$.

- **Step 3**: Sample the discrimination parameter a_{tk} for the measurement occasion t from the full conditional distribution $[a_{tk}\,|b_{tk}, \boldsymbol{\theta}_{t\cdot}, \boldsymbol{Y}_{tk}]$.

- **Step 4**: Sample the level-2 random coefficients $\boldsymbol{\pi}_i$ from $[\boldsymbol{\pi}_i\,|\boldsymbol{\theta}_i, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\Omega}]$. Here, $\boldsymbol{\theta}_i \triangleq \boldsymbol{\theta}_{\cdot i} = (\theta_{1i}, \theta_{2i}, \cdots, \theta_{Ti})^{'}$.

- **Step 5**: Sample the level-3 regression coefficients $\boldsymbol{\beta}$ from $[\boldsymbol{\beta}\,|\boldsymbol{\pi}, \boldsymbol{\Omega}]$.

- **Step 6**: Sample the level-2 residual variance σ^2 from $[\sigma^2\,|\boldsymbol{\theta}, \boldsymbol{\pi}, v, \omega]$. Here, the prior for σ^2 is an inverse-Gamma(v, ω) distribution.

- **Step 7**: Sample the level-3 covariance matrix $\boldsymbol{\Omega}$ from $[\boldsymbol{\Omega}\,|\boldsymbol{\pi}, \boldsymbol{\beta}, \lambda, \Xi]$. Here, the prior for is an inverse-Wishart(λ, Ξ) distribution.

For Steps 1 to 3, the Metropolis-Hastings Gibbs algorithm is used to draw samples from the full conditional posterior distributions because the parameters of interest do not have closed form of the corresponding posterior distribution. Note that since the discrimination parameters should be positive, we use the log-normal distribution as the proposal distribution to ensure that the candidate samples are greater than zero. The proposal distribution of discrimination parameters is assumed as a log-normal distribution with mean equal to the current estimation and variance chosen to give an acceptance rate of 25 to 40 percent. For Steps 4 to 7, it is easy and efficient to use the Gibbs algorithm through the use of conjugate priors. Further detailed information on the combined Bayesian algorithm is provided in the Appendix and the corresponding MATLAB program is available upon request.

3.2 Model selection

It is well known that two widely used model selection criteria are the Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criterion (BIC) (Schwartz, 1978), which depend on the effective number of parameters in a model as a measure of model complexity. However, as a drawback of these measures, they are often difficult to calculate for random-effect models, as the effective number of parameters is heavily dependent on higher-level variance parameters. When the variance in random effects approaches zero, all random effects are equal and the model reduces to a simple linear model with one mean parameter. However, when the variance goes to infinity, the number of free parameters approaches the number of random effects. To overcome the above problems, Spiegelhalter *et al.* (2002) proposed the deviance information criterion (DIC) for conducting model comparisons when the number of parameters is not clearly defined in a random-effect model. The DIC is calculated as a sum of deviance measure and penalty term for the effective number of parameters based on a measure of model complexity. In the Bayesian IRT literature, DIC is one of the most popular model comparion methods and widely used for multilevel models. The penalty term has the following form:

$$\begin{aligned} p_D &= E\left(-2\log p\left(Y\left|\boldsymbol{\theta},\boldsymbol{a},\boldsymbol{b}\right.\right)\right) + 2\log p\left(Y\left|\overline{\boldsymbol{\theta}},\overline{\boldsymbol{a}},\overline{\boldsymbol{b}}\right.\right) \\ &= \overline{D\left(\boldsymbol{\Psi}\right)} - D\left(\overline{\boldsymbol{\Psi}}\right). \end{aligned} \tag{5}$$

The deviance function is given by $D\left(\boldsymbol{\Psi}\right) = -2\log\left[\prod_{t=1}^{T}\prod_{i=1}^{n}\prod_{k=1}^{K} p\left(Y_{tik}\left|\theta_{ti},a_{tk},b_{tk}\right.\right)\right]$. $\overline{D\left(\boldsymbol{\Psi}\right)} = (-2)\frac{1}{M}\sum_{m=1}^{M}\log\left[\prod_{t=1}^{T}\prod_{i=1}^{n}\prod_{k=1}^{K} p\left(Y_{tik}\left|\theta_{ti}^{(m)},a_{tk}^{(m)},b_{tk}^{(m)}\right.\right)\right]$ is the posterior mean deviance and $D\left(\overline{\boldsymbol{\Psi}}\right)$ is the estimated deviance for the posterior estimate of $\overline{\boldsymbol{\Psi}}$. Only the computation of the first term of its penalty term utilizes the whole posterior distribution. Then the DIC is given as

$$\text{DIC} = \overline{D\left(\boldsymbol{\Psi}\right)} + p_D = \overline{D\left(\boldsymbol{\Psi}\right)} - D\left(\overline{\boldsymbol{\Psi}}\right). \tag{6}$$

Within the competing models, those with lower DIC values are preferred over those with higher DIC values.

Additionally, a more fully Bayesian approach is also used to the model assessment. That is the widely applicable information criterion (WAIC; Watanabe, 2010, 2013;

Gelman, Hwang, & Vehtari, 2014). The penalty term has the following form:

$$
\begin{aligned}
p_{WAIC} &= \sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=1}^{K} var_{post}\left[\log p\left(Y_{tik}\left|\theta_{ti}, a_{tk}, b_{tk}\right.\right)\right] \\
&= \sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{\frac{1}{M-1}\sum_{m=1}^{M}\left[\log p\left(Y_{tik}\left|\theta_{ti}^{(m)}, a_{tk}^{(m)}, b_{tk}^{(m)}\right.\right)\right.\right. \\
&\left.\left. -\frac{1}{M}\sum_{m=1}^{M}\log p\left(Y_{tik}\left|\theta_{ti}^{(m)}, a_{tk}^{(m)}, b_{tk}^{(m)}\right.\right)\right]^{2}\right\}
\end{aligned}
\tag{7}
$$

Let

$$
\begin{aligned}
\widehat{\text{lppd}} &= \text{the estimate of the log pointwise predictive density} \\
&= \sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=1}^{K}\log\left[\frac{1}{M}\sum_{m=1}^{M} p\left(Y_{tik}\left|\theta_{ti}^{(m)}, a_{tk}^{(m)}, b_{tk}^{(m)}\right.\right)\right].
\end{aligned}
\tag{8}
$$

Therefore, the WAIC can be written as

$$
\text{WAIC} = -2\left(\widehat{\text{lppd}} - p_{WAIC}\right). \tag{9}
$$

The model with a smaller WAIC has a better fit to the data. As can be seen from equation (7), the computation of the penalty term utilizes the whole posterior distribution other than point estimates which is why WAIC is considered full Bayesian. The theoretical superiority is acknowledged(Vehtari, Gelman, & Gabry, 2017; Luo & Al-Harbi, 2017), how such a strength translate into our simulation remain unknow.

4 Simulation Study

4.1 Simulation study 1

Simulation design

The simulation study was conducted to evaluate the recovery performance of the combined Markov chain Monte Carlo (MCMC) sampling algorithm. Three time points were considered (i.e., $t = 1, 2, 3$). When estimating model parameters, 20% items per occasion were treated as anchor items, which were assumed to be known and pre-linked. The following manipulated conditions were considered: (a) test length per occasion, $K = 20$ or 30 (i.e., there were 4 or 6 anchor items at each measurement occasion); and (b) the number of individuals, $N = 500, 1,000$ or $2,000$. Fully crossing different levels of these two factors yielded 6 conditions (2 test lengths $\times$ 3 sample sizes). Response data were simulated using the level-1 TS-IRT model given by Equation (1). For illustrative purpose, we used the quadratic growth model to describe the level-2 individual development trajectory, and the level-3 model that included two explanatory variables was considered. The structural model can be written as

$$\begin{cases} \theta_{ti} = \pi_{0i} + \pi_{1i} d_{ti} + \pi_{2i} d_{ti}^2 + e_{ti}, \\ \pi_{0i} = \beta_{00} + \beta_{01} x_{1i} + \beta_{02} x_{2i} + u_{0i}, \\ \pi_{1i} = \beta_{10} + \beta_{11} x_{1i} + \beta_{12} x_{2i} + u_{1i}, \\ \pi_{2i} = \beta_{20} + \beta_{21} x_{1i} + \beta_{22} x_{2i} + u_{2i}. \end{cases} \tag{10}$$

In Equation (10), $e_{ti} \sim N(0, \sigma^2)$, $t = 1, 2, 3$;

$$\begin{pmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{\Omega} \right), \text{ where } \mathbf{\Omega} = \begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{10} & \tau_{11} & \tau_{12} \\ \tau_{20} & \tau_{21} & \tau_{22} \end{pmatrix};$$

and d_{ti} were the time-specific covariates. True item discrimination parameters a_{tk} for different time points were generated from $\log(N(\exp(1),\ 0.15))$, $t = 1, 2, 3$. The item difficulty parameters b_{ik} were respectively generated from three normal distributions, i.e., $b_{1k} \sim N(0, 0.05)$, $b_{2k} \sim N(0.25, 0.05)$, and $b_{3k} \sim N(0.5, 0.05)$. The ability parameters of individuals $\boldsymbol{\theta}_i$ were generated from the normal distribution $N(\boldsymbol{A}_i \boldsymbol{\pi}_i,\ \sigma^2 \mathbf{I}_{T\times T})$, where the true value of the level-2 residual variance was set to 0.15 (i.e., $\sigma^2 = 0.15$) and $\boldsymbol{D_i}$ was a time-based loading matrix for examinee i (for further details, please see step 4 in the appendix), and where the level-2 random regression coefficients $\boldsymbol{\pi}_i$ were induced by a normal distribution with mean vector $\boldsymbol{X}_i\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Omega}$. Therefore, to generate $\boldsymbol{\pi}_i$, we only need to know the true values of the fixed effect $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Omega}$ where $\boldsymbol{\beta} = (0\ 0.15\ 0.05; 0.35\ -0.05\ 0.5; 0.3\ \ -0.225\ 0.15)$ and $\boldsymbol{\Omega} = (0.1\ 0.05\ 0.025; 0.05\ 0.1\ 0.005; 0.025\ 0.005\ 0.1)$. Explanatory variables $\boldsymbol{X}$ were drawn from $N(0.5, 1)$.

Prior distributions

We assume that priors of the discrimination and difficulty parameters were taken to be $a_{tk} \sim \log N(0, 0.5)$ and $b_{tk} \sim N(0, 2)$ from Patz & Junker(1999a, 1999b). The fixed effect $\boldsymbol{\beta}$ followed the normal prior distribution $N(0, 100)$. The prior to the variance of the level-2 residual was assumed to follow an inverse gamma distribution with shape parameter $v = 0.001$ and rate parameter $\omega = 0.001$. The prior to the level-3 covariance matrix $\boldsymbol{\Omega}$ was set to be an inverse Wishart distribution with small degrees of freedom $\lambda = 4$ and identity matrix Ξ.

Convergence diagnostics

As an illustation, convergence diagnostics consider a situation in which the test length was 60 for three time points, and the individual sample size was set to 1,000. The following two methods were used to check the convergence of our algorithm: the Gelman-Rubin method (Gelman, 1996; Gelman & Rubin, 1992) and the Raftery-Lewis diagnostic method (Raftery & Lewis, 1996). The convergence of the MCMC sampler was checked by monitoring 5 chain trace plots of parameters for consecutive sequences of 10,000 iterations. The first 2500 iterations were discarded as burn-in period.

Figures 1 and 2 represented trace and autocorrelation plots for the fixed-effect parameter vector $\boldsymbol{\beta}$, level-2 variance parameter σ^2, and level-3 variance-covariance parameter $\boldsymbol{\Omega}$, respectively. The Brooks-Gelman ratio diagnostic $\widehat{R}$ (as an updated Gelman-Rubin statistic) plots were also used to monitor the convergence and stability (Gelman, 1996;

Brooks & Gelman, 1998). From Figure 3, it can be seen that nine plots of $\widehat{R}$ were all close to 1 rapidly and finally less than 1.2, which supported the convergence of the MCMC sampler (Lunn et al., 2000).

Parameter recovery

The accuracy of the parameter estimates was measured by five evaluation criteria, i.e., Bias, Root Mean Squared Error (RMSE), Standard deviation (SD), Standard error (SE) and coverage probability (CP) of the 95% highest posterior density intervals (HPDI) statistics. Let η be the parameter of interest. Assume that $M = 500$ data sets were generated. Also, let $\widehat{\eta}^{(m)}$ and $\mathrm{SD}^{(m)}(\eta)$ denoted the posterior mean and the posterior standard deviation of η obtained from the mth simulated data set for $m = 1, \ldots, M$.

The Bias for parameter η is defined as

$$\mathrm{Bias}(\eta) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\eta}^{(m)} - \eta\right), \tag{11}$$

and the RMSE for parameter η is defined as

$$\mathrm{RMSE}(\eta) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\eta}^{(m)} - \eta\right)^2}. \tag{12}$$

The simulation SE is the square root of the sample variance of the posterior estimates over different simulated data sets. It can be defined as

$$\text{Simulation SE}(\eta) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\eta}^{(m)} - \frac{1}{M}\sum_{\ell=1}^{M}\widehat{\eta}^{(\ell)}\right)^2}. \tag{13}$$

and the average of posterior standard deviation can be defined as

$$\mathrm{SD}(\eta) = \frac{1}{M}\sum_{m=1}^{M}\mathrm{SD}^{(m)}(\eta). \tag{14}$$

The coverage probability can be defined as

$$\text{CP}\,(\eta) = \frac{\#\text{ of }95\%\text{ HPDI containing }\eta\text{ in }M\text{ simulated data sets}}{M}. \tag{15}$$

Results

The average Bias, RMSE, SD, SE and CP for discrimination and difficulty parameters at each time point were shown in Tables 1. From Table 1, the following conclusions can be obtained. (1) Given the total test length, when the number of individuals increased from 500 to 2000, the average Bias, RMSE, SD and SE for discrimination and difficulty parameters obviously decreased. For example, the total test length was 60 items and the three time points were considered, when the number of individuals increased from 500 to 2000, the average Bias of all discrimination parameters decreased from 0.018 to 0.004, the average RMSE of all discrimination parameters decreased from 0.013 to 0.067, the average SD of all discrimination parameters decreased from 0.156 to 0.076, and the average of SE of all discrimination parameters decreased from 0.158 to 0.093. (2) The average SD were slightly less than the average SE, but they were very close. This indicated that the fluctuation of posterior mean between different replications was large compared with the fluctuation of posterior mean in each replication. (3) At different time points, the average CP of the discrimination and difficulty parameters were about 0.95. (4) When the total test length increased from 60 to 90, the average Bias, RMSE, SD and SE shown that the recovery results of the discrimination and difficulty parameters were close to the case that total test length was 60, which indicated that our algorithm was stable and did not reduce the accuracy due to the increase in the number of items.

The recovery performance of structure parameters for six kinds of simulation design was shown in Table 2. From Table 2, it can be found that the Bias of the fixed effect parameters (βs) had a range of $-0.011\sim0.006$ under all six conditions. The Bias had a range of $-0.021\sim -0.016$ for the level-2 variance parameter (σ^2), and $-0.039\sim0.094$ for the level-3 covariance parameters (τ). The RMSE had a range of 0.009∼0.100 for the fixed effect parameters , 0.021∼0.024 for the level-2 variance parameter, and 0.008∼0.101 for the level-3 covariance parameters. Additionally, the SD of the fixed effect parameters had a range of 0.010∼0.101. The SD had a range of 0.014∼0.021 for the level-2 variance parameter, 0.007∼0.067 for the level-3 covariance parameters. The SE had a range of 0.009∼0.100 for the fixed effect parameters, 0.007∼0.013 for the level-2 variance, and 0.007∼0.036 for the level-3 covariance parameters. Moreover, the CP of the fixed effect parameters had a range of 0.914∼0.966 under six different design conditions. The CP had a range of 0.784∼0.926 for the level-2 variance parameter. The CP had a range of 0.802∼0.958 for the level-3 covariance parameters. In summary, it is obvious that the Bayesian sampling algorithm provided accurate estimates of the item and structure parameters in term of five indexes evaluation results.

4.2 Simulation study 2

The purpose of this simulation was to show our Bayesian sampling algorithm was effective to recover various prior distributions of the item parameters, where the sensitivity analysis based on item parameter prior distribution with a larger variance was addressed.

Simulation Design

As an illustration, the number of individuals was fixed on 1000. Three time points were considered and test length per occasion was $K = 20$ (i.e., there were 4 anchor items at each measurement occasion). Response data were generated from the level-1 time-specific IRT model given by Equation (1). The growth model and the level-3 model were same as the simulation study 1. The true values of parameters were also same as the simulation study 1. Next, the four types of priors were given by the following: (i) $a_j \sim \log N(0, 0.5)$ and $b_j \sim N(0, 0.5)$; (ii) $a_j \sim \log N(0, 1)$ and $b_j \sim N(0, 1)$; (iii) $a_j \sim \log N(0, 10)$ and $b_j \sim N(0, 10)$; (iv) $a_j \sim \log N(0, 100)$ and $b_j \sim N(0, 100)$.

Table 1: Evaluating the accuracy of the item parameters in simulation study 1.

	No. of items=60														
	No. of individuals 500					No. of individuals 1000					No. of individuals 2000				
Item parameter	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
Discrimination $\boldsymbol{a}_{1\cdot}$	0.023	0.153	0.179	0.177	0.955	0.011	0.111	0.125	0.133	0.951	0.006	0.079	0.088	0.102	0.951
Discrimination $\boldsymbol{a}_{2\cdot}$	0.043	0.212	0.258	0.237	0.951	0.022	0.148	0.174	0.173	0.954	0.011	0.108	0.126	0.130	0.944
Discrimination $\boldsymbol{a}_{3\cdot}$	0.018	0.134	0.156	0.158	0.955	0.008	0.095	0.108	0.118	0.950	0.004	0.067	0.076	0.093	0.951
Difficulty $\boldsymbol{b}_{1\cdot}$	-0.006	0.133	0.152	0.157	0.949	-0.007	0.095	0.108	0.115	0.950	-0.003	0.068	0.076	0.091	0.948
Difficulty $\boldsymbol{b}_{2\cdot}$	0.000	0.095	0.107	0.117	0.953	-0.000	0.067	0.074	0.090	0.949	0.000	0.046	0.052	0.073	0.947
Difficulty $\boldsymbol{b}_{3\cdot}$	0.001	0.104	0.116	0.127	0.949	0.001	0.073	0.082	0.094	0.949	0.000	0.051	0.057	0.075	0.949
	No. of items=90														
	No. of individuals 500					No. of individuals 1000					No. of individuals 2000				
Item parameter	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
Discrimination $\boldsymbol{a}_{1\cdot}$	0.022	0.150	0.173	0.174	0.951	0.011	0.108	0.121	0.131	0.947	0.005	0.076	0.085	0.101	0.948
Discrimination $\boldsymbol{a}_{2\cdot}$	0.040	0.205	0.245	0.229	0.947	0.019	0.140	0.165	0.163	0.955	0.010	0.101	0.115	0.123	0.951
Discrimination $\boldsymbol{a}_{3\cdot}$	0.016	0.131	0.151	0.154	0.949	0.008	0.092	0.104	0.115	0.949	0.004	0.065	0.073	0.090	0.949
Difficulty $\boldsymbol{b}_{1\cdot}$	-0.007	0.127	0.147	0.151	0.957	-0.005	0.094	0.104	0.116	0.946	-0.002	0.066	0.073	0.088	0.947
Difficulty $\boldsymbol{b}_{2\cdot}$	-0.002	0.092	0.103	0.115	0.947	-0.000	0.064	0.071	0.088	0.951	-0.001	0.045	0.050	0.071	0.947
Difficulty $\boldsymbol{b}_{3\cdot}$	0.001	0.099	0.112	0.121	0.954	0.001	0.070	0.079	0.093	0.951	0.001	0.049	0.055	0.074	0.949

Note that the Bias, RMSE, SD, SE and CP denote the average Bias, RMSE, SD, SE and CP for all discrimination and difficulty parameters at each time point.

Table 2: Evaluating the accuracy of the fixed and random effect parameters in simulation study 1.

	No. of items=60														
	No. of individuals 500					No. of individuals 1000					No. of individuals 2000				
Fixed effect	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
β_{00}	-0.003	0.053	0.053	0.053	0.946	-0.002	0.040	0.037	0.040	0.914	0.000	0.027	0.026	0.027	0.946
β_{01}	0.005	0.022	0.022	0.021	0.952	0.002	0.016	0.015	0.016	0.958	0.001	0.011	0.011	0.011	0.952
β_{02}	-0.001	0.022	0.022	0.022	0.956	-0.001	0.015	0.015	0.014	0.958	-0.000	0.011	0.011	0.011	0.960
β_{10}	-0.000	0.052	0.051	0.052	0.932	0.002	0.037	0.036	0.037	0.966	0.002	0.026	0.025	0.026	0.940
β_{11}	0.001	0.027	0.028	0.027	0.948	-0.001	0.018	0.020	0.018	0.964	0.000	0.014	0.014	0.014	0.932
β_{12}	-0.011	0.032	0.029	0.030	0.928	-0.003	0.019	0.021	0.019	0.958	-0.002	0.015	0.014	0.014	0.954
β_{20}	0.002	0.100	0.101	0.100	0.928	-0.002	0.075	0.071	0.075	0.928	-0.003	0.051	0.050	0.051	0.940
β_{21}	0.006	0.042	0.040	0.042	0.922	0.003	0.029	0.028	0.029	0.938	0.001	0.020	0.020	0.020	0.952
β_{22}	-0.002	0.039	0.041	0.039	0.950	-0.001	0.030	0.028	0.031	0.938	0.000	0.018	0.020	0.018	0.966
Random effect	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
σ^2 (level-2 var.)	-0.016	0.021	0.021	0.013	0.926	-0.019	0.022	0.017	0.011	0.878	-0.019	0.021	0.015	0.009	0.822
τ_{00}	0.020	0.027	0.024	0.018	0.956	0.019	0.024	0.019	0.014	0.936	0.018	0.021	0.016	0.010	0.900
τ_{10}	-0.012	0.018	0.016	0.013	0.902	-0.007	0.013	0.012	0.011	0.924	-0.005	0.010	0.009	0.008	0.926
τ_{11}	0.028	0.035	0.026	0.021	0.896	0.022	0.026	0.020	0.015	0.898	0.018	0.022	0.016	0.012	0.852
τ_{20}	-0.039	0.044	0.031	0.019	0.868	-0.035	0.038	0.026	0.015	0.844	-0.031	0.033	0.022	0.011	0.838
τ_{21}	0.010	0.025	0.028	0.023	0.978	0.008	0.020	0.021	0.019	0.946	0.005	0.015	0.015	0.014	0.952
τ_{22}	0.094	0.101	0.067	0.036	0.894	0.085	0.091	0.058	0.031	0.804	0.076	0.080	0.051	0.026	0.802
	No. of items=90														
	No. of individuals 500					No. of individuals 1000					No. of individuals 2000				
Fixed effect	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
β_{00}	-0.004	0.044	0.045	0.044	0.960	-0.003	0.034	0.032	0.034	0.930	-0.002	0.023	0.023	0.023	0.934
β_{01}	0.005	0.022	0.021	0.021	0.946	0.002	0.015	0.015	0.015	0.932	0.001	0.010	0.010	0.011	0.958
β_{02}	-0.001	0.021	0.021	0.021	0.954	-0.001	0.015	0.015	0.015	0.944	-0.001	0.009	0.010	0.009	0.956
β_{10}	-0.001	0.043	0.044	0.043	0.960	0.000	0.032	0.031	0.032	0.948	0.001	0.021	0.022	0.021	0.946
β_{11}	0.002	0.025	0.026	0.025	0.956	0.001	0.018	0.018	0.018	0.948	0.000	0.013	0.012	0.013	0.940
β_{12}	-0.009	0.028	0.026	0.027	0.936	-0.004	0.019	0.018	0.019	0.938	-0.002	0.013	0.013	0.013	0.960
β_{20}	0.001	0.089	0.085	0.089	0.938	0.001	0.060	0.060	0.060	0.952	0.001	0.043	0.042	0.043	0.952
β_{21}	0.006	0.038	0.037	0.037	0.948	0.002	0.027	0.026	0.027	0.954	0.001	0.018	0.018	0.018	0.936
β_{22}	-0.001	0.039	0.037	0.039	0.954	-0.001	0.026	0.026	0.026	0.946	0.000	0.019	0.018	0.019	0.948
Random effect	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
σ^2 (level-2 var.)	-0.021	0.024	0.018	0.011	0.836	-0.021	0.023	0.016	0.009	0.790	-0.020	0.021	0.014	0.007	0.784
τ_{00}	0.016	0.023	0.021	0.015	0.958	0.018	0.021	0.018	0.012	0.928	0.017	0.019	0.015	0.009	0.912
τ_{10}	-0.010	0.016	0.014	0.012	0.902	-0.006	0.011	0.010	0.009	0.922	-0.004	0.008	0.007	0.007	0.916
τ_{11}	0.023	0.029	0.022	0.017	0.928	0.019	0.024	0.018	0.014	0.876	0.017	0.020	0.014	0.010	0.836
τ_{20}	-0.034	0.038	0.027	0.016	0.896	-0.033	0.035	0.023	0.013	0.860	-0.030	0.032	0.021	0.009	0.832
τ_{21}	0.008	0.020	0.024	0.018	0.992	0.005	0.017	0.018	0.016	0.954	0.004	0.013	0.013	0.012	0.950
τ_{22}	0.089	0.095	0.061	0.033	0.854	0.080	0.085	0.054	0.029	0.824	0.073	0.076	0.049	0.021	0.802

Results

The Bayesian sampling algorithm was iterated 10,000 times. The first 2,500 iterations were discarded as burn-in period. 500 replications were considered in this simulation. The recovery performance of item parameters for four kinds of simulation design was shown in Table 3. It can be found that the average Bias of the discrimination parameters had a range of 0.008∼0.029 under four conditions (−0.031∼0.004 for difficulty parameters). Additionally, the average RMSE of the discrimination parameters had a range of 0.095∼0.174 under four conditions (0.067∼0.099 for difficulty parameters). The average SD and SE of the discrimination parameters had the range of 0.108∼0.191 and 0.118∼0.201 under four conditions, and the average SD and SE of the difficulty parameters had the range of 0.073∼0.110 and 0.087∼0.121 under four conditions. The average SD were slightly less than the average SE, but they were very close. Moreover, we found that when the prior variances of the discrimination and difficulty parameters increased from 0.5 to 10, the average RMSE of the discrimination and difficulty parameters increased slightly, which indicated that there was almost no change in the estimation accuracy when the prior changed from informative prior to non-informative prior (variance increased from 0.5 to 10). When the prior variances of the discrimination and difficulty parameters increased from 10 to 100, the average Bias, RMSE, SD, SE and CP of the discrimination and difficulty parameters were almost the same for both cases. This indicated that when the prior variance researchs 10, the prior was "flat" enough to provide relatively little information.

The recovery performance of structure parameters for four kinds of prior design was shown Table 4. From Table 4, it can be found that the Bias of the fixed effect parameters had a range of −0.018∼0.005 under all four conditions. The Bias had a range of −0.021∼ −0.019 for the level-2 variance parameter, and −0.034∼0.086 for the level-3 covariance parameters . The RMSE had a range of 0.014∼0.074 for the fixed effect parameters , 0.022∼0.024 for the level-2 variance parameter, and 0.013∼0.091 for the level-3 covariance parameters. Additionally, the SD of the fixed effect parameters had

Table 3: Evaluating the accuracy of item parameters based on the different prior distributions in simulation study 2.

	(i)					(ii)				
Item parameter	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
Discrimination $\boldsymbol{a}_{1\cdot}$	0.011	0.108	0.123	0.130	0.954	0.013	0.116	0.128	0.140	0.945
Discrimination $\boldsymbol{a}_{2\cdot}$	0.022	0.148	0.174	0.173	0.955	0.026	0.166	0.186	0.193	0.943
Discrimination $\boldsymbol{a}_{3\cdot}$	0.008	0.095	0.108	0.118	0.948	0.009	0.099	0.110	0.123	0.945
Difficulty $\boldsymbol{b}_{1\cdot}$	-0.005	0.090	0.105	0.112	0.954	-0.007	0.097	0.109	0.119	0.948
Difficulty $\boldsymbol{b}_{2\cdot}$	-0.015	0.067	0.073	0.087	0.947	-0.003	0.067	0.074	0.089	0.950
Difficulty $\boldsymbol{b}_{3\cdot}$	-0.031	0.078	0.079	0.090	0.918	-0.004	0.072	0.081	0.094	0.947
	(iii)					(iv)				
Item parameter	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
Discrimination $\boldsymbol{a}_{1\cdot}$	0.013	0.119	0.130	0.143	0.945	0.013	0.119	0.130	0.143	0.944
Discrimination $\boldsymbol{a}_{2\cdot}$	0.028	0.173	0.191	0.201	0.940	0.029	0.174	0.191	0.201	0.940
Discrimination $\boldsymbol{a}_{3\cdot}$	0.009	0.100	0.111	0.124	0.944	0.009	0.100	0.111	0.124	0.945
Difficulty $\boldsymbol{b}_{1\cdot}$	-0.008	0.099	0.110	0.121	0.947	-0.008	0.099	0.110	0.121	0.946
Difficulty $\boldsymbol{b}_{2\cdot}$	0.001	0.068	0.075	0.091	0.947	0.001	0.068	0.075	0.090	0.948
Difficulty $\boldsymbol{b}_{3\cdot}$	0.004	0.074	0.082	0.095	0.949	0.004	0.074	0.082	0.095	0.948

Note that the Bias, RMSE, SD, SE and CP denote the average Bias, RMSE, SD, SE and CP for the parameters at each time point.

Table 4: Evaluating the accuracy of the fixed and random effect parameters in simulation study 2.

	(i)					(ii)				
Fixed effect	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
β_{00}	-0.016	0.041	0.036	0.038	0.898	-0.004	0.040	0.037	0.039	0.924
β_{01}	0.002	0.016	0.015	0.016	0.960	0.003	0.016	0.015	0.016	0.946
β_{02}	-0.001	0.014	0.015	0.014	0.962	-0.001	0.014	0.016	0.014	0.952
β_{10}	-0.018	0.039	0.035	0.034	0.918	-0.001	0.036	0.036	0.036	0.968
β_{11}	-0.001	0.018	0.020	0.018	0.960	-0.001	0.018	0.020	0.018	0.960
β_{12}	-0.003	0.020	0.020	0.019	0.956	-0.004	0.020	0.021	0.019	0.952
β_{20}	-0.003	0.070	0.069	0.070	0.930	-0.003	0.073	0.071	0.073	0.934
β_{21}	0.003	0.029	0.028	0.029	0.938	0.004	0.029	0.028	0.029	0.940
β_{22}	-0.001	0.030	0.028	0.030	0.938	-0.001	0.030	0.028	0.030	0.934
Random effect	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
σ^2 (level-2 var.)	-0.019	0.022	0.017	0.011	0.864	-0.020	0.023	0.017	0.011	0.858
τ_{00}	0.019	0.024	0.019	0.014	0.934	0.017	0.022	0.019	0.014	0.948
τ_{10}	-0.007	0.013	0.012	0.011	0.922	-0.007	0.013	0.012	0.011	0.934
τ_{11}	0.022	0.027	0.020	0.015	0.890	0.023	0.027	0.020	0.015	0.884
τ_{20}	-0.034	0.038	0.025	0.015	0.858	-0.033	0.037	0.025	0.015	0.868
τ_{21}	0.008	0.020	0.021	0.019	0.956	0.008	0.020	0.021	0.019	0.948
τ_{22}	0.085	0.090	0.058	0.031	0.814	0.085	0.091	0.057	0.031	0.788
	(iii)					(iv)				
Fixed effect	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
β_{00}	-0.000	0.041	0.037	0.040	0.916	-0.000	0.040	0.037	0.040	0.916
β_{01}	0.003	0.016	0.015	0.016	0.950	0.003	0.016	0.015	0.016	0.948
β_{02}	-0.001	0.014	0.015	0.014	0.956	-0.001	0.015	0.015	0.015	0.956
β_{10}	0.005	0.037	0.036	0.036	0.968	0.005	0.037	0.036	0.036	0.968
β_{11}	-0.001	0.018	0.020	0.018	0.960	-0.001	0.018	0.020	0.018	0.962
β_{12}	-0.005	0.020	0.021	0.019	0.954	-0.004	0.020	0.020	0.019	0.954
β_{20}	-0.002	0.074	0.071	0.074	0.928	-0.003	0.074	0.071	0.074	0.930
β_{21}	0.003	0.029	0.028	0.029	0.940	0.004	0.029	0.028	0.028	0.946
β_{22}	-0.001	0.030	0.028	0.030	0.938	-0.002	0.030	0.028	0.030	0.936
Random effect	Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
σ^2 (level-2 var.)	-0.021	0.024	0.017	0.011	0.828	-0.021	0.024	0.017	0.011	0.834
τ_{00}	0.016	0.022	0.019	0.014	0.948	0.016	0.022	0.019	0.014	0.948
τ_{10}	-0.007	0.014	0.012	0.011	0.922	-0.007	0.014	0.012	0.011	0.922
τ_{11}	0.023	0.027	0.020	0.015	0.890	0.023	0.027	0.020	0.015	0.886
τ_{20}	-0.033	0.036	0.025	0.015	0.880	-0.033	0.036	0.025	0.015	0.874
τ_{21}	0.009	0.021	0.021	0.019	0.944	0.009	0.021	0.021	0.019	0.934
τ_{22}	0.085	0.091	0.057	0.031	0.782	0.086	0.091	0.057	0.031	0.786

a range of 0.015~0.071. The SD is 0.017 for the level-2 variance parameter under all four conditions, 0.012~0.058 for the level-3 covariance parameters. The SE had a range of 0.014~0.074 for the fixed effect parameters, 0.011~0.012 for the level-2 variance, and 0.011~0.031 for the level-3 covariance parameters. The recovery results of the structure parameters were almost the same under the four simulation conditions.

4.3 Simulation study 3

In this section, simulation study was designed to evaluate the performance of the two criteria in terms of selection the true model. We used the DIC and WAIC tools to identify a TS-IRT model combined with three different longitudinal multilevel models. The true LMTS-IRT model differed by (1)whether linear growth or quadratic growth was used as the true individual growth model; (2)whether significant individual covariates were included. The simulation study was described in detail below.

Simulation design

The number of time points was fixed at 3, the total number of items was set to 60 and there had 20 items including 4 anchor items at each time point. In addition, the number of individuals ($N = 500, 1000, 2000$) were considered. The same true values and the prior distributions were used as in simulation study 1. Three longitudinal multilevel models were given by

$$\textbf{Model 1}. \begin{cases} \theta_{ti} = \pi_{0i} + \pi_{1i} d_{ti} + e_{ti}, \\ \pi_{0i} = \beta_{00} + \beta_{01} x_{1i} + \beta_{02} x_{2i} + u_{0i}, \\ \pi_{1i} = \beta_{10} + \beta_{11} x_{1i} + \beta_{12} x_{2i} + u_{1i}, \end{cases} \tag{16}$$

where $e_{ti} \sim N(0, \sigma^2), t = 1, 2, 3$, $\begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & 0 \\ 0 & \tau_{11} \end{pmatrix}\right)$, and

$$\textbf{Model 2.} \begin{cases} \theta_{ti} = \pi_{0i} + \pi_{1i} d_{ti} + \pi_{2i} d_{ti}^2 + e_{ti}, \\ \pi_{0i} = \beta_{00} + \beta_{01} x_{1i} + \beta_{02} x_{2i} + u_{0i}, \\ \pi_{1i} = \beta_{10} + u_{1i}, \\ \pi_{2i} = \beta_{20} + u_{2i}. \end{cases} \tag{17}$$

where $e_{ti} \sim N(0, \sigma^2), t = 1, 2, 3$, $\begin{pmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{\Omega}\right)$, and

$$\textbf{Model 3.} \begin{cases} \theta_{ti} = \pi_{0i} + \pi_{1i} d_{ti} + \pi_{2i} d_{ti}^2 + e_{ti}, \\ \pi_{0i} = \beta_{00} + \beta_{01} x_{1i} + \beta_{02} x_{2i} + u_{0i}, \\ \pi_{1i} = \beta_{10} + \beta_{11} x_{1i} + \beta_{12} x_{2i} + u_{1i}, \\ \pi_{2i} = \beta_{20} + \beta_{21} x_{1i} + \beta_{22} x_{2i} + u_{2i}. \end{cases} \tag{18}$$

where $e_{ti} \sim N(0, \sigma^2), t = 1, 2, 3$, $\begin{pmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{\Omega}\right)$.

Nine simulated datasets (3 sample sizes $\times$ 3 growth trajectories) were generated from the TS-IRT model combined with longitudinal multilevel models(TS-IRT$\oplus$Model 1, TS-IRT$\oplus$Model 2 and TS-IRT$\oplus$Model 3). To compare the performances of different model selection methods, we ran 500 replications in each condition and computed the proportion of times when the generating model was selected as the true model.

Results

From Table 5, the results indicated that the percentages were fairly consistent between DIC and WAIC. When data were generated from TS-IRT$\oplus$Model 1, and those chose TS-IRT$\oplus$Model 1 with probalility higher than 92%. When data were generated from TS-IRT$\oplus$Model 3, and those chose TS-IRT$\oplus$Model 3 with probalility higher than

98.2%. However, When data were generated from TS-IRT$\oplus$Model 2, the percentages of two criteria cannot easily distinguish models(TS-IRT$\oplus$Model 2 and TS-IRT$\oplus$Model 3) that differ by multilevel covariates. This might be because the unremarkable difference between the TS-IRT$\oplus$Model 2 and TS-IRT$\oplus$Model 3 in the process of model selection. By calculating the specific values of the DIC and WAIC, we found that the DIC was low difference between the two models, and WAIC was low difference between the two models too. In the case of three sample sizes (N =500,1000 and 2000), Figure 4 showed that the medians of DIC differences between TS-IRT$\oplus$Model 2 and TS-IRT$\oplus$Model 3 were 3.844, 5.053 and 4.172, respectively. The medians of WAIC differences between TS-IRT$\oplus$Model 2 and TS-IRT$\oplus$Model 3 were 4.159, 5.444 and 4.673, respectively. Considering the very low difference, both DIC and WAIC were difficult to accurately select the true model, additional indexs might be needed. Other similar kinds of situations also occured in educational psychology (Zhang, Wang, & Tao, 2018). In our simulation study, the inclusion of covariates were considered ,the 95%HPDI of β_{11}, β_{12}, β_{21} and β_{22} can be calculated as a variable selection index(Zhang, Wang, & Tao, 2018) to evaluate whether the inclusion of covariates were needed in the model. This was because TS-IRT$\oplus$Model 2 to TS-IRT$\oplus$Model 3 differ essentially on whether the certain covariates were included. The proportions of the 95% HPDI of β_{11}, β_{12}, β_{21} and β_{22} contained zero were higher than 93.4% in the TS-IRT$\oplus$Model 3. The results indicated that these parameters were not significantly different from 0 and were not included in the model. Therefore, the TS-IRT$\oplus$Model 2 was an appropriate model to fit the 500 data sets which were generated from TS-IRT$\oplus$Model 2. In addition, as the number of individuals increased, the percentages of correct selection increased in most cases. Specifically, although WAIC seemed to perform slightly better than DIC, there were some conditions in which WAIC perform slightly worse. For example, when the generating model is TS-IRT$\oplus$Model 2, DIC has a slightly higher percentage of choosing the true model.

Table 5: The percentage of correct selection for the different simulated data sets using DIC and WAIC.

The number of individuals N=500						
	Model assessment methods					
	DIC			WAIC		
	Generation model			Generation model		
Calibration Model	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Model 1	**92**	0	1	**93.8**	0	1.8
Model 2	0	**36**	0	0	**34**	0
Model 3	8	64	**99**	6.2	66	**98.2**
The number of individuals N=1,000						
	Model assessment methods					
	DIC			WAIC		
	Generation model			Generation model		
Calibration Model	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Model 1	**92.6**	0	0	**92.8**	0	0
Model 2	0	**32**	0	0	**32**	0
Model 3	7.4	68	**100**	7.2	68	**100**
The number of individuals N=2,000						
	Model assessment methods					
	DIC			WAIC		
	Generation model			Generation model		
Calibration Model	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Model 1	**94.1**	0	0	**95.3**	0	0
Model 2	0	**38.4**	0	0	**35.4**	0
Model 3	5.9	61.6	**100**	4.7	64.6	**100**

5 Analysis of the longitudinal education quality assessment data

The dataset analyzed came from the Student Development Program (SDP) initiated by the Changchun Education Bureau that includes short-term and long-term plans. Compared to the long-term plan (three academic years from grade 1 to grade 3), the short-term plan (half a semester in an academic year) used in this study was focused mainly on the development of achievement in mathematics measured over a relatively short period of time. The short-term plan was designed to modify current teaching programs in a timely manner, and to put forward corresponding teaching strategies for different groups (genders or family socioeconomic statuses) of students with different growth trajectories.

The test data included a two-stage cluster sample of 3,109 students in grade 2 of junior middle schools. The students were from 16 different schools. The number of enrolled students ranged from 124 to 255 for different schools. The sampling population was first classified according to district, and schools were then selected at random. Second, students were selected at random from each school. Achievement in mathematics was measured over four time points (FSE, the first sectional examination; MTE, a middle-term exam; TSE, the third sectional examination; and FE, a final exam). Moreover, all 3,109 students were assessed at exactly the same time over the course of the study. Students took 24 items at each time point. Each set of items included 4 anchor items that do not overlap across time points. This lack of overlapping items across time points was designed to eliminate potential practical effects and to prevent the occurrence of security breaches. The anchor items were known and pre-linked. Here, we focused on a core sample of 2,000 students from 3,109 students. In addition, the level-2 background covariates of individuals were measured. At the individual level, gender (0=male, 1=female) and socioeconomic status (SES) were measured. The SES was measured based on the parents' degrees of education and scaled as five-point Likert items ranging from 0 to 4 (0=lowest, 4=highest).

5.1 Longitudinal multilevel IRT models

We considered four competing LMTS-IRT models to fit the real data. The level-1 model was a two-parameter TS-IRT model used to define the relationship between observable item responses and latent constructs. The TS-IRT model was the same but with four different longitudinal multilevel models.

Model 4 consists of the level-2 linear growth model and multilevel model. The random intercept π_{0i} in model 4 is explained by two background variables (*SES* and *Gender*) at level 3. The model has the following form:

$$\textbf{Model 4}. \begin{cases} \theta_{ti} = \pi_{0i} + \pi_{1i} Time_i + e_{ti}, \\ \pi_{0i} = \beta_{00} + \beta_{01} Gender_i + \beta_{02} SES_i + u_{0i}, \\ \pi_{1i} = \beta_{10} + u_{1i}. \end{cases} \tag{19}$$

where the error e_{ti} is normally distributed with mean zero and variance σ^2. The error terms at level 3, u_{0i} and u_{1i}, are bivariate normally distributed with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Omega}_1$, and they are independent of the level-2 residuals.

Model 5 is an extended version of model 4 by including two variables (*SES* and *Gender*) at level 3 to explain the random slope. Model 5 has the following form:

$$\textbf{Model 5}. \begin{cases} \theta_{ti} = \pi_{0i} + \pi_{1i} Time_i + e_{ti}, \\ \pi_{0i} = \beta_{00} + \beta_{01} Gender_i + \beta_{02} SES_i + u_{0i}, \\ \pi_{1i} = \beta_{10} + \beta_{11} Gender_i + \beta_{12} SES_i + u_{1i}. \end{cases} \tag{20}$$

Model 6 consists of the level-2 quadratic growth model and level-3 multilevel model. The random intercept π_{0i} and random slopes for the first (π_{1i}) and second (π_{2i}) order polynomial time effects, where the random intercept is defined conditionally on the *Gender* and *SES* variables. Model 6 is given by

$$\textbf{Model 6}. \begin{cases} \theta_{ti} = \pi_{0i} + \pi_{1i}Time_i + \pi_{2i}Time_i^2 + e_{ti}, \\ \pi_{0i} = \beta_{00} + \beta_{01}Gender_i + \beta_{02}SES_i + u_{0i}, \\ \pi_{1i} = \beta_{10} + u_{1i}, \\ \pi_{2i} = \beta_{20} + u_{2i}. \end{cases} \tag{21}$$

Model 7 is an extended version of model 6 by including two background variables (*SES* and *Gender*) at level 3 to explain the random slopes. Model 7 has the following form:

$$\textbf{Model 7}. \begin{cases} \theta_{ti} = \pi_{0i} + \pi_{1i}Time_i + \pi_{2i}Time_i^2 + e_{ti}, \\ \pi_{0i} = \beta_{00} + \beta_{01}Gender_i + \beta_{02}SES_i + u_{0i}, \\ \pi_{1i} = \beta_{10} + \beta_{11}Gender_i + \beta_{12}SES_i + u_{1i}, \\ \pi_{2i} = \beta_{20} + \beta_{21}Gender_i + \beta_{22}SES_i + u_{2i}. \end{cases} \tag{22}$$

The combined sampling procedure was applied to estimate parameters of various models. For each chain, 10,000 iterations were run with the first 2,500 iterations as the burn-in period.

5.2 Model selection and parameter estimation

First, the DIC and WAIC tools were used to identify the competing LMTS-IRT models. From Table 6, combining model 7 with the TS-IRT model generated the smallest effective number of model parameters, which was preferred given the DIC and WAIC values among the four competing models. It can be found that the quadratic growth model was more appropriate for fitting the real data than the linear growth model. In addition, the level-2 random-effect coefficients, which were modeled by individual-level covariates (level-3 *Gender* and *SES*), led to a serious reduction in the effective number of model parameters inferred from the p_D and p_{WAIC} values in Table 6.

According to the above model selection results, model 6 combined with the TS-IRT model as the best-fitting model is used to analyze the real data. The expectation a posteriori estimation, standard deviation, and 95% HPDI of the structural parameters were shown in Table 7. Figure 5 represented the posterior means and 95% HPDIs of the item discrimination and difficulty parameters, respectively. As the anchor items were known and pre-linked, there were totally 80 items need to be estimated. Now, we considered the following two practical issues.

Conditional on the level-3 *SES*, how should the male students perform compared to female students in terms of mathematics performance as measured at the four time points? Figure 6 showed the differences between male and female students in terms of mathematics performance given the level-3 *SES* (*SES*=0,...,4). Over time, the male students' mathematics abilities (circle) were generally better than those of the female students (pentagram). For the first two time points, differences between the male and female students in terms of mathematical ability were not remarkable. The findings revealed that the male students may have strong logical thinking and spatial thinking capacities that had not been fully identified through the preliminary assessment. Moreover, improvements in ability for the male and female students from families of moderate to high *SES* were found to occur faster than those of the other three categories (steeper growth trajectory). In addition, the students who are of the same *Gender* but have different *SES* do have different effects. According to Figure 7, for the male and female students, the average growth rates of the five curves were not the same. Over time, all of students' mathematical abilities improved. However, the higher one's *SES* was, the greater one's capacity becomed. Furthermore, the capacities of the female students with the lowest *SES* (i.e., *SES*=0) improved more slowly than those of the other four categories.

The analysis of growth trajectories may help one gain a stronger understanding of the development of student achievement over time. Both educators and students should properly understand *Gender**SES* differences and teachers should consciously manage to improve female student training in logical thinking and spatial thinking capacities in

Table 6: The results of Bayesian model assessment for the real data.

Model specification	p_D	DIC	p_{WAIC}	WAIC
The linear latent growth model				
Model 4: Random intercept and slope. Intercept is explatined by *Gender* and *SES* variables	5,682.5	142,604.3	5,303.8	142,557.2
Model 5: Random intercept and slope. Intercept and slope are explatined by *Gender* and *SES* variables	5,384.6	142,350.1	4,991.9	142,240.4
The quadratic latent growth model				
Model 6: Random intercept, first and second order and slopes. Intercept is explatined by *Gender* and *SES* variables	5,386.3	142,042.5	5,066.0	142,051.3
Model 7: Random intercept, first and second order and slopes. Intercept and slopes are explatined by *Gender* and *SES* variables	5,035.3	141,721.7	4,748.3	141,710.1

Table 7: Parameter estimates of the longitudinal multilevel model parameters for real data.

Fixed effect	Coefficient	Standard deviation	HPDI
β_{00}	0.027	0.018	$[-0.007, 0.061]$
β_{01}	-0.041	0.012	$[-0.064, -0.018]$
β_{02}	0.510	0.012	$[0.487, 0.534]$
β_{10}	1.459	0.020	$[1.431, 1.510]$
β_{11}	-0.110	0.011	$[-0.132, -0.087]$
β_{12}	0.506	0.013	$[0.482, 0.532]$
β_{20}	0.015	0.016	$[-0.014, 0.047]$
β_{21}	-0.154	0.011	$[-0.175, -0.133]$
β_{22}	0.018	0.012	$[-0.004, 0.041]$
Random effect	Coefficient	Standard deviation	HPDI
σ^2 (level-2 var.)	0.142	0.008	$[0.126, 0.156]$
τ_{00}	0.117	0.009	$[0.097, 0.135]$
τ_{10}	0.058	0.007	$[0.045, 0.072]$
τ_{11}	0.011	0.009	$[0.089, 0.125]$
τ_{20}	0.015	0.006	$[0.004, 0.027]$
τ_{21}	-0.025	0.006	$[-0.032, -0.013]$
τ_{22}	0.108	0.008	$[0.093, 0.125]$

junior middle school period.

6 Concluding Remarks

The developed LMTS-IRT model has three levels. At level 1, a TS-IRT model is selected to characterize item responses across time points. At level 2, a latent ability growth model that takes into account variations in latent traits across measurement occasions among persons is formulated. In the latent ability growth model, a polynomial growth curve is specified to describe how the expected value of a response variable changes over time. At level 3, a multilevel regression model is incorporated to describe variations in growth trajectories between persons. The simulation results show that our combined Bayesian algorithm provides accurate estimates of the model parameters in terms of s-maller bias and RMSE values. Simultaneously, the SD and SE are close to each other and the CP of 95% HPDI is around 95% for item parameters and fixed effect parameters. Therefore, the algorithm is effective and can be used to analyze the real data. In our simulation, DIC and WAIC are used to assess the competing models.

In the analysis of the longitudinal mathematical achievement data, some phenomena well worthy of consideration are revealed: first, male and female students with similar family *SES* do not show remarkable differences in ability during early periods of learning. However, over time, the mathematical capacities of male students become superior to those of female students. In addition, family *SES* has an important effect on students' mathematical abilities. The findings can help educators modify current teaching programs and put forward corresponding teaching strategies for different groups (*Gender* or *SES*) of students with different development trajectories. Therefore, it is expected that the analysis results may guide the development and improvement of educational quality monitoring mechanisms. The results of DIC and WAIC are similar, and select the same best model among a set of candidate models.

The current study has its limitation. Firstly, the CP for level-2 variance and level-3 covariance parameters were low to 78%. The Inverse Gamma distribution is generally considered as an uninformative prior of a single variance(level-2 variance), but studies have shown that when the variance is very small, Inverse Gamma distribution will indeed lead to the underestimation of the variance (Browne & Draper, 2006; Gelman, 2006). This may be the reason for the low CP value of the level 2 variance. For level-3 covariance, the typically used Inverse Wishart prior with small df and identity matrix $\boldsymbol{\Xi}$ is relatively uninformative. In many cases, this type of prior will have the smallest impact on the result. When the variances are quite small, the Inverse -Wishart prior distribution is informative so that the estimates for the variances will be sensitive to the Inverse -Wishart prior specification, resulting in over- or under-estimation for the variances depending on the specification of the prior distribution(Schuurman, Grasman, & Hamaker, 2016b; Chung et al., 2015). This may be the reason for the low CP value of the level 3 covariance matrix. In education and psychology , covariance structures are of great interests to researchers. However, forming new types of priors for covariance matrices can be very difficult. A popular way to form new priors for a covariance matrix is based on the matrix decomposition. Barnard et al. (2000) introduced a separation strategy to decompose a covariance matrix, and Liu, Zhang, and Grimm (2016) investigated the influence of separation strategy priors. They found that the use of separation strategy priors took much longer time than with Inverse-Wishart priors to obtain posterior samples. Moreover, with the increase of the dimension of covariance matrix, the use of separation strategy priors might cause some practical issues. In the existing educational and psychological literature, almost all studies have applied the Inverse -Gamma and Inverse-Wishart priors in Bayesian estimation. We will draw more attention to the choice of priors on the variance and covariance matrix in the future studies. Secondly, from an empirical perspective, we should assess the effect of more covariates and explore the effect of missing data, because longitudinal research with complete data are rare. Thirdly, more model selection methods can be used and expanded to select models for those more complex IRT models.

7 Appendix

Step 1: Sample the ability parameter θ_{ti} for the ith individual for the measurement occasion t. We independently draw θ_{ti}^* from the normal proposal distribution, i.e., $\theta_{ti}^* \sim N\left(\theta_{ti}^{(r-1)}, v_\theta^2\right)$. The acceptance probability is measured as

$$\alpha\left(\theta_{ti}^{(r-1)}, \theta_{ti}^*\right) = \min\left\{1, \frac{p\left(\boldsymbol{Y}_{ti}\left|\theta_{ti}^*, \boldsymbol{a}_{t\cdot}^{(r-1)}, \boldsymbol{b}_{t\cdot}^{(r-1)}\right.\right) p\left(\theta_{ti}^*\left|\boldsymbol{\pi}_i^{(r-1)}, (\sigma^2)^{(r-1)}, \boldsymbol{d}_{ti}\right.\right)}{p\left(\boldsymbol{Y}_{ti}\left|\theta_{ti}^{(r-1)}, \boldsymbol{a}_{t\cdot}^{(r-1)}, \boldsymbol{b}_{t\cdot}^{(r-1)}\right.\right) p\left(\theta_{ti}^{(r-1)}\left|\boldsymbol{\pi}_i^{(r-1)}, (\sigma^2)^{(r-1)}, \boldsymbol{d}_{ti}\right.\right)}\right\}. \tag{23}$$

Otherwise, the value of the preceding iteration is retained, i.e., $\theta_{ti} = \theta_{ti}^{(r-1)}$. Here, $\boldsymbol{Y}_{ti} = (Y_{ti1}, Y_{ti2}, \cdots, Y_{tiK})$. In Equation (23), $p\left(\boldsymbol{Y}_{ti}\left|\theta_{ti}^*, \boldsymbol{a}_{t\cdot}^{(r-1)}, \boldsymbol{b}_{t\cdot}^{(r-1)}\right.\right) = \prod_{k=1}^{K} (p_{tik})^{y_{tik}} (1 - p_{tik})^{1-y_{tik}}$.

Step 2: Sample the difficulty parameter b_{tk} for the measurement occasion t. We independently draw b_{tk}^* from the normal proposal distribution, i.e., $b_{tk}^* \sim N\left(b_{tk}^{(r-1)}, v_b^2\right)$. The acceptance probability is measured as

$$\alpha\left(b_{tk}^{(r-1)}, b_{tk}^*\right) = \min\left\{1, \frac{p\left(\boldsymbol{Y}_{tk}\left|\boldsymbol{\theta}_{t\cdot}^{(r)}, a_{tk}^{(r-1)}, b_{tk}^*\right.\right) p\left(b_{tk}^*\right)}{p\left(\boldsymbol{Y}_{tk}\left|\boldsymbol{\theta}_{t\cdot}^{(r)}, a_{tk}^{(r-1)}, b_{tk}^{(r-1)}\right.\right) p\left(b_{tk}^{(r-1)}\right)}\right\}. \tag{24}$$

Otherwise, the value of the preceding iteration is retained, i.e., $b_{tk} = b_{tk}^{(r-1)}$. Here, $\boldsymbol{Y}_{tk} = (Y_{t1k}, Y_{t1k}, \cdots, Y_{tnk})$. In Equation (24), $p\left(\boldsymbol{Y}_{tk}\left|\boldsymbol{\theta}_{t\cdot}, a_{tk}, b_{tk}\right.\right) = \prod_{i=1}^{n} (p_{tik})^{y_{tik}} (1 - p_{tik})^{1-y_{tik}}$. In addition, $p(b_{tk})$ is a normal prior distribution, i.e., $p(b_{tk}) \sim N(\mu_b, \sigma_b^2)$.

Step 3: Sample the discrimination parameter a_{tk} for the measurement occasion t. We independently draw a_{tk}^* from the log-normal proposal distribution, i.e., $a_{tk}^* \sim \log N\left(\log a_{ik}^{(r-1)}, v_a^2\right)$. The acceptance probability is measured as

$$\alpha\left(a_{tk}^{(r-1)}, a_{tk}^*\right) = \min\left\{1, \frac{p\left(\boldsymbol{Y}_{tk}\left|\boldsymbol{\theta}_{t\cdot}^{(r)}, a_{tk}^*, b_{tk}^{(r)}\right.\right) p\left(a_{tk}^*\right) a_{tk}^*}{p\left(\boldsymbol{Y}_{tk}\left|\boldsymbol{\theta}_{t\cdot}^{(r)}, a_{tk}^{(r-1)}, b_{tk}^{(r)}\right.\right) p\left(a_{tk}^{(r-1)}\right) a_{tk}^{(r-1)}}\right\}. \tag{25}$$

Otherwise, the value of the preceding iteration is retained, i.e., $a_{tk} = a_{tk}^{(r)}$. In Equation (25), . In addition, $p\left(a_{tk}\right)$ is a log-normal prior distribution, i.e., $p\left(a_{tk}\right) \sim \log N\left(\mu_a, \sigma_a^2\right)$.

Step 4: Sample the level-2 random coefficients $\boldsymbol{\pi}_i = \left(\pi_{0i}, \pi_{1i}, ..., \pi_{Hi}\right)'$ given $\boldsymbol{\theta}$, σ^2, $\boldsymbol{\beta}$, and $\boldsymbol{\Omega}$.

$$\theta_{ti} = \pi_{0i} + \pi_{1i} d_{ti} + \pi_{2i} d_{ti}^2 + ... + \pi_{Hi} d_{ti}^H + e_{ti},$$

$$\begin{pmatrix} \theta_{1i} \\ \theta_{2i} \\ \vdots \\ \theta_{Ti} \end{pmatrix} = \begin{pmatrix} 1 & d_{1i} & d_{1i}^2 & \cdots & d_{1i}^H \\ 1 & d_{2i} & d_{2i}^2 & \cdots & d_{2i}^H \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & d_{Ti} & d_{Ti}^2 & \cdots & d_{Ti}^H \end{pmatrix} \begin{pmatrix} \pi_{0i} \\ \pi_{1i} \\ \vdots \\ \pi_{Hi} \end{pmatrix} + \begin{pmatrix} e_{1i} \\ e_{2i} \\ \vdots \\ e_{Ti} \end{pmatrix}. \tag{26}$$

Equation (26) can be rewritten as

$$\boldsymbol{\theta}_i = \boldsymbol{D}_i \boldsymbol{\pi}_i + \boldsymbol{e}_i. \tag{27}$$

According to Equation (27), $\boldsymbol{\pi}_i$ is the vector of random regression coefficients following a normal distribution with mean vector $\widetilde{\boldsymbol{\pi}_i} = \left(\boldsymbol{D}_i' \boldsymbol{D}_i\right)^{-1} \left(\boldsymbol{D}_i' \boldsymbol{\theta}_i\right)$ and covariance matrix $\boldsymbol{\Sigma}_i = \left(\boldsymbol{D}_i' \boldsymbol{D}_i\right)^{-1} \sigma^2$, and the random regression coefficients $\boldsymbol{\pi}_i$ are induced by a normal prior distribution with mean vector $\boldsymbol{X}_i \boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Omega}$. Here, we define $\boldsymbol{x}_i = \left(1, x_{1i}, x_{2i}, ..., x_{Si}\right)'$, and $\boldsymbol{X}_i$ is the Kronecker product of an $(H+1)$ identity matrix and $\boldsymbol{x}_i$. That is, $\boldsymbol{X}_i = \mathbf{I}_{(H+1)} \otimes \boldsymbol{x}_i'$. Let $\boldsymbol{\beta}_h = \left(\beta_{h0}, \beta_{h1}, ..., \beta_{hS}\right)'$ and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', ..., \boldsymbol{\beta}_H'\right)'$, the full conditional posterior distribution of $\boldsymbol{\pi}_i$ is given by

$$p\left(\boldsymbol{\pi}_i \left| \boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\beta}, \boldsymbol{T}\right.\right) \propto p\left(\boldsymbol{\theta}_i \left| \boldsymbol{\pi}_i, \boldsymbol{\Sigma}_i\right.\right) p\left(\boldsymbol{\pi}_i \left| \boldsymbol{\beta}, \boldsymbol{\Omega}\right.\right)$$

$$\propto \exp\left\{-\frac{\left(\boldsymbol{\pi}_i - \widetilde{\boldsymbol{\pi}_i}\right)' \boldsymbol{D}_i' \boldsymbol{D}_i \left(\boldsymbol{\pi}_i - \widetilde{\boldsymbol{\pi}_i}\right)}{2\sigma^2}\right\} \exp\left\{-\frac{\left(\boldsymbol{\pi}_i - \boldsymbol{X}_i \boldsymbol{\beta}\right)' \boldsymbol{\Omega}^{-1} \left(\boldsymbol{\pi}_i - \boldsymbol{X}_i \boldsymbol{\beta}\right)}{2}\right\}, \tag{28}$$

$$\boldsymbol{\pi}_i \left| \boldsymbol{\theta}_i^{(r)}, \left(\sigma^2\right)^{(r-1)}, \boldsymbol{\beta}^{(r-1)}, \boldsymbol{\Omega}^{(r-1)} \right. \sim N\left(\left(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Omega}^{-1}\right)^{-1} \left(\sigma^{-2} \boldsymbol{D}_i' \boldsymbol{\theta}_i + \boldsymbol{\Omega}^{-1} \boldsymbol{X}_i \boldsymbol{\beta}\right), \left(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Omega}^{-1}\right)^{-1}\right).$$

Step 5: Sample $\boldsymbol{\beta}$, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^{'}, \boldsymbol{\beta}_2^{'}, ..., \boldsymbol{\beta}_H^{'}\right)^{'}$. $\boldsymbol{\beta}$ is the matrix of regression coefficients of regression of $\boldsymbol{\pi}_i$ on $\boldsymbol{X}_i$. The full conditional posterior distribution of $\boldsymbol{\beta}$ is given by

$$
\begin{aligned}
p\left(\boldsymbol{\beta}\left|\boldsymbol{\pi}, \boldsymbol{\Omega}\right.\right) &\propto \prod_{i=1}^{n} p\left(\boldsymbol{\pi}_i\left|\boldsymbol{\beta}, \boldsymbol{\Omega}\right.\right) p\left(\boldsymbol{\beta}\left|\boldsymbol{\Omega}\right.\right) \\
&\propto \exp\left\{-\frac{\sum_{i=1}^{n}\left(\boldsymbol{\pi}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)^{'}\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\pi}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)}{2}\right\} p\left(\boldsymbol{\beta}\left|\boldsymbol{\Omega}\right.\right),
\end{aligned}
$$

$$
\boldsymbol{\beta}\left|\boldsymbol{\pi}^{(r)}, \boldsymbol{\Omega}^{(r-1)}\right. \sim N\left(\left(\sum_{i=1}^{n}\boldsymbol{X}_i^{'}\boldsymbol{\Omega}^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^{'}\boldsymbol{\Omega}^{-1}\boldsymbol{\pi}_i, \left(\sum_{i=1}^{n}\boldsymbol{X}_i^{'}\boldsymbol{\Omega}^{-1}\boldsymbol{X}_i\right)^{-1}\right). \tag{29}
$$

Step 6: Sample the level-2 residual variance σ^2. A prior for is an $\text{Inverse} - \text{Gamma}\left(v, \omega\right)$ distribution. The full conditional posterior distribution of is given by

$$
p\left(\sigma^2\left|\boldsymbol{\theta}, \boldsymbol{\pi}, v, \omega\right.\right) \propto p\left(\boldsymbol{\theta}\left|\boldsymbol{\pi}, \sigma^2\mathbf{I}_{(T\times T)}\right.\right) p\left(\sigma^2\right)
$$

$$
\propto \left(\sigma^2\right)^{-\frac{n\times T}{2}} \exp\left\{-\frac{\sum_{i=1}^{n}\left(\boldsymbol{\theta}_i - \boldsymbol{D}_i\boldsymbol{\pi}_i\right)^{'}\left(\sigma^2\mathbf{I}_{(T\times T)}\right)^{-1}\left(\boldsymbol{\theta}_i - \boldsymbol{D}_i\boldsymbol{\pi}_i\right)}{2}\right\}\left(\sigma^2\right)^{-(v+1)}\exp\left(-\frac{\omega}{\sigma^2}\right).
$$

Let $F = \sum_{i=1}^{n}\left(\boldsymbol{\theta}_i - \boldsymbol{D}_i\boldsymbol{\pi}_i\right)^{'}\left(\boldsymbol{\theta}_i - \boldsymbol{D}_i\boldsymbol{\pi}_i\right)/2$, then we have

$$
\sigma^2\left|\boldsymbol{\theta}^{(r)}, \boldsymbol{\pi}^{(r)}, v, \omega\right. \sim \text{Inverse} - \text{Gamma}\left(v + nT/2, \omega + F\right). \tag{30}
$$

Step 7: Sample $\boldsymbol{\Omega}$. $\boldsymbol{\Omega}$ is the covariance matrix of ability at level 3. A prior for $\boldsymbol{\Omega}$ is an $\text{Inverse} - \text{Wishart}\left(\lambda, \Xi\right)$ distribution. The full conditional posterior distribution of $\boldsymbol{\Omega}$

is given by

$$p\left(\boldsymbol{\Omega}\left|\boldsymbol{\pi},\boldsymbol{\beta},\lambda,\boldsymbol{\Xi}\right.\right)\propto p\left(\boldsymbol{\pi}_i\left|\boldsymbol{\beta},\boldsymbol{\Omega}\right.\right)p\left(\boldsymbol{\Omega}\left|\lambda,\boldsymbol{\Xi}\right.\right)$$

$$\propto|\boldsymbol{\Omega}|^{-\frac{n}{2}}\exp\left\{-\frac{\sum_{i=1}^{n}\left(\boldsymbol{\pi}_i-\boldsymbol{X}_i\boldsymbol{\beta}\right)'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\pi}_i-\boldsymbol{X}_i\boldsymbol{\beta}\right)}{2}\right\}|\boldsymbol{\Omega}|^{-\frac{\lambda+H+2}{2}}\exp\left\{-\frac{trace\left(\boldsymbol{\Omega}^{-1}\boldsymbol{\Xi}\right)}{2}\right\}.$$

Let $F_1=\sum_{i=1}^{n}\left(\boldsymbol{\pi}_i-\boldsymbol{X}_i\boldsymbol{\beta}\right)\left(\boldsymbol{\pi}_i-\boldsymbol{X}_i\boldsymbol{\beta}\right)'$, then we have

$$\boldsymbol{\Omega}\left|\boldsymbol{\pi}^{(r)},\boldsymbol{\beta}^{(r)},\lambda,\boldsymbol{\Xi}\right.\sim\text{Inverse}-\text{Wishart}\left(\lambda+n,F_1+\boldsymbol{\Xi}\right). \tag{31}$$

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant number 11571069) and Natural Science Foundation of Changchun Normal University.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *2nd International Symposium on Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3-16.

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis*, *95*(1), 1-22.

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

Bacci, S. (2012). Longitudinal data: different approaches in the context of item-response theory models. *Journal of Applied Statistics*, *39*, 2047-2065.

Barnard, J., McCulloch, R., & Meng, X. (2000). Modeling covariances matrices in terms of standard deviations and correlations with applications to shrinkage. *Statistica Sinica*, textit10, 1281 – 1311.

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

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective.* Hoboken, NJ: Wiley Interscience.

Brooks, S. P., & Gelman A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434-455.

Browne, W. J., &Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473 – 514.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics*, *40*, 136 – 157.

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

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66, 269-286.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. T. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*, pp. 131-143. London: Chapman and Hall.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*, 997-1016.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457-511.

Gelman, A. (2006). Comment on Browne and Draper, Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 514 - 534.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Hahn, E. D. (2014). *Bayesian methods for management and business: Pragmatic solutions for real problems.* NJ: Wiley Hoboken.

Harville, D. A. (1977). Maximum likelihood approaches to variance components estimation and related problems. *Journal of the American Statistical Association*, *72*, 320-340.

Hastings, W. K. (1970). Monte Carlo sampling-based methods using Markov chains and their applications. *Biometrika*, *57*, 97-109.

Kim, S., & Camilli, G. (2014). An item response theory approach to longitudinal analysis with application to summer setback in preschool language/literacy. *Large-scale Assessments in Education*, *2*(1), 1-17.

Laird, N. M., & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, *38*, 963-974.

Liu, H., Zhang, Z., & Grimm, K. J. (2016). Comparison of Inverse-Wishart and SeparationStrategy Priors for Bayesian Estimation of Covariance Parameter Matrix in Growth Curve Analysis. *Structural Equation Modeling*, *23*(3), 354-367.

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

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

Luo, Y., & Al-Harbi, K. (2017). Performances of LOO and WAIC as IRT model selection methods. *Psychological Test and Assessment Modeling*, *59*, 183-205.

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

Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81-117.

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

Patz, R.J., & Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342-366.

Raftery, A. E., & Lewis, S. M. (1996). Implementing MCMC. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*, pp. 115-130. London: Chapman and Hall.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, 2nd ed. Thousand Oaks, CA: Sage.

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E.L. (2016b). A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, *51(2 – 3)*, 185 – 206.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & 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.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussions). *The Annals of Statistics*, *22*, 1701-1762.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leaveone-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413-1432.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*, 318-336.

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

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571-3594.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, *14*, 867-897.

Zhang, X., Tao, J., & Wang, C. (2019). Bayesian model selection methods for multilevel IRT models: A comparison of five DIC-based indices. *Journal of Educational Measurement*, *56*, 3-27.

8 Authors' addresses

Shuang Qu

1.School of Mathematics and Statistics,

Northeast Normal University,

Changchun, 130024, Jilin, China

2.School of Mathematics,

Changchun Normal University,

Changchun, 130032, Jilin, China

E-mail address:qus687@nenu.edu.cn

Jiwei Zhang

3.School of Mathematics and Statistics,

Yunnan University,

Kunming, 650091, Yunnan, China

E-mail address:zhangjw713@nenu.edu.cn

Jian Tao

1.School of Mathematics and Statistics,

Northeast Normal University,

Changchun, 130024, Jilin, China

E-mail address:taoj@nenu.edu.cn

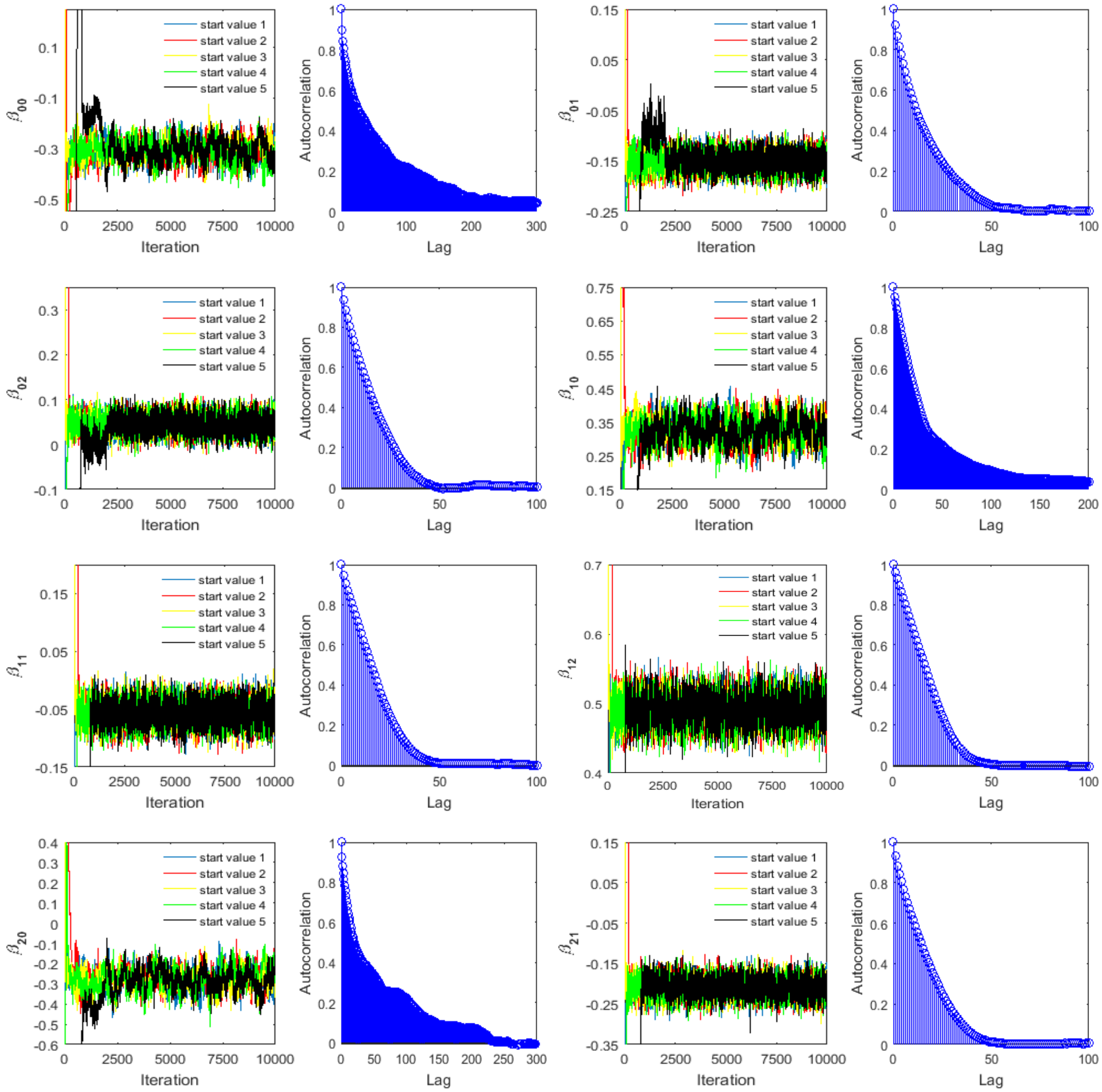


Figure 1: The trace and autocorrelation plots for the fixed-effect parameters $\boldsymbol{\beta}$. Note that the first 2500 iterations are discarded as burn-in time.

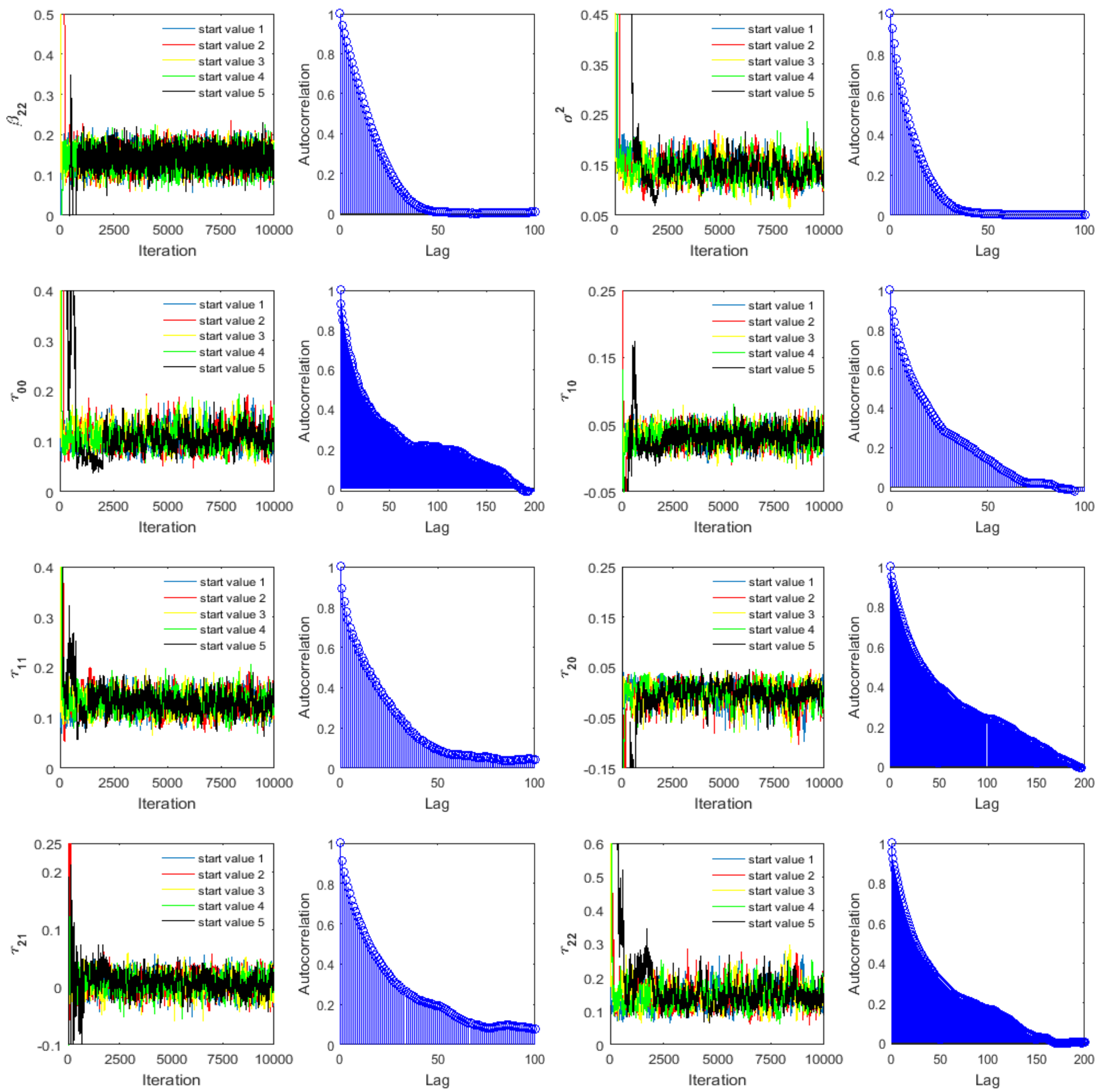


Figure 2: The trace and autocorrelation plots for the fixed-effect parameters $\boldsymbol{\beta}$, level-2 variance parameter σ^2, and the level-3 covariance parameters $\boldsymbol{\Omega}$. Note that the first 2500 iterations are discarded as burn-in time.

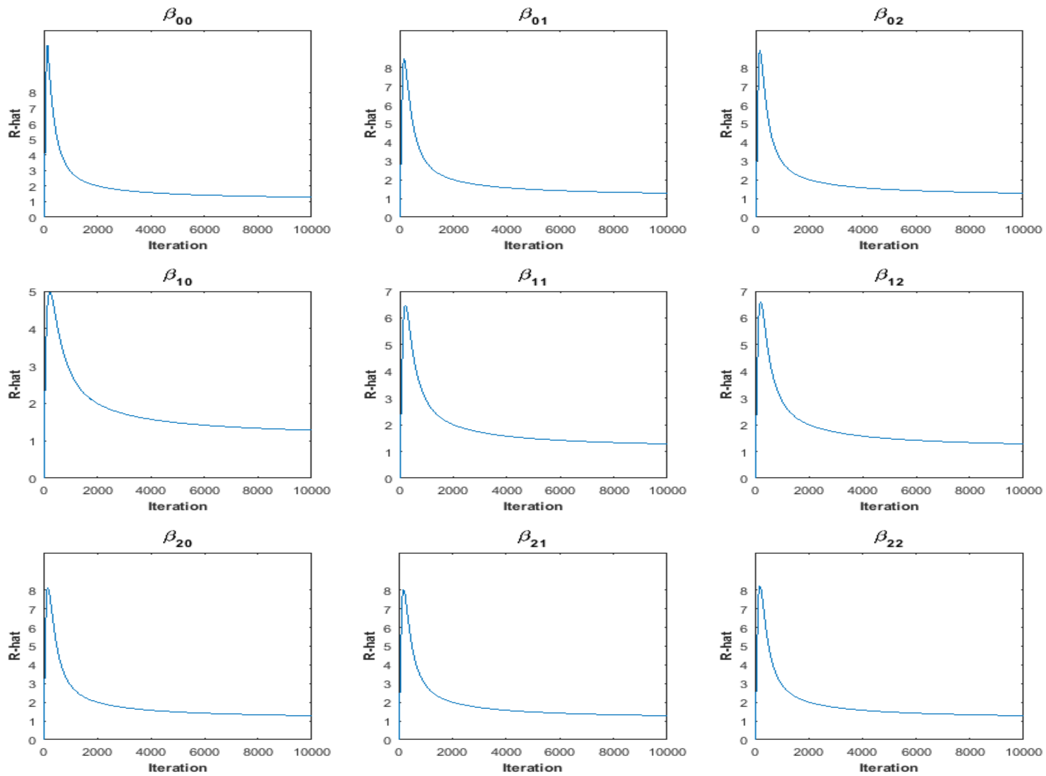
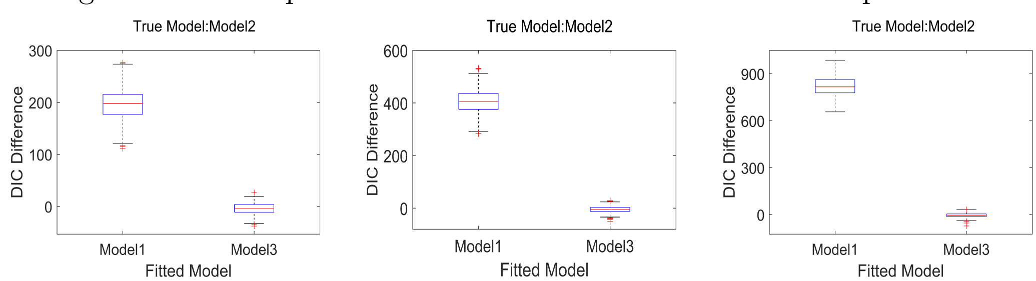
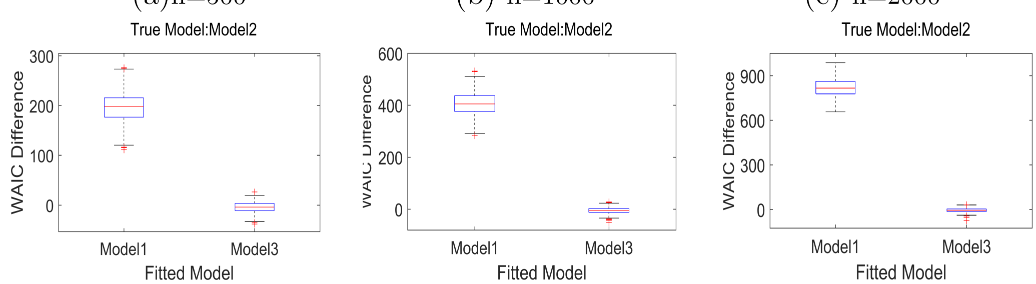


Figure 3: The sequence of $\widehat{R}$ values of for multilevel model parameters.



(a)n=500 (b) n=1000 (c) n=2000



(d)n=500 (e) n=1000 (f) n=2000

Figure 4: Boxpolts of DIC and WAIC based on the true model 2 in the simulation study 3.n=the number of individuals.

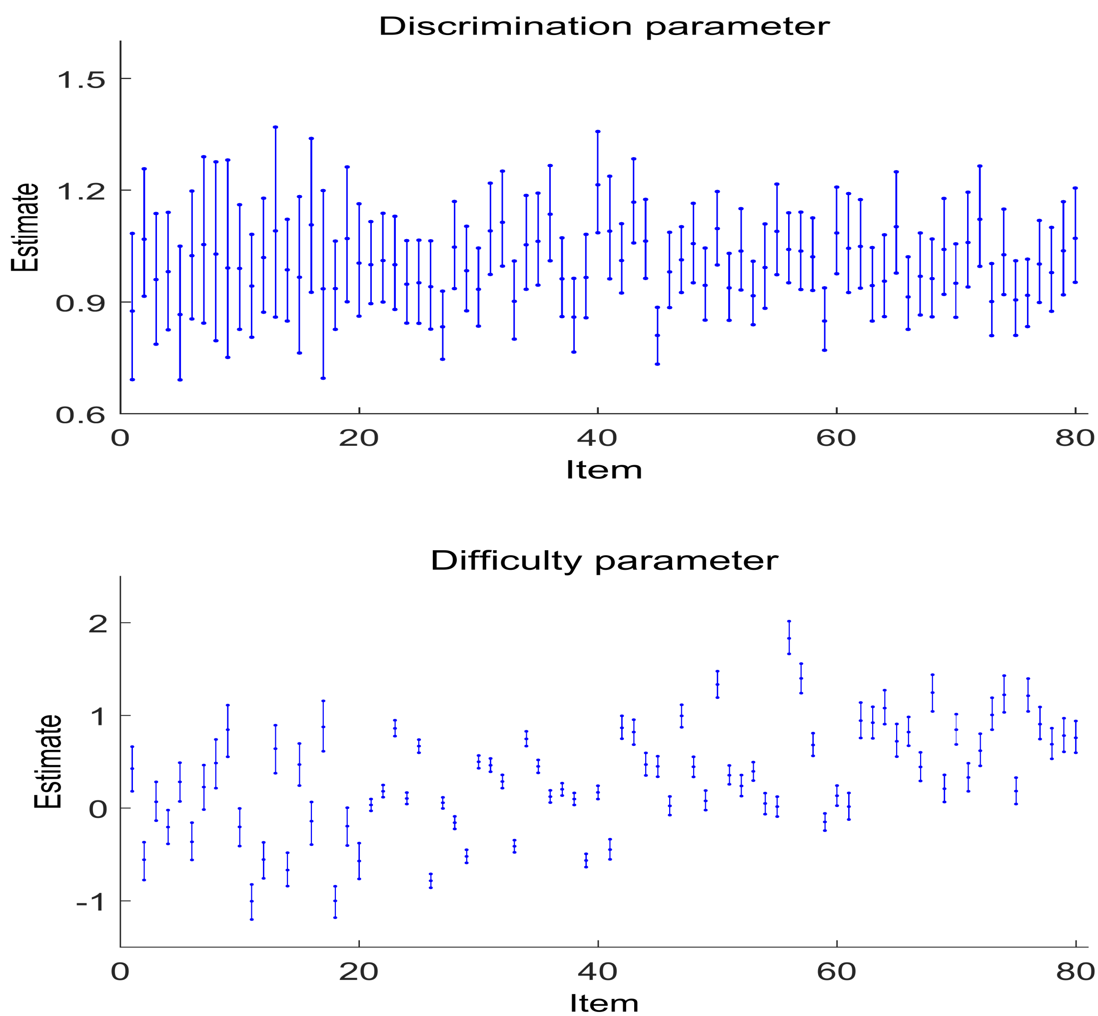


Figure 5: Posterior means and 95% HPDIs for the discrimination and difficulty parameters of SDP data.

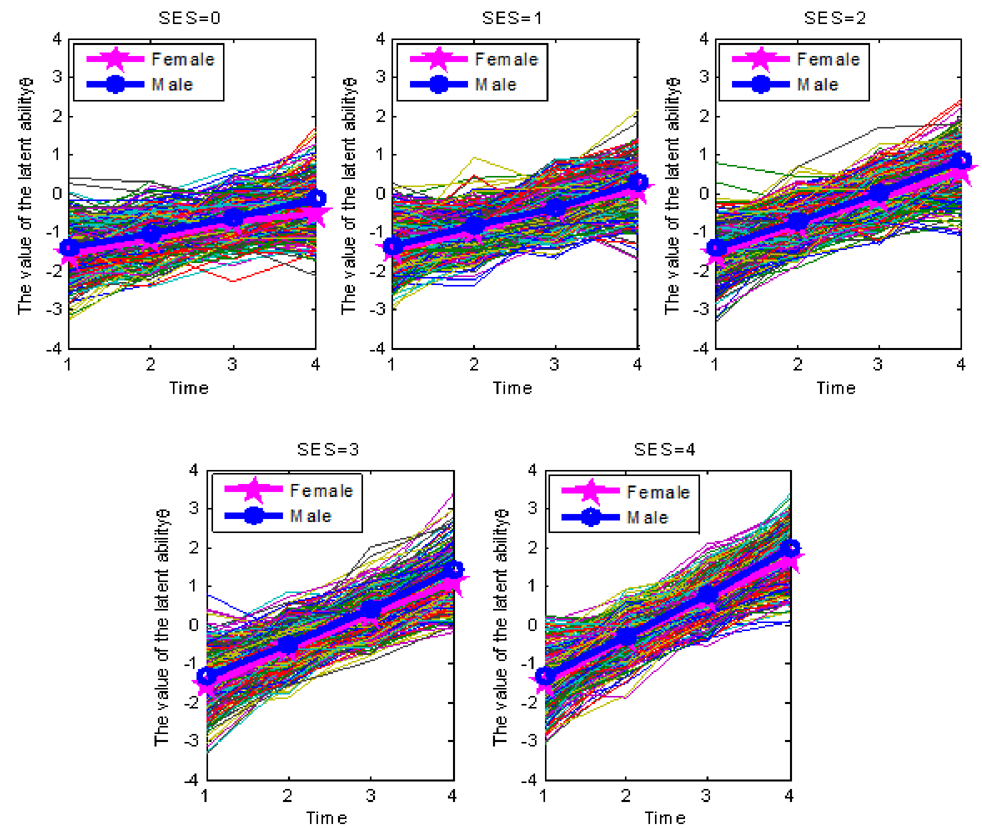


Figure 6: The development trajectories of latent ability for male and female students given a family SES.

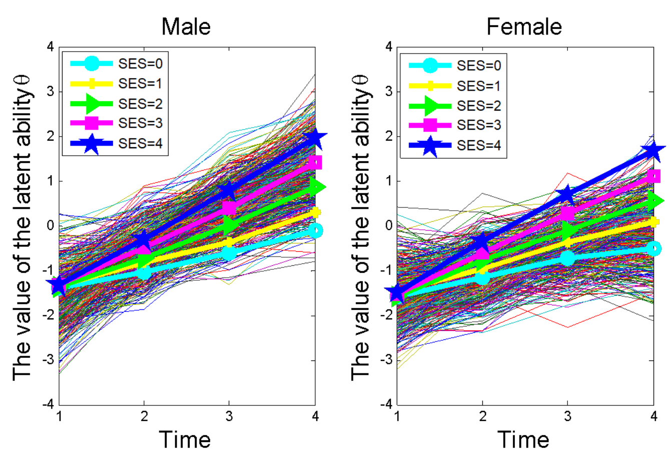


Figure 7: The development trajectories of latent ability for students for different family SES.