Statistics and Its Interface Volume 15 (2022) 463–474

A Gibbs sampler for estimating the graded item response model with Likert-scale data via the Pólya–Gamma distribution: a calculationally efficient data-augmentation scheme

Zhaoyuan Zhang, Jiwei Zhang*, and Jing Lu*

This paper reports the use of a highly effective Pólya–Gamma Gibbs sampling algorithm [32] based on auxiliary variables to estimate the parameters of the graded response model (GRM; [34]) that has been used widely in educational and psychological assessments. As its name suggests, the algorithm can be viewed as an extension of the traditional Gibbs sampling algorithm, overcoming the defect that the latter is ineffective for Bayesian non-conjugate models. By introducing auxiliary variables, non-conjugate models are transformed into conjugate ones, and posterior sampling is easier to implement with the help of the traditional Gibbs sampling algorithm. Also, the algorithm avoids the Metropolis–Hastings sampling algorithm's tedious adjustment of tuning parameters to achieve an appropriate acceptance probability. Two simulation studies are conducted, and data from the Sexual Compulsivity Scale are subjected to detailed analysis to further illustrate the proposed methodology.

Keywords and phrases: Auxiliary variables, Bayesian estimation methods, Graded response model, Item response theory, Pólya–Gamma Gibbs sampling algorithm.

1. INTRODUCTION

In educational and psychological assessments, many researchers prefer to use surveys, questionnaires, and scales with Likert-type items that often consist of multiple, ordered response categories, such as “strongly disagree”, “disagree”, “undecided”, “agree”, and “strongly agree” ([7], [10], [35], [40]). To analyze these items with a polytomous format, various complex nonlinear item response theory (IRT) models have also been developed, such as the graded response model (GRM; [34]), nominal response model [8], rating scale model [3], partial credit model [26], generalized partial credit model [28], and sequential response model ([43], [44]). Of these models, we focus on the GRM as the most widely used IRT model for ordinal polytomous response data in psychological measurements. However, parameter estimation has been a major concern in the application of the GRM. In fact, simultaneous estimations of items and examinee's latent ability result in statistical complexities in the estimation task.

Within the fully Bayesian framework, Markov-chain Monte Carlo (MCMC) methods are extremely general and flexible and have proved useful in parameter estimation and model comparisons. Bayesian procedures have been developed for dichotomous IRT models ([1], [6], [23], [25], [30], [31], [33], [37], [38], [39], [40], [47], [49], [50]), and the polytomous GRM ([2], [15], [23], [28], [46], [51]). In fact, two Bayesian methods are often used to estimate the parameters of the GRM. One is the Gibbs sampling algorithm ([17], [19], [42]) based on auxiliary variables to estimate the parameters of the GRM with probit link function ([2], [15], [23]), and the other is the Metropolis–Hastings (MH) sampling algorithm ([12], [13], [21], [27]) to estimate the parameters of the GRM with logit link function ([30], [31], [46]).

In the present study, an efficient Pólya–Gamma Gibbs sampling algorithm [32] in a fully Bayesian framework is proposed for estimating the parameters of the GRM. Compared with the traditional MH and Gibbs sampling algorithms, we analyze the advantages of the Pólya–Gamma Gibbs sampling algorithm from multiple perspectives. First, the Pólya–Gamma Gibbs sampling algorithm avoids the retrospective tuning in the MH sampling algorithm if either we do not know how to choose a proper tuning parameter or no value for the tuning parameter is appropriate. Second, the Pólya–Gamma Gibbs sampling algorithm can transform a non-conjugate model into a conjugate one by using augmented auxiliary variables, and posterior sampling is easier to implement with the help of the traditional Gibbs sampling algorithm. Third, the prior specifications and prior sensitivity are important aspects of Bayesian inferences [20]. In fact, the Pólya–Gamma Gibbs sampling algorithm is insensitive to the specification of prior distribution and can still obtain satisfactory results even if the improper or misspecification priors are adopted. For a discussion about

*Jiwei Zhang and Jing Lu are co-corresponding authors.

different types of prior distributions, please see the supplementary material, http://intlpress.com/site/pub/files/_supp/sii/2022/0015/0004/SII-2022-0015-0004-s001.pdf.

The rest of this paper is organized as follows. In Section 2, the GRM is presented to explain the polytomous item response data. In Section 3, we discuss how the Pólya–Gamma Gibbs sampling algorithm implement sampling on the GRM. In Section 4, we present two simulation examples that focus on the performance of the Pólya–Gamma Gibbs sampling algorithm in parameter recovery based on different sample sizes, and we present the accuracy of parameter estimation for the Pólya–Gamma Gibbs sampling algorithm and the MH sampling algorithm. In Section 5, the performance of the Pólya–Gamma Gibbs sampling algorithm in a practical situation is shown by means of an empirical example. Finally, some concluding remarks are presented in Section 6.

2. MODEL AND MODEL IDENTIFICATION

2.1 Graded response model

The GRM [34], is used to fit polytomous item response data. The probability that examinee i scores in category k on item j is modeled by the GRM as

$$P_{jk}(\theta_i) = p(Y_{ij} = k \,|\theta_i, a_j, \boldsymbol{b}_j) = \Psi^*_{j,k}(\theta_i) - \Psi^*_{j,k+1}(\theta_i). \tag{1}$$

In Eq. (1), Y_{ij} is the response of examinee i answering item j, where $i = 1, \ldots, N$, $j = 1, \ldots, J$, and $k = 0, 1, \ldots, K$. θ_i is the latent ability for examinee i. a_j is the discrimination parameter for item j, and $\boldsymbol{b}_j$ is a $(K + 1)$-dimensional vector of threshold parameters for item j, i.e., $\boldsymbol{b}_j = (b_{j,0}, b_{j,1}, b_{j,2}, \ldots, b_{j,K})'$. $\Psi^*_{j,k}(\theta_i)$ is the boundary probability for examinee i having a score larger or equal to k on item j, and the boundary curve is given by

$$\Psi^*_{j,k}(\theta_i) = \frac{\exp(a_j\theta_i - b_{j,k})}{1 + \exp(a_j\theta_i - b_{j,k})}, \tag{2}$$

where $k = 0, \ldots, K$, $\Psi^*_{j,0}(\theta_i) = 1$, and $\Psi^*_{j,K}(\theta_i) = 0$. Furthermore, the boundaries between the response categories are represented by an ordered vector of thresholds

$$b_{j,0} < b_{j,1} < b_{j,2} < \cdots < b_{j,k} < \cdots < b_{j,K}, \tag{3}$$

where $b_{j,0} = -\infty, b_{j,K} = +\infty$. Therefore, in total there are $K - 1$ threshold parameters and one discrimination parameter for each item.

2.2 Model identification

In Eq. (2), the linear parts of the GRM can be written as

$$a_j\theta_i - b_{j,k},\ k = 1, 2, \ldots, K,$$

where we fix the mean population level of ability to zero to eliminate the trade-off between ability θ and threshold parameter b in location, i.e., $\mu_\theta = 0$. Meanwhile, to eliminate the trade-off between ability θ and discrimination parameter a in scale, we restrict the variance population level of ability to one, i.e., $\sigma^2_\theta = 1$. For a similar identification limitation method, see [9], [15], [16] and [24].

3. BAYESIAN ESTIMATION METHODS

3.1 Pólya–Gamma Gibbs sampling algorithm

[32] proposed a new data-augmentation strategy for fully Bayesian inference in logistic regression. The data-augmentation approach appeals to a new class of Pólya–Gamma distribution rather than the data-augmentation algorithm by [2] based on a truncated normal distribution. Next, we introduce the Pólya–Gamma distribution.

Definition. Let $\{T_k\}_{k=1}^{+\infty}$ be a sequence of i.i.d. random variables from a Gamma distribution with parameters λ and 1, i.e., $T_k \sim \text{Gamma}(\lambda, 1)$. A random variable W follows a Pólya–Gamma distribution with parameters $\lambda > 0$ and $\tau \in R$, denoted $W \sim \text{PG}(\lambda, \tau)$, if

$$W \overset{D}{=} \frac{1}{2\pi} \sum_{k=1}^{+\infty} \frac{T_k}{\left(k - \frac{1}{2}\right)^2 + \frac{\tau^2}{4\pi^2}},$$

where $\overset{D}{=}$ denotes equality in distribution. In fact, the Pólya–Gamma distribution is an infinite mixture of Gamma distributions, thereby providing the plausibility for sampling from such distributions. Within the Bayesian framework, in the method of auxiliary variables, realizations from a complicated distribution can be obtained by augmenting the auxiliary variables of interest by one or more additional variables such that the full conditional posterior distributions are tractable and easy to simulate from. Next, we give the full conditional posterior distribution based on the auxiliary variable Z_{ij}:

$$\begin{aligned}
&p(\boldsymbol{Z} \,|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{Y}) \\
&\propto p(\boldsymbol{Z} \,|\boldsymbol{a}, \boldsymbol{\theta})\, p(\boldsymbol{Y} \,|\boldsymbol{b}, \boldsymbol{Z}, \boldsymbol{W}) \\
&\propto \prod_{i=1}^{N}\prod_{j=1}^{J} \left\{ f(Z_{ij}) \sum_{k=1}^{K} \mathrm{I}(Y_{ij} = k)\, \mathrm{I}(b_{j,k-1} < Z_{ij} \le b_{j,k}) \right\} \\
&\propto \prod_{i=1}^{N}\prod_{j=1}^{J} \left\{ \frac{\exp(-Z_{ij} + a_j\theta_i)}{[1 + \exp(-Z_{ij} + a_j\theta_i)]^2} \right. \\
&\quad \left. \times \sum_{k=1}^{K} \mathrm{I}(Y_{ij} = k)\, \mathrm{I}(b_{j,k-1} < Z_{ij} \le b_{j,k}) \right\} \\
&\propto \prod_{i=1}^{N}\prod_{j=1}^{J} \left\{ \left(\frac{1}{2}\right)^2 \int_0^{+\infty} \exp\left[-\frac{W_{ij}(-Z_{ij} + a_j\theta_i)^2}{2} \right] p(W_{ij})\, dW_{ij} \right. \\
&\quad \left. \times \sum_{k=1}^{K} \mathrm{I}(Y_{ij} = k)\, \mathrm{I}(b_{j,k-1} < Z_{ij} \le b_{j,k}) \right\},
\end{aligned}$$

where the indicator function $\mathrm{I}(A)$ equals 1 if A is true or 0 if A is false. $p(W_{ij})$ is $p(W_{ij};\beta=2,d=0)$, where $p(W;\beta=2,d=0)$ denotes the density of the auxiliary random variable $W\sim\mathrm{PG}(\beta=2,d=0)$. The last inequality can be obtained by the Pólya–Gamma method [36], which is useful when working with logistic likelihoods, and has the form

$$\frac{[\exp(\psi)]^{\eta}}{[1+\exp(\psi)]^{\beta}} = \left(\frac{1}{2}\right)^{\beta}\exp(\kappa\psi)\int_{0}^{+\infty}\exp\left(-\frac{W\beta^{2}}{2}\right)p(W;\beta,0)\,dW,$$

where $\kappa=\eta-\frac{\beta}{2}$. The specific expression of $p(W;\beta,d)$ is

$$\begin{aligned} p(W;\beta,d) &= \left\{\cosh^{\beta}\left(\frac{d}{2}\right)\right\}\frac{2^{\beta-1}}{\Gamma(\beta)} \\ &\times\sum_{m=1}^{+\infty}\left\{\frac{(-1)^{m}\Gamma(m+\beta)(2m+\beta)}{\Gamma(m+1)\sqrt{2\pi W^{3}}}\right. \\ &\left.\times\exp\left[-\frac{(2m+\beta)^{2}}{8W}-\frac{Wd^{2}}{2}\right]\right\}, \end{aligned}$$

where cosh denotes the hyperbolic cosine. Then the joint posterior distribution of Z_{ij} and W_{ij} is

$$\begin{aligned} &p(\boldsymbol{Z},\boldsymbol{W}\,|\boldsymbol{a},\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{Y}) \\ &\propto\prod_{i=1}^{N}\prod_{j=1}^{J}\left\{\left(\frac{1}{2}\right)^{2}\exp\left[-\frac{W_{ij}(-Z_{ij}+a_j\theta_i)^{2}}{2}\right]\right. \\ &\left.\times p(W_{ij};\beta{=}2,d{=}0)\sum_{k=1}^{K}\mathrm{I}(Y_{ij}=k)\,\mathrm{I}(b_{j,k-1}<Z_{ij}\le b_{j,k})\right\}. \end{aligned}$$

Therefore, the full conditional posterior distribution of the auxiliary variable Z_{ij} is a truncated normal distribution, i.e.,

$$\begin{aligned} &Z_{ij}\,|W_{ij},a_j,\theta_i,\boldsymbol{b}_j,Y_{ij}=k \\ &\sim N\left(a_j\theta_i,\ \frac{1}{W_{ij}}\right)\mathrm{I}(b_{j,k-1}<Z_{ij}\le b_{j,k}). \end{aligned}$$

We use the inverse transform technique to sample the auxiliary variable Z_{ij} based on the truncated normal distribution. First, we sample the random variable U_{ij} from a uniform distribution, i.e., $U_{ij}\sim\mathrm{Uniform}(0,1)$. When the response $Y_{ij}=k$, the following equation can be established:

$$U_{ij}=\frac{\Phi\left[\sqrt{W_{ij}}\,(Z_{ij}-a_j\theta_i)\right]-\Phi\left[\sqrt{W_{ij}}\,(b_{j,k-1}-a_j\theta_i)\right]}{\Phi\left[\sqrt{W_{ij}}\,(b_{j,k}-a_j\theta_i)\right]-\Phi\left[\sqrt{W_{ij}}\,(b_{j,k-1}-a_j\theta_i)\right]},$$

or equivalently

$$\begin{aligned} &Z_{ij} \\ &=\Phi^{-1}\Big\{U_{ij}\left\{\Phi\left[\sqrt{W_{ij}}\,(b_{j,k}-a_j\theta_i)\right]\right. \\ &\qquad\left.-\,\Phi\left[\sqrt{W_{ij}}\,(b_{j,k-1}-a_j\theta_i)\right]\right\} \\ &\quad+\Phi\left[\sqrt{W_{ij}}\,(b_{j,k-1}-a_j\theta_i)\right]\Big\}\times\frac{1}{\sqrt{W_{ij}}}+a_j\theta_i, \end{aligned}$$

where $\Phi(\bullet)$ is a normal cumulative distribution with mean 0 and variance 1. The full conditional posterior distribution of the auxiliary variable W_{ij} is

$$\begin{aligned} &p(W_{ij}\,|Z_{ij},a_j,\theta_i) \\ &\propto\left(\frac{1}{2}\right)^{2}\exp\left[-\frac{W_{ij}(-Z_{ij}+a_j\theta_i)^{2}}{2}\right]p(W_{ij};\beta=2,d=0). \end{aligned}$$

Based on [32], we obtain

$$W_{ij}\,|Z_{ij},a_j,\theta_i\sim\mathrm{PG}(2,\ \ -Z_{ij}+a_j\theta_i).$$

Next, we update the discrimination parameter a_j for each item j in the GRM. The prior of a_j is assumed to follow a truncated normal distribution, i.e., $a_j\sim N(\mu_a,\sigma_a^{2})\,\mathrm{I}(a_j>0)$. The full conditional posterior distribution of the discrimination parameter a_j is

$$\prod_{i=1}^{N}\left\{\exp\left[-\frac{W_{ij}(-Z_{ij}+a_j\theta_i)^{2}}{2}-\frac{(a_j-\mu_a)^{2}}{2\sigma_a^{2}}\right]\right\}\mathrm{I}(a_j>0).$$

Therefore, the full conditional posterior distribution of a_j follows a truncated normal distribution with mean

$$\mathrm{Var}_{a_j}\times\left(\mu_a\sigma_a^{-2}+\left(\sum_{i=1}^{N}\theta_i^{2}W_{ij}\right)\left[\frac{\left(\sum_{i=1}^{N}\theta_i Z_{ij}W_{ij}\right)}{\left(\sum_{i=1}^{N}\theta_i^{2}W_{ij}\right)}\right]\right)$$

and variance

$$\mathrm{Var}_{a_j}=\left(\sigma_a^{-2}+\left(\sum_{i=1}^{N}\theta_i^{2}W_{ij}\right)\right)^{-1}.$$

Similarly, we update the ability parameter θ_i for each examinee i. The prior distribution of θ_i is assumed to follow a normal distribution with mean μ_θ and variance σ_θ^{2}. The full conditional posterior distribution of the discrimination parameter θ_i is

$$\prod_{j=1}^{J}\left\{\exp\left[-\frac{W_{ij}(-Z_{ij}+a_j\theta_i)^{2}}{2}-\frac{(\theta_i-\mu_\theta)^{2}}{2\sigma_\theta^{2}}\right]\right\}.$$

Therefore, the full conditional posterior distribution of θ_i follows a normal distribution with mean

$$\mathrm{Var}_{\theta_i} \times \left(\mu_\theta \sigma_\theta^{-2} + \left(\sum_{j=1}^{J} a_j^2 W_{ij} \right) \left[\frac{\left(\sum_{j=1}^{J} a_j Z_{ij} W_{ij} \right)}{\left(\sum_{j=1}^{J} a_j^2 W_{ij} \right)} \right] \right)$$

and variance

$$\mathrm{Var}_{\theta_i} = \left(\sigma_\theta^{-2} + \left(\sum_{j=1}^{J} a_j^2 W_{ij} \right) \right)^{-1}.$$

In fact, two methods are often used to update the threshold parameters. One is the Gibbs sampling algorithm based on auxiliary variables to draw the posterior samples from a uniform distribution ([2], [15], [41]). The specific implementation process is to derive the full conditional distribution using a conjugate prior that takes the order constraint in Eq. (3) into account. Define a uniformly distributed variable V_{ij} over $[0, 1]$ such that

$$V_{ij} \le p\left(Z_{ij} \le b_{j,k} \left| \boldsymbol{Y}, \boldsymbol{b}_j, \theta_i, a_j \right.\right) \mathrm{I}\left(i \in \Delta_1\right),$$
$$V_{ij} > p\left(Z_{ij} > b_{j,k} \left| \boldsymbol{Y}, \boldsymbol{b}_j, \theta_i, a_j \right.\right) \mathrm{I}\left(i \in \Delta_2\right),$$

where the set $\Delta_1 = \{i : Y_{ij} = k\}$ and the set $\Delta_2 = \{i : Y_{ij} = k + 1\}$. Accordingly, the full conditional distribution of $b_{j,k}$ is uniform using a diffuse prior with equal probability for each possible parameter value, i.e.,

$$b_{j,k} \left| \boldsymbol{Z}, \boldsymbol{b}_{j,(-k)}, \boldsymbol{\theta}_i, a_j \right. \sim \mathrm{Uniform}\left(\Omega_L,\ \Omega_U\right),$$

where $\Omega_L = \max\left(\max_{i:Y_{ij}=k} Z_{ij}, b_{j,k-1}\right)$, $\Omega_U = \min\left(\min_{i:Y_{ij}=k+1} Z_{ij}, b_{j,k+1}\right)$, and $\boldsymbol{b}_{j,(-k)}$ is the set of threshold parameters for item j without $b_{j,k}$. In fact, any prior distribution for $b_{j,k}$ can be used as long as the values sampled from it satisfy a reasonable range of parameter support set, i.e.,

$$b_{j,k} \left| \boldsymbol{Z}, \boldsymbol{b}_{j,(-k)}, \boldsymbol{\theta}_i, a_j \right. \sim f_{prior}\left(b_{j,k}\right) \mathrm{I}\left(\Omega_L,\ \Omega_U\right),$$

where Ω_L and Ω_U are respectively the upper and lower bounds of the truncated prior distribution $f_{prior}\left(b_{j,k}\right)$. The R code for implementing the Pólya–Gamma Gibbs sampling algorithm is given in the supplementary material, http://intlpress.com/site/pub/files/_supp/sii/2022/0015/0004/SII-2022-0015-0004-s001.pdf.

4. SIMULATION STUDIES

4.1 Simulation 1

This simulation study was conducted to evaluate the recovery performance of the combined MCMC sampling algorithm based on different simulation conditions.

Simulation designs

The following manipulated conditions were considered: (a) three-point, four-point, and five-point Likert-type response scales, i.e., $K = 3, 4, 5$; (b) test length $J = 20, 30, 40$; (c) number of examinees $N = 500, 1000, 2000$. Fully crossing the different levels of these three factors yields 27 conditions (3 response scales × 3 test lengths × 3 sample sizes). For the GRMs, true values of item discrimination parameters a_j were generated from a uniform distribution [29], i.e., $a_j \sim \mathrm{Uniform}\,(0.5, 1.5)$, $j = 1, 2, \ldots, J$. It is known that the values of threshold parameters usually lie between -3 and 3 ([4], [5], [29]), i.e., (i) when $K = 3$, $b_{j1} \sim \mathrm{Uniform}\,(-3, 0)$ and $b_{j2} \sim \mathrm{Uniform}\,(0, 3)$, $j = 1, 2, \ldots, J$; (ii) when $K = 4$, $b_{j1} \sim \mathrm{Uniform}\,(-3, -1)$, $b_{j2} \sim \mathrm{Uniform}\,(-1, 1)$, and $b_{j3} \sim \mathrm{Uniform}\,(1, 3)$, $j = 1, 2, \ldots, J$; (iii) when $K = 5$, $b_{j1} \sim \mathrm{Uniform}\,(-3, -1.5)$, $b_{j2} \sim \mathrm{Uniform}\,(-1.5, 0)$, $b_{j3} \sim \mathrm{Uniform}\,(0, 1.5)$, and $b_{j4} \sim \mathrm{Uniform}\,(1.5, 3)$, $j = 1, 2, \ldots, J$. The true values of ability parameter were generated from a standard normal distribution. The polytomous item response data were generated from the GRM.

Prior distributions

The prior distributions of the GRM parameters for different response categories were set as follows. We used noninformative prior distributions for the item parameters, i.e., $a_j \sim N\left(0, 10^5\right) \mathrm{I}\,(0, +\infty)$; (i) $b_{j1} \sim N\left(0, 10^5\right) \mathrm{I}\,(-3, 3)$ and $b_{j2} \sim N\left(0, 10^5\right) \mathrm{I}\,(b_{j1}, 3)$; (ii) $b_{j1} \sim N\left(0, 10^5\right) \mathrm{I}\,(-3, 3)$, $b_{j2} \sim N\left(0, 10^5\right) \mathrm{I}\,(b_{j1}, 3)$, and $b_{j3} \sim N\left(0, 10^5\right) \mathrm{I}\,(b_{j2}, 3)$; (iii) $b_{j1} \sim N\left(0, 10^5\right) \mathrm{I}\,(-3, 3)$, $b_{j2} \sim N\left(0, 10^5\right) \mathrm{I}\,(b_{j1}, 3)$, $b_{j3} \sim N\left(0, 10^5\right) \mathrm{I}\,(b_{j2}, 3)$, and $b_{j4} \sim N\left(0, 10^5\right) \mathrm{I}\,(b_{j3}, 3)$. The prior of ability parameters was assumed to follow a standardized normal distribution. We considered 25 replications in each simulation condition. This choice was based on previous research in educational psychological assessments; for example, [45] used 10 replications for each simulation condition, and [48] used 30. In fact, in the present study, the accuracy of parameter estimation was guaranteed in 25 replications. If we were to consider too many replications, then it would be difficult to check the $\widehat{R}$ values (potential scale reduction factor (PSRF, [11]) calculated from each simulated dataset (replication) to ensure the parameter convergence. The work becomes huge when the simulated conditions increase.

Convergence diagnostics

To evaluate the convergence of parameter estimations, as an illustration, we consider only the case in which the response scale involves three-point Likert-type response data, the test length is fixed at 30, and there are 1000 examinees. We used two methods to check the convergence of our algorithm. One was the "eyeball" method of monitoring the convergence by visually inspecting the history plots of the generated sequences, and the other was the Gelman–Rubin method ([11], [18]) for checking the convergence of the parameters.

The convergence of the Bayesian algorithm was checked by monitoring the trace plots of the parameters for consecutive sequences of 20000 iterations. The first 10000 iterations

Table 1. Accuracy of parameters based on four different simulated conditions in simulation study 1

	3-Point Likert-Type Response Scale								
	Test Length 20			Test Length 30			Test Length 40		
Sample Size	500	1000	2000	500	1000	2000	500	1000	2000
Bias									
$\boldsymbol{a}$	−0.018	−0.026	−0.021	0.014	−0.012	−0.026	−0.004	−0.012	−0.014
$\boldsymbol{b}_{\bullet 1}$	−0.005	−0.022	0.009	−0.011	0.006	−0.003	−0.012	0.005	−0.006
$\boldsymbol{b}_{\bullet 2}$	−0.000	−0.025	0.002	−0.002	0.000	−0.002	0.003	0.013	−0.013
RMSE									
$\boldsymbol{a}$	0.144	0.100	0.077	0.137	0.094	0.077	0.144	0.104	0.070
$\boldsymbol{b}_{\bullet 1}$	0.118	0.089	0.063	0.118	0.083	0.054	0.130	0.089	0.070
$\boldsymbol{b}_{\bullet 2}$	0.122	0.094	0.063	0.130	0.089	0.063	0.141	0.089	0.077
SD									
$\boldsymbol{a}$	0.136	0.096	0.067	0.132	0.092	0.064	0.131	0.092	0.064
$\boldsymbol{b}_{\bullet 1}$	0.102	0.072	0.049	0.101	0.071	0.049	0.108	0.075	0.052
$\boldsymbol{b}_{\bullet 2}$	0.106	0.073	0.051	0.110	0.077	0.052	0.114	0.079	0.054
SE									
$\boldsymbol{a}$	0.140	0.094	0.069	0.132	0.093	0.068	0.140	0.101	0.066
$\boldsymbol{b}_{\bullet 1}$	0.110	0.084	0.064	0.113	0.081	0.058	0.119	0.086	0.067
$\boldsymbol{b}_{\bullet 2}$	0.114	0.086	0.064	0.124	0.085	0.064	0.135	0.088	0.073
CP									
$\boldsymbol{a}$	0.940	0.942	0.902	0.941	0.937	0.989	0.915	0.912	0.913
$\boldsymbol{b}_{\bullet 1}$	0.988	0.987	0.981	0.989	0.989	0.984	0.989	0.988	0.981
$\boldsymbol{b}_{\bullet 2}$	0.989	0.983	0.984	0.989	0.989	0.984	0.988	0.989	0.980

Note that the Bias, RMSE, SD, SE and CP denote the average Bias, RMSE, SD, SE and CP for the item parameters. $\boldsymbol{a}$ represents all discrimination parameters, $\boldsymbol{b}_{\bullet 1}$ represents all b_{j1} $(j = 1, \ldots, J.)$, and $\boldsymbol{b}_{\bullet 2}$ represents all b_{j2} $(j = 1, \ldots, J.)$.

were set as the burn-in period. As an illustration, four chains started at overdispersed starting values were run for each replication. The trace plots and posterior histograms for randomly selected item parameters are shown in the supplementary material, http://intlpress.com/site/pub/files/_supp/sii/2022/0015/0004/SII-2022-0015-0004-s001.pdf, where the posterior histograms are based on 10000 simulated values after a 10000 burn-in period. In addition, the PSRF values of all item parameters were less than 1.2.

Accuracy evaluation criteria

The accuracy of the parameter estimates was measured using four evaluation criteria, i.e., bias, root mean squared error (RMSE), standard deviation (SD), standard error (SE), and coverage probability of the 95% highest posterior density (HPD) intervals (CP). Let η be the parameter of interest. Assume that $M = 25$ 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 the parameter is defined as $\mathrm{Bias}(\eta) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\eta}^{(m)} - \eta\right)$. The RMSE for the parameter is defined as $\mathrm{RMSE}(\eta) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\eta}^{(m)} - \eta\right)^2}$. The simulation SE is the square root of the sample variance of the posterior estimates over different simulated data sets and is defined as

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

the average of the posterior standard deviation is defined as

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

and the coverage probability is defined as

$$\mathrm{CP}(\eta) = \frac{\#\text{ of }95\%\text{ HPD intervals containing }\eta\text{ in }M\text{ simulated data sets}}{M}.$$

Recovery of item parameters

The average bias, RMSE, SD, SE, and CP for item parameters based on 27 different simulation conditions are given in Tables 1, 2, and 3. The following conclusions can be drawn. 1) Given the total test length and response scale, when the number of individuals increases from 500 to 2000, the average RMSE, SD, and SE for discrimination and threshold parameters decrease. For example, for a total test length of 20 items and a fixed three-point Likert response scale, when the number of examinees increases from 500 to 2000, the average RMSE of all discrimination parameters decreases from 0.144 to 0.077, the average SD of all discrimination parameters decreases from 0.136 to 0.067, and the average SE of all discrimination parameters decreases from 0.140 to 0.069. For the threshold parameters, the average RMSE of all b_{j1} $(j = 1, \ldots, 20)$ parameters decreases from 0.118 to 0.063, the average SD of all b_{j1} $(j = 1, \ldots, 20)$ parameters decreases from 0.102 to 0.049, and the average SE of all b_{j1} $(j = 1, \ldots, 20)$ parameters decreases from 0.110 to

Table 2. Accuracy of parameters based on the different simulated conditions in simulation study 1

	4-Point Likert-Type Response Scale								
	Test Length 20			Test Length 30			Test Length 40		
Sample Size	500	1000	2000	500	1000	2000	500	1000	2000
Bias									
$\boldsymbol{a}$	−0.006	−0.018	−0.030	0.003	−0.019	−0.026	0.008	−0.016	−0.013
$\boldsymbol{b}_{\bullet 1}$	−0.040	0.009	0.019	0.018	−0.010	0.006	−0.020	−0.012	−0.004
$\boldsymbol{b}_{\bullet 2}$	−0.021	0.011	0.017	0.027	−0.004	0.005	0.003	−0.003	−0.003
$\boldsymbol{b}_{\bullet 3}$	0.005	0.012	0.008	0.049	0.000	−0.002	0.027	0.004	−0.001
RMSE									
$\boldsymbol{a}$	0.126	0.100	0.070	0.130	0.100	0.070	0.130	0.089	0.063
$\boldsymbol{b}_{\bullet 1}$	0.158	0.114	0.083	0.126	0.100	0.070	0.126	0.094	0.063
$\boldsymbol{b}_{\bullet 2}$	0.122	0.089	0.063	0.104	0.083	0.054	0.104	0.077	0.054
$\boldsymbol{b}_{\bullet 3}$	0.118	0.094	0.077	0.141	0.100	0.070	0.126	0.094	0.063
SD									
$\boldsymbol{a}$	0.119	0.084	0.059	0.117	0.082	0.057	0.117	0.082	0.058
$\boldsymbol{b}_{\bullet 1}$	0.131	0.090	0.061	0.114	0.082	0.056	0.110	0.078	0.053
$\boldsymbol{b}_{\bullet 2}$	0.098	0.069	0.047	0.093	0.066	0.045	0.093	0.066	0.045
$\boldsymbol{b}_{\bullet 3}$	0.106	0.078	0.054	0.120	0.083	0.058	0.111	0.078	0.053
SE									
$\boldsymbol{a}$	0.122	0.095	0.066	0.125	0.093	0.063	0.128	0.083	0.064
$\boldsymbol{b}_{\bullet 1}$	0.145	0.107	0.076	0.114	0.096	0.069	0.113	0.089	0.060
$\boldsymbol{b}_{\bullet 2}$	0.114	0.084	0.063	0.095	0.078	0.057	0.098	0.074	0.056
$\boldsymbol{b}_{\bullet 3}$	0.114	0.092	0.073	0.126	0.092	0.069	0.113	0.088	0.061
CP									
$\boldsymbol{a}$	0.926	0.987	0.985	0.982	0.989	0.989	0.922	0.925	0.989
$\boldsymbol{b}_{\bullet 1}$	0.987	0.986	0.981	0.984	0.986	0.982	0.926	0.988	0.987
$\boldsymbol{b}_{\bullet 2}$	0.985	0.986	0.978	0.985	0.986	0.982	0.936	0.988	0.984
$\boldsymbol{b}_{\bullet 3}$	0.989	0.986	0.980	0.983	0.988	0.985	0.928	0.988	0.986

Note that the Bias, RMSE, SD, SE and CP denote the average Bias, RMSE, SD, SE and CP for the item parameters. $\boldsymbol{a}$ represents all discrimination parameters, $\boldsymbol{b}_{\bullet 1}$ represents all $b_{j1}\,(j = 1, \ldots, J.)$, $\boldsymbol{b}_{\bullet 2}$ represents all $b_{j2}\,(j = 1, \ldots, J.)$ and $\boldsymbol{b}_{\bullet 3}$ represents all $b_{j3}\,(j = 1, \ldots, J.)$.

0.064. The average RMSE of all $b_{j2}\,(j = 1, \ldots, 20)$ parameters decreases from 0.122 to 0.063, the average SD of all $b_{j2}\,(j = 1, \ldots, 20)$ parameters decreases from 0.106 to 0.051, and the average SE of all $b_{j2}\,(j = 1, \ldots, 20)$ parameters decreases from 0.114 to 0.064. 2) Under the 27 simulated conditions, the average CPs of the discrimination and threshold parameters are about 0.970. 3) When the number of examinees is fixed at 500 (1000 or 2000), the Likert response scale is fixed at three points (four or five points), and the number of items is fixed at 20, the average RMSE, SD, and SE show that the recovery results of the discrimination and threshold parameters are close to those in the case that the total test length is 30 (40), which indicates that the Bayesian algorithm is stable and does not reduce the accuracy because of the increase in the number of items. In summary, the Pólya–Gamma Gibbs sampling algorithm provides accurate estimates of the item parameters in terms of various numbers of examinees and items.

Recovery of ability parameters

Next, we evaluate the recovery of the latent ability parameter using five accuracy evaluation criteria in Table 3. The following conclusions can be drawn from Table 4. 1) For 500 (1000 or 2000) examinees and the Likert response scale fixed at three points (four or five points), when the number of items increases from 20 to 40, the average RMSE, SD, and SE for ability parameters decrease. 2) Under the 27 simulated conditions, the average CPs of the ability parameters are about 0.950. In summary, it is shown again that the Pólya–Gamma Gibbs sampling algorithm is effective and the estimates of the parameters are accurate under various simulation conditions.

4.2 Simulation 2

In this simulation study, we compared the MH sampling algorithm and the Pólya–Gamma Gibbs sampling algorithm from two perspectives: accuracy and convergence.

Simulation designs

In this simulation, the number of examinees was $N = 1000$, a five-point Likert-type response scale was used, and the test length was $J = 40$. The true values of the item and person parameters in the GRM were the same as those in simulation study 1. We specified the following non-informative priors to the MH and Pólya–Gamma Gibbs sampling algorithms, i.e., $a_j \sim N\left(0, 10^5\right) \mathrm{I}\,(0, +\infty)$, $b_{j1} \sim N\left(0, 10^5\right) \mathrm{I}\,(-3, 3)$, $b_{j2} \sim N\left(0, 10^5\right) \mathrm{I}\,(b_{j1}, 3)$, $b_{j3} \sim N\left(0, 10^5\right) \mathrm{I}\,(b_{j2}, 3)$, and $b_{j4} \sim N\left(0, 10^5\right) \mathrm{I}\,(b_{j3}, 3)$, $j =$

Table 3. Accuracy of parameters based on the different simulated conditions in simulation study 1

	5-Point Likert-Type Response Scale								
	Test Length 20			Test Length 30			Test Length 40		
Sample Size	500	1000	2000	500	1000	2000	500	1000	2000
Bias									
$\boldsymbol{a}$	−0.004	−0.009	−0.021	0.021	−0.018	−0.027	−0.000	−0.009	−0.010
$\boldsymbol{b}_{\bullet 1}$	−0.013	−0.009	0.005	−0.038	−0.022	−0.011	−0.039	−0.013	−0.009
$\boldsymbol{b}_{\bullet 2}$	0.001	−0.001	−0.002	−0.021	−0.020	−0.013	−0.019	−0.012	−0.011
$\boldsymbol{b}_{\bullet 3}$	0.019	0.006	−0.006	0.002	−0.012	−0.016	0.006	0.000	−0.012
$\boldsymbol{b}_{\bullet 4}$	0.047	0.011	−0.010	0.021	−0.008	−0.017	0.038	0.011	−0.010
RMSE									
$\boldsymbol{a}$	0.118	0.083	0.063	0.122	0.083	0.063	0.122	0.077	0.054
$\boldsymbol{b}_{\bullet 1}$	0.141	0.100	0.077	0.144	0.104	0.070	0.137	0.094	0.070
$\boldsymbol{b}_{\bullet 2}$	0.109	0.077	0.063	0.109	0.083	0.054	0.104	0.077	0.054
$\boldsymbol{b}_{\bullet 3}$	0.114	0.077	0.063	0.109	0.089	0.063	0.114	0.083	0.054
$\boldsymbol{b}_{\bullet 4}$	0.144	0.100	0.077	0.137	0.104	0.077	0.141	0.104	0.070
SD									
$\boldsymbol{a}$	0.112	0.079	0.055	0.109	0.077	0.053	0.109	0.077	0.054
$\boldsymbol{b}_{\bullet 1}$	0.125	0.090	0.062	0.123	0.088	0.061	0.116	0.083	0.058
$\boldsymbol{b}_{\bullet 2}$	0.092	0.065	0.045	0.092	0.066	0.045	0.091	0.065	0.044
$\boldsymbol{b}_{\bullet 3}$	0.093	0.066	0.045	0.096	0.067	0.045	0.098	0.069	0.047
$\boldsymbol{b}_{\bullet 4}$	0.124	0.086	0.058	0.123	0.086	0.060	0.125	0.088	0.060
SE									
$\boldsymbol{a}$	0.114	0.083	0.059	0.116	0.083	0.055	0.118	0.076	0.055
$\boldsymbol{b}_{\bullet 1}$	0.131	0.096	0.074	0.129	0.097	0.071	0.119	0.087	0.065
$\boldsymbol{b}_{\bullet 2}$	0.102	0.071	0.058	0.102	0.078	0.057	0.094	0.074	0.055
$\boldsymbol{b}_{\bullet 3}$	0.103	0.073	0.059	0.102	0.082	0.057	0.105	0.080	0.056
$\boldsymbol{b}_{\bullet 4}$	0.125	0.094	0.071	0.125	0.098	0.071	0.126	0.097	0.066
CP									
$\boldsymbol{a}$	0.928	0.917	0.988	0.924	0.901	0.987	0.921	0.943	0.918
$\boldsymbol{b}_{\bullet 1}$	0.924	0.989	0.984	0.926	0.989	0.985	0.907	0.919	0.986
$\boldsymbol{b}_{\bullet 2}$	0.988	0.989	0.979	0.989	0.984	0.982	0.915	0.988	0.983
$\boldsymbol{b}_{\bullet 3}$	0.988	0.989	0.978	0.985	0.984	0.983	0.908	0.988	0.984
$\boldsymbol{b}_{\bullet 4}$	0.918	0.989	0.981	0.925	0.988	0.984	0.937	0.989	0.987

Note that the Bias, RMSE, SD, SE and CP denote the average Bias, RMSE, SD, SE and CP for the item parameters. $\boldsymbol{a}$ represents all discrimination parameters, $\boldsymbol{b}_{\bullet 1}$ represents all $b_{j1}\ (j = 1, \ldots, J.)$, $\boldsymbol{b}_{\bullet 2}$ represents all $b_{j2}\ (j = 1, \ldots, J.)$, $\boldsymbol{b}_{\bullet 3}$ represents all $b_{j3}\ (j = 1, \ldots, J.)$ and $\boldsymbol{b}_{\bullet 4}$ represents all $b_{j4}\ (j = 1, \ldots, J.)$.

$1, 2, \ldots, 40$, and the prior of ability parameters was assumed to follow a standardized normal distribution because of the model identification limitation, i.e., $\theta_i \sim N(0, 1)$, $i = 1, \ldots, 1000$. In fact, we know that an improper proposal distribution for the MH sampling algorithm can seriously reduce the acceptance probability of sampling, with most of the posterior samples being rejected. Therefore, low sampling efficiency is usually unavoidable, and the reduction in the number of valid samples may lead to incorrect inference results. In contrast, the Pólya–Gamma Gibbs sampling algorithm takes the acceptance probability as 1 to draw the samples from full conditional posterior distributions. The following proposal distributions for the discrimination, threshold, and ability parameters were considered in the process of implementing the MH sampling algorithm:

- Case 1: $a_j^* \sim N(a_j^{(r-1)}, 0.1)$, $b_{j,k}^* \sim N\big(b_{j,k}^{(r-1)}, 0.1\big) \times \mathrm{I}\big(b_{j,k-1}^* < b_{j,k}^* < b_{j,k+1}^{(r-1)}\big)$, $j = 1, \ldots, J.$ $k = 1, \ldots, K-1$, and $\theta_i^* \sim N(\theta_i^{(r-1)}, 0.1)$, $i = 1, \ldots, N$;
- Case 2: $a_j^* \sim N(a_j^{(r-1)}, 1)$, $b_{j,k}^* \sim N\big(b_{j,k}^{(r-1)}, 1\big)\mathrm{I}\big(b_{j,k-1}^* < b_{j,k}^* < b_{j,k+1}^{(r-1)}\big)$, $j = 1, \ldots, J.$ $k = 1, \ldots, K-1$, and $\theta_i^* \sim N(\theta_i^{(r-1)}, 1)$, $i = 1, \ldots, N$.

To compare the convergence of all parameters for the MH sampling algorithm with different proposal distributions and the Pólya–Gamma Gibbs sampling algorithm, the convergence of item and person parameters was evaluated by judging whether the PSRF values were less than 1.2. Figure 1 shows that the discrimination and threshold parameters had already converged by 2000 iterations for the Pólya–Gamma Gibbs sampling algorithm. For the MH sampling algorithm, some parameters had not converged after 5000 iterations with the proposal distributions with a variance of 0.1. The convergence with the proposal distributions with a variance of 1 was worse than that with those with a variance of 0.1, with some parameters not even having converged after 8000 iterations. Moreover, the bias and RMSE are used to evaluate the performances of the two algorithms in Table 5. It

Table 4. Accuracy of person parameters based on four different simulated conditions in simulation study 1

Category	No. of examinees	No. of items	Bias	RMSE	SD	SE	CP
3	500	20	0.004	0.313	0.295	0.279	0.943
		30	0.003	0.264	0.245	0.217	0.953
		40	0.000	0.225	0.210	0.192	0.963
	1000	20	−0.012	0.311	0.295	0.262	0.943
		30	0.008	0.262	0.248	0.214	0.964
		40	0.011	0.221	0.210	0.199	0.963
	2000	20	0.009	0.309	0.294	0.276	0.940
		30	0.000	0.258	0.248	0.238	0.954
		40	−0.004	0.221	0.210	0.197	0.940
4	500	20	−0.019	0.284	0.264	0.261	0.931
		30	0.015	0.236	0.223	0.242	0.963
		40	0.003	0.207	0.191	0.189	0.943
	1000	20	0.004	0.279	0.265	0.280	0.938
		30	−0.009	0.236	0.225	0.243	0.953
		40	−0.001	0.202	0.193	0.218	0.973
	2000	20	0.010	0.277	0.266	0.277	0.944
		30	−0.000	0.232	0.225	0.232	0.984
		40	−0.000	0.202	0.192	0.185	0.963
5	500	20	0.009	0.266	0.252	0.266	0.973
		30	−0.000	0.225	0.210	0.230	0.952
		40	0.008	0.197	0.181	0.195	0.953
	1000	20	0.002	0.266	0.253	0.241	0.943
		30	−0.010	0.223	0.214	0.201	0.973
		40	0.004	0.192	0.182	0.178	0.963
	2000	20	−0.003	0.262	0.254	0.259	0.944
		30	−0.007	0.223	0.215	0.198	0.954
		40	−0.001	0.189	0.182	0.175	0.953

Note that the Bias, RMSE, SD, SE and CP denote the average Bias, RMSE, SD, SE and CP for the ability parameters.

Table 5. Accuracy of parameter estimation using the two algorithms in the simulation study 2

	Pólya–Gamma Gibbs algorithm		MH algorithm under Case 1		MH algorithm under Case 2	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
$\boldsymbol{a}$	−0.009	0.077	0.103	0.170	0.185	0.219
$\boldsymbol{b}_{\bullet 1}$	−0.013	0.094	0.089	0.167	0.129	0.178
$\boldsymbol{b}_{\bullet 2}$	−0.012	0.077	0.027	0.094	0.027	0.070
$\boldsymbol{b}_{\bullet 3}$	0.000	0.083	−0.047	0.089	−0.061	0.094
$\boldsymbol{b}_{\bullet 4}$	0.011	0.104	−0.123	0.176	−0.173	0.214
$\boldsymbol{\theta}$	0.004	0.192	−0.016	0.254	−0.016	0.225

Note that the Bias and RMSE denote the average Bias and RMSE for the item parameters. $\boldsymbol{a}$ represents all discrimination parameters, $\boldsymbol{b}_{\bullet 1}$ represents all b_{j1} $(j = 1, \ldots, 40)$, $\boldsymbol{b}_{\bullet 2}$ represents all b_{j2} $(j = 1, \ldots, 40)$, $\boldsymbol{b}_{\bullet 3}$ represents all b_{j3} $(j = 1, \ldots, 40)$, $\boldsymbol{b}_{\bullet 4}$ represents all b_{j4} $(j = 1, \ldots, 40)$, and $\boldsymbol{\theta}$ represents all ability parameters.

has been shown that the selection of the proposal distribution has an important influence on the accuracy of parameter estimation, and the process of finding the proper tuning parameter is time consuming. In addition, we investigate the efficiency of the two algorithms from the perspective of the time consumed in implementing them. On a desktop computer (Intel(R) Xeon(R) E5-2695 V2 CPU) with a 2.4-GHz dual-core processor and 192 GB of RAM memory, the Pólya–Gamma Gibbs sampling algorithm and MH algorithm respectively consumed 3.8573 hours and 4.7456 hours when MCMC was run for 20 000 iterations for a replication experiment, where the MH algorithm was used to implement Case 1. In summary, the Pólya–Gamma Gibbs sampling algorithm is more effective than the MH algorithm in estimating model parameters.

5. EMPIRICAL EXAMPLE

To illustrate the applicability of the GRM in psychological assessment, we analyzed data from the Sexual Compulsivity Scale (SCS). Item response data were obtained from the Open Source Psychometrics Project (https://openpsychometrics.org), with the respondents constituting

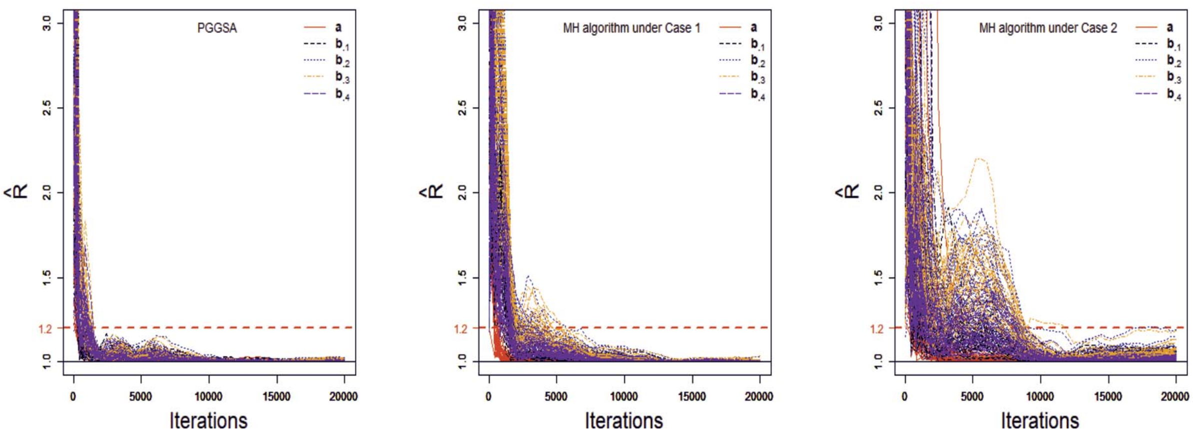

Figure 1. The trace plots of PSRF values for the simulation study 2.

Table 6. The sexual compulsivity scale items

Item	Question
1	My sexual appetite has gotten in the way of my relationships
2	My sexual thoughts and behaviors are causing problems in my life
3	My desires to have sex have disrupted my daily life
4	I sometimes fail to meet my commitments and responsibilities because of my sexual behaviors
5	I sometimes get so horny I could lose control
6	I find myself thinking about sex while at work
7	I feel that sexual thoughts and feelings are stronger than I am
8	I have to struggle to control my sexual thoughts and behavior
9	I think about sex more than I would like to
10	It has been difficult for me to find sex partners who desire having sex as much as I want to

a self-selected sample who had agreed to their responses being stored and made publicly available. All test items can be viewed on the aforementioned website, and the questions for all the test items are given in Table 6. The Sexual Compulsivity Scale [22] includes 10 items and was administered to $N = 3243$ respondents (with median age 33 and 55% male). All items were presented based on a four-point Likert-type response scale, with options 1 (‘Not at all like me’), 2 (‘Slightly like me’), 3 (‘Mainly like me’), and 4 (‘Very much like me’).

In the Bayesian computation, we used 20000 MCMC samples after a burn-in of 10000 iterations to compute all posterior estimates. The convergence of the chains was checked by PSRF, and we found that the PSRF values of all item parameters were less than 1.2.

Analysis of item parameters

The estimates of the item parameters are given in Table 7, from which we find that the expected a posteriori (EAP) estimations of the six item discrimination parameters are less than 2. This indicates that these items perform poorly at distinguishing among abilities compared with the other four items. The five items with the highest discrimination are items 8, 3, 7, 2, and 5 in turn. The EAP estimations of discrimination parameters for the five items are 2.714, 2.509, 2.379, 1.682, and 1.779, respectively. In addition, the EAP estimations of the threshold parameters (b_{j1}, b_{j2}, b_{j3}) are greatest for items 4, 7, and 2 in turn, which indicates that these items pose more difficulty than do the other seven items. This means that most respondents chose the first option (‘Not at all like me’) for these three items. By contrast, based on the EAP estimations of the threshold parameters, we find that most respondents chose the fourth option (‘Very much like me’) for items 6, 10, and 9. The SD range is 0.005–0.009 for the discrimination parameters and 0.000–0.005 for the threshold parameters.

Analysis of person parameters

The frequency histograms of the response scores and the posterior estimates of ability parameters for the 3243 respondents are shown in Figure 2. Most of the estimates of the ability parameters are near zero, and there are slightly fewer examinees with high ability (estimates between 0 and 2) than those with low ability (estimates between −2 and 0). The histogram of the EAPs of ability parameters is consistent with the frequency histogram of response scores, i.e., there are slightly more examinees with low response scores than those with high ones. Once again, it is verified that the estimation results are accurate.

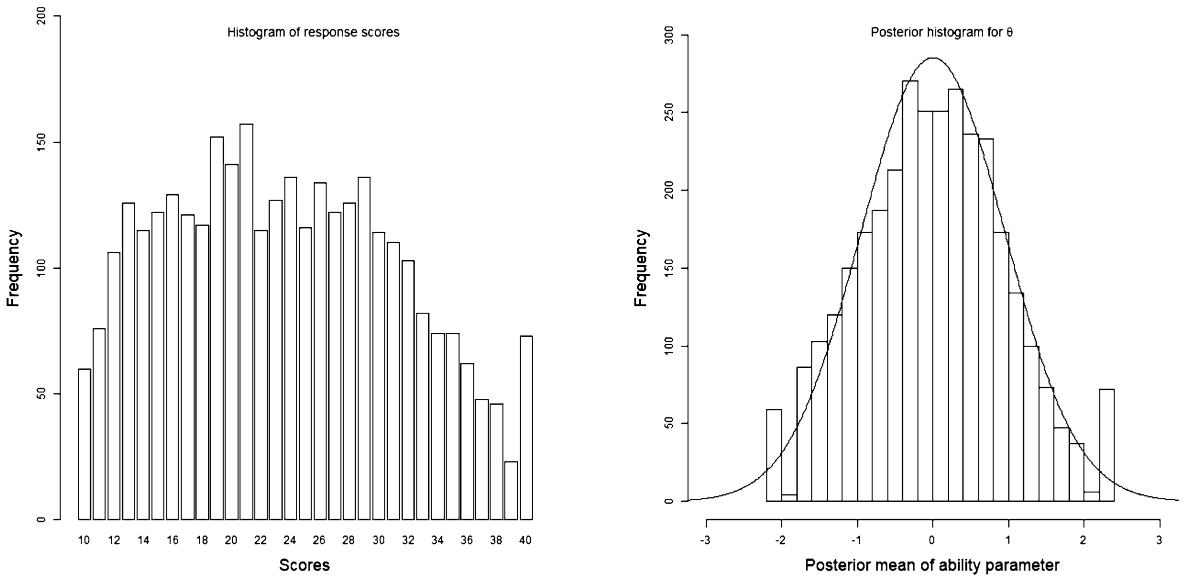

Figure 2. The frequency histograms of the response scores and the posterior estimates of ability parameters for 3243 respondents.

Table 7. The results of item parameter estimation in empirical example analysis

Parameter	EAP	SD	HPDI	Parameter	EAP	SD	HPDI
a_1	1.682	0.005	$[1.539, 1.824]$	$b_{1,1}$	-1.303	0.001	$[-1.379, -1.204]$
a_2	2.231	0.007	$[2.060, 2.410]$	$b_{2,1}$	-1.162	0.002	$[-1.225, -1.073]$
a_3	2.509	0.009	$[2.320, 2.700]$	$b_{3,1}$	-1.386	0.002	$[-1.470, -1.298]$
a_4	1.779	0.006	$[1.620, 1.936]$	$b_{4,1}$	-0.370	0.000	$[-0.420, -0.306]$
a_5	1.903	0.007	$[1.745, 2.070]$	$b_{5,1}$	-1.309	0.001	$[-1.391, -1.223]$
a_6	1.308	0.005	$[1.177, 1.448]$	$b_{6,1}$	-2.983	0.000	$[-2.999, -2.953]$
a_7	2.379	0.007	$[2.219, 2.549]$	$b_{7,1}$	-1.092	0.001	$[-1.169, -1.022]$
a_8	2.714	0.008	$[2.535, 2.888]$	$b_{8,1}$	-1.368	0.002	$[-1.464, -1.267]$
a_9	1.546	0.006	$[1.398, 1.696]$	$b_{9,1}$	-1.520	0.002	$[-1.610, -1.425]$
a_{10}	1.147	0.005	$[1.011, 1.280]$	$b_{10,1}$	-1.744	0.001	$[-1.858, -1.638]$
Parameter	EAP	SD	HPDI	Parameter	EAP	SD	HPDI
$b_{1,2}$	0.531	0.002	$[0.445, 0.606]$	$b_{1,3}$	2.021	0.002	$[1.943, 2.162]$
$b_{2,2}$	0.896	0.001	$[0.865, 1.009]$	$b_{2,3}$	2.851	0.004	$[2.718, 2.953]$
$b_{3,2}$	0.851	0.001	$[0.784, 1.000]$	$b_{3,3}$	2.769	0.005	$[2.656, 2.932]$
$b_{4,2}$	1.456	0.002	$[1.392, 1.570]$	$b_{4,3}$	2.919	0.003	$[2.845, 2.999]$
$b_{5,2}$	0.773	0.001	$[0.699, 0.841]$	$b_{5,3}$	2.209	0.004	$[2.101, 2.359]$
$b_{6,2}$	-1.175	0.001	$[-1.252, -1.094]$	$b_{6,3}$	0.301	0.001	$[0.224, 0.376]$
$b_{7,2}$	0.938	0.003	$[0.802, 0.924]$	$b_{7,3}$	2.888	0.002	$[2.790, 2.999]$
$b_{8,2}$	0.694	0.003	$[0.612, 0.819]$	$b_{8,3}$	2.664	0.003	$[2.554, 2.821]$
$b_{9,2}$	0.045	0.001	$[-0.021, 0.130]$	$b_{9,3}$	1.451	0.002	$[1.365, 1.534]$
$b_{10,2}$	-0.051	0.000	$[-0.107, -0.005]$	$b_{10,3}$	1.050	0.001	$[0.985, 1.134]$

Note: EAP is the expected a posteriori estimation, SD denotes the standard deviation, and HPDI denotes the 95% highest probability density interval.

6. CONCLUDING REMARKS

In this study, a novel and effective Pólya–Gamma Gibbs sampling algorithm based on auxiliary variables was used to estimate the parameters of the GRM. The Bayesian algorithm avoids the tedious multidimensional integral operation of marginal maximum likelihood estimation. Within a fully Bayesian framework, compared with the traditional Gibbs sampling algorithm and the MH sampling algorithm, the Pólya–Gamma Gibbs sampling algorithm (i) avoids the problem that the MH sampling algorithm relies heavily on the tuning parameters of the proposal distribution for different data sets and (ii) overcomes the disadvantage of the MH algorithm being sensitive to step size. It is known that the Gibbs sampling algorithm becomes ineffective for Bayesian non-conjugate models. By comparison, the Pólya–Gamma Gibbs sampling algorithm transforms a non-conjugate model into a conjugate one by using augmented auxiliary variables. With the help of the traditional Gibbs sampling algorithm, posterior sampling is easier to implement. Moreover, the Pólya–Gamma Gibbs sampling algorithm allows the use of informative and non-informative prior distributions, with satisfactory results being obtained even if an inappropriate prior distribution is used.

However, the computational burden of the Pólya–Gamma Gibbs sampling algorithm can be great, especially if many examinees, items, or missing data are considered or a large MCMC sample size is used. Therefore, would be desirable to develop a stand alone R package associated with C++ or Fortran software for a more-extensive large-scale assessment program.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 12001091), China Postdoctoral Science Foundations (Grant No. 2021M690587 and Grant No. 2021T140108), the Fundamental Research Funds for the Central Universities of China (Grant No. 2412020QD025), Yili Normal University 2021 Annual Research Project (Grant No. 2021YSBS012).

Received 25 May 2021

REFERENCES

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

[2] Albert, J. H., and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679. MR1224394

[3] Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, **2**, 581–594.

[4] Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann. MR1198885

[5] Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker. MR1198885

[6] 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

[7] Bjorner, J. B., Kosinski, M., and Ware, K. J. E., Jr. (2003). Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the Headache Impact Test (HIT$^{\mathrm{TM}}$). *Quality of Life Research*, **12**, 913–933.

[8] Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29–51.

[9] Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, **46**, 443–459. MR0668311

[10] Bolt, D. M., Hare, R. D., Vitale, J. E., and Newman, J. P. (2004). A multigroup item response theory analysis of the psychopathy checklist-revised. *Psychological Assessment*, **16**, 155–168.

[11] Brooks, S. P., and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455. MR1665662

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

[13] Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**, 327–335.

[14] Cowles, M. K. (1996). Accelerating Monte Carlo Markov Chain convergence for cumulative-link generalized linear models. Statistics and Computing, **6**, 101–111.

[15] Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer. MR2657265

[16] Fox, J.-P., and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, **66**, 269–286. MR1836937

[17] Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409. MR1141740

[18] Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

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

[20] Ghosh, M., Ghosh, A., Chen, M., and Agresti, A. (2000). Noninformative priors for one parameter item response models. *Journal of Statistical Planning and Inference*, **88**, 99–115. MR1767562

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

[22] Kalichman, S. C., and Rompa, D. (1995). Sexual sensation seeking and sexual compulsivity scales: validity, and predicting HIV risk behavior. *Journal of Personality Assessment*, **65**, 586–601.

[23] Kuo, T. C., and Sheng, Y. (2015). Bayesian estimation of a multi-unidimensional graded response IRT model. *Behaviormetrika*, **42**, 79–94.

[24] Lord, F. M. (1980). *Applications of item response theory to practical testing scores*. Reading: Addison-Wesley.

[25] Lu, J., Zhang, J. W., 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] Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149–174. MR0691827

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

[28] Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, **14**, 59–71.

[29] Natesan, P., Limbers, C., and Varni, J. W. (2010). Bayesian estimation of graded response multilevel models using Gibbs sampling: formulation and illustration. *Educational and Psychological Measurement*, **70**, 420–439.

[30] Patz, R. J., and Junker, B. W. (1999a). A straight forward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, **24**, 146–178.

[31] Patz, R. J., and 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.

[32] Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, **108**, 1339–1349. MR3174712

[33] Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation*, **72**, 217–232. MR1909259

[34] Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf. MR2272494

[35] Scherbaum, C. A., Cohen-Charash, Y., and Kern, M. J. (2006). Measuring general self-efficacy: a comparison of three measures using item response theory. *Educational and Psychological Measurement*, **66**, 1047–1063. MR2297383

[36] Scott, J. G., and Pillow, J. W. (2013). *Fully Bayesian inference for neural models with negative-binomial spiking*, pp. 1898–1906 in Advances in Neural Information Processing Systems 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Retrieved from http://papers.nips.cc/book/advances-in-neural-information-processing-systems-25-2012.

[37] Sheng, Y., and Headrick, T. C. (2012). A Gibbs sampler for the multidimensional item response model. *ISRN Applied Mathematics*, **2012**, 1–14. MR2929714

[38] Sheng, Y., and Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item Response theroy models. *Educational and Psychological Mesurement*, **67**, 899–919. MR2405522

[39] Sheng, Y., and Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, **68**, 413–430. MR2432233

[40] Sheng, Y., and Wikle, C. K. (2009). Bayesian IRT models in incorporating general and specific abilities. *Behaviormetrika*, **36**, 27–48. MR2649630

[41] Sorensen, D. A., Andersen, S., Gianola, D., and Korsgaard, I. (1995). Bayesian inference in threshold models using Gibbs sampling. *Genetics Selection Evolution Gse*, **27**, 229–249.

[42] Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–550. MR0898357

[43] Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, **43**, 39–55. MR1065199

[44] Tutz, G. (1991). Sequential models in categorical regression. *Computational Statistics and Data Analysis*, **11**, 275–295. MR1116203

[45] Wang, C., Fan, Z., Chang, H.-H., and Douglas, J. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, **38**, 381–417.

[46] Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory.* Monterey, CA: CTB/McGraw-Hill.

[47] Yao, L., and Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, **31**, 83–105. MR2339193

[48] Zhan, P., Jiao, H., and Liao, D. (2017). Cognitive diagnosis modelling incorporating item response times. *British Journal Mathematical Statistics Psychology*, **71**, 262–286.

[49] Zhang, J., Lu, J., and Tao, J. (2019). 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. *Statistics and Its Interface*, **12**, 35–48. MR3876408

[50] Zhang, J., Lu, J., Du, H., and Zhang, Z. (2020). Gibbs-slice sampling algorithm for estimating the four-parameter logistic model. *Frontiers in Psychology*, **11**, 2121.

[51] Zhu, X., and Stone, C. A. (2011). Assessing fit of unidimensional graded response models using Bayesian methods. *Journal of Educational Measurement*, **48**, 81–97.

Zhaoyuan Zhang
1 School of Mathematics and Statistics
Yili Normal University
Yining, 835000, Xinjiang, China
2 Institute of Applied Mathematics
Yili Normal University
Yining, 835000, Xinjiang, China
E-mail address: zhangzy328@nenu.edu.cn

Jiwei Zhang
School of Mathematics and Statistics
Yunnan University
Kunming, 650091, Yunnan, 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