frontiers in Psychology

ORIGINAL RESEARCH
published: 18 September 2020
doi: 10.3389/fpsyg.2020.02121

Gibbs-Slice Sampling Algorithm for Estimating the Four-Parameter Logistic Model

Jiwei Zhang[1], *Jing Lu*[2*], *Hang Du*[2,3] *and Zhaoyuan Zhang*[2,4]

[1] *Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, School of Mathematics and Statistics, Yunnan University, Kunming, China,* [2] *Key Laboratory of Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun, China,* [3] *School Affiliated to Longhua Institute of Educational Science, Shenzhen, China,* [4] *School of Mathematics and Statistics, Yili Normal University, Yili, China*

The four-parameter logistic (4PL) model has recently attracted much interest in educational testing and psychological measurement. This paper develops a new Gibbs-slice sampling algorithm for estimating the 4PL model parameters in a fully Bayesian framework. Here, the Gibbs algorithm is employed to improve the sampling efficiency by using the conjugate prior distributions in updating asymptote parameters. A slice sampling algorithm is used to update the 2PL model parameters, which overcomes the dependence of the Metropolis–Hastings algorithm on the proposal distribution (tuning parameters). In fact, the Gibbs-slice sampling algorithm not only improves the accuracy of parameter estimation, but also enhances sampling efficiency. Simulation studies are conducted to show the good performance of the proposed Gibbs-slice sampling algorithm and to investigate the impact of different choices of prior distribution on the accuracy of parameter estimation. Based on Markov chain Monte Carlo samples from the posterior distributions, the deviance information criterion and the logarithm of the pseudomarginal likelihood are considered to assess the model fittings. Moreover, a detailed analysis of PISA data is carried out to illustrate the proposed methodology.

Keywords: Bayesian inference, four-parameter logistic model, item response theory, model assessment, potential scale reduction factor, slice sampling algorithm

OPEN ACCESS

Edited by:
Taiki Takahashi,
Hokkaido University, Japan

Reviewed by:
Concha Bielza,
Polytechnic University of Madrid, Spain
Nazeer Muhammad,
COMSATS University, Pakistan

***Correspondence:**
Jing Lu
luj282@nenu.edu.cn

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *17 December 2019*
Accepted: *30 July 2020*
Published: *18 September 2020*

Citation:
Zhang J, Lu J, Du H and Zhang Z (2020) Gibbs-Slice Sampling Algorithm for Estimating the Four-Parameter Logistic Model. Front. Psychol. 11:2121. doi: 10.3389/fpsyg.2020.02121

1. INTRODUCTION

Over the past four decades, item response theory (IRT) models have been extensively used in educational testing and psychological measurement (Lord and Novick, 1968; Van der Linden and Hambleton, 1997; Embretson and Reise, 2000; Baker and Kim, 2004). These are latent variable modeling techniques, in which the response probability is used to construct the interaction between an individual's "ability" and item level stimuli (difficulty, guessing, etc.), where the focus is on the pattern of responses rather than on composite or total score variables and linear regression theory. Specifically, IRT attempts to model individual ability using question-level performance instead of aggregating test-level performance, and it focuses more on the information provided by an individual on each question. In social sciences, IRT has been applied to attachment (Fraley et al., 2000), personality (Ferrando, 1994; Steinberg and Thissen, 1995; Gray-Little et al., 1997; Rouse et al., 1999), psychopathology (Reise and Waller, 2003; Loken and Rulison, 2010; Waller and Reise, 2010; Waller and Feuerstahler, 2017), attention deficit hyperactivity disorder (Lanza et al., 2005), and delinquency (Osgood et al., 2002), among others.

To explore these applications, it is necessary to establish how the appropriate IRT models should be built and what valuable educational psychological phenomena can be examined to guide practice. In the field of dichotomous IRT models, the one-parameter logistic (1PL) model and the Rasch model (Rasch, 1960), as well as their extensions, the two-parameter logistic model (2PL) (Birnbaum, 1957) and the three-parameter logistic model (3PL) (Birnbaum, 1968), have attracted increasing attention in recent years because of their attractive mathematical properties. However, compared with the widely used 1PL, 2PL, and 3PL models, the four-parameter logistic (4PL) model has languished in obscurity for nearly 30 years (Barton and Lord, 1981), although its importance has gradually been realized by many researchers over the past decade (Hessen, 2005; Loken and Rulison, 2010; Waller and Reise, 2010; Green, 2011; Liao et al., 2012; Yen et al., 2012; Magis, 2013; Waller and Feuerstahler, 2017). This growing interest can be attributed to the need to deal with a number of problems encountered in educational psychology, which can be explained well and indeed solved using the 4PL model. For example, in computerized adaptive testing (CAT), high-ability examinees might on occasion miss items that they should be able to answer correctly, owing to a number of reasons, including anxiety, carelessness, unfamiliarity with the computer environment, distraction by poor testing conditions, or even misreading of the question (Hockemeyer, 2002; Rulison and Loken, 2009). Chang and Ying (2008) demonstrated that the ability determined using the traditional 2PL model is underestimated when the examinee mistakenly answers several items at the beginning of the CAT. In addition, Rulison and Loken (2009) found that using the 3PL model could severely penalize a high-ability examinee who makes a careless error on an easy item (Barton and Lord, 1981; Rulison and Loken, 2009). In psychopathology studies, researchers found that subjects with severe psychopathological disorders may be reluctant to self-report their true attitudes, behaviors, and experiences, so it is obviously inappropriate to use the traditional 3PL model with lower asymptotic parameter to explain such behaviors (Reise and Waller, 2003; Waller and Reise, 2010). Descriptions of the applications of the 4PL model in other areas can be found in Osgood et al. (2002) and Tavares et al. (2004). In addition to the development of the 4PL model in terms of its applications, its theoretical properties have been investigated in some depth. For example, Ogasawara (2012) discussed the asymptotic distribution of the ability, and Magis (2013) systematically studied the properties of the information function and proposed a method for determining its maximum point.

The main reason why the 4PL model has not been more widely used is that an upper asymptotic parameter is added to the 3PL model, which makes parameter estimation more difficult. However, with the rapid development of computer technology in recent years, the estimation problem for complex models has been solved. At the same time, the development of statistical software makes it easier for psychometricians to study complex models such as the 4PL model. Several researchers have used existing software to estimate the 4PL model. For example, Waller and Feuerstahler (2017) investigated 4PL model item and person parameter estimations using marginal maximum likelihood (MML) with the *mirt* (Chalmers, 2012) package, which uses MML via the expectation-maximization (EM) algorithm to estimate simple item response theory models. This is a different approach to that adopted here, where we use a Gibbs-slice sampling algorithm based on augmented data (auxiliary variables). Our Gibbs-slice sampling algorithm is in a fully Bayesian framework, and the posterior samples are drawn from the full conditional posterior distribution, whereas the MML–EM algorithm used in the *mirt* package is in a frequentist framework. Parameter estimates are obtained by an integral operation in the process of implementing the EM algorithm. Loken and Rulison (2010) used WinBUGS (Spiegelhalter et al., 2003) to estimate the 4PL model parameters in a Bayesian framework. However, convergence of parameter estimation is not completely achieved in the case of some non-informative prior distributions for WinBUGS. The reason for this may be that WinBUGS does not explicitly impose the monotonicity restriction $c < d$ on the 4PL model, i.e., it does not assume that the lower asymptote parameter c is smaller than the upper asymptote parameter d. (The introduction of parameters in the 4PL model will be described in section 2, and further discussion of these two parameters can be found in Culpepper, 2016 and Junker and Sijtsma, 2001). Thus, the prior Gibbs samplers do not strictly enforce an identification condition, and this leads to estimator non-convergence. More specifically, the prior distributions of the upper and lower asymptote parameters are given by the following informative priors (Loken and Rulison, 2010, p. 513):

$$c_j \sim N(0.22, 0.05), \qquad d_j \sim N(0.84, 0.05).$$

If we choose the non-informative prior distributions

$$c_j \sim N(0.22, 10^5), \qquad d_j \sim N(0.84, 10^5),$$

then, from the value ranges of the upper and lower asymptote parameters, we find that the lower asymptote parameter can be larger than the upper asymptote parameter, $d_j < c_j$, which violates the model identification condition $c_j < d_j$ (this condition will be introduced in detail in section 2). In this case, using WinBUGS to infer the model parameters may lead to biased estimates when the sample size (the number of examinees) is small and the prior distributions then play an important role. To solve the above problems in using WinBUGS, Loken and Rulison (2010) employed strong informative prior distributions to obtain good recovery (Culpepper, 2016, p. 1,143). However, Culpepper (2016, p. 1,161) pointed out that the use of informative prior distribution may lead to serious deviations if it happens to be centered at the wrong values. Therefore, he proposed that recovery should also be dealt with by using some non-informative priors.

In the present study, a novel and highly effective Gibbs-slice sampling algorithm in the Bayesian framework is proposed to estimate the 4PL model. The Gibbs-slice sampling algorithm overcomes the defects of WinBUGS that affect the convergence of parameter estimation based on the monotonicity restriction. Moreover, the algorithm can obtain good recovery results by using various types of prior distribution. In the following

sections, we will introduce the theoretical foundation of the slice sampling algorithm in detail, and we will then analyze the advantages of the slice sampling algorithm over two traditional Bayesian algorithms.

The rest of this paper is organized as follows. Section 2 contains a short introduction to the 4PL model, its reparameterized form, and model identification restrictions. In section 3, the theoretical foundation of the slice sampling algorithm is presented and its advantages compared with traditional Bayesian algorithms are analyzed. In section 4, three simulation studies focus respectively on the performance of parameter recovery, an analysis of the flexibility and sensitivity of different prior distributions for the slice sampling algorithm, and an assessment of model fittings using two Bayesian model selection criteria. In section 5, the quality of the Gibbs-slice sampling algorithm is investigated using an empirical example. We conclude the article with a brief discussion in section 6.

2. MODELS AND MODEL IDENTIFICATIONS

The 1PL and 2PL models have been widely used to fit binary item response data. Birnbaum (1968) modified the 2PL model to give the now well-known 3PL model, which includes a lower asymptote parameter to represent the contribution of guessing to the probability of correct response. To characterize the failure of high-ability examinees to answer easy items, Barton and Lord (1981) introduced an upper asymptote parameter into the 3PL model, giving the 4PL model:

$$\begin{aligned} P_{ij} &= P(y_{ij} = 1 \mid a_j, b_j, c_j, d_j, \theta_i) \\ &= c_j + (d_j - c_j)\frac{\exp[1.7a_j(\theta_i - b_j)]}{1 + \exp[1.7a_j(\theta_i - b_j)]} \end{aligned} \tag{1}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, J$, where N is the total number of examinees participating in the test and J is the test length. Here, y_{ij} is the binary response of the ith examinee with latent ability level θ_i to answer the jth item and is coded as 1 for a correct response and 0 for an incorrect response, P_{ij} is the corresponding probability of correct response, a_j is the item discrimination parameter, b_j is the item difficulty parameter, c_j is the item lower asymptote (pseudo-guessing) parameter, and d_j is the item upper asymptote parameter. The 4PL model reduces to the other models as special cases: $d_j = 1$ gives the 3PL model, $c_j = 0$ gives the 2PL model, and $a_j = 1$ gives the 1PL model. Following Culpepper (2016), we reparameterize the traditional 4PL model to construct a new 4PL model by defining a slipping parameter similar to that in cognitive diagnostic tests:

$$\begin{aligned} P_{ij} &= P(y_{ij} = 1 \mid a_j, b_j, c_j, \gamma_j, \theta_i) \\ &= c_j + (1 - \gamma_j - c_j)\frac{\exp[1.7a_j(\theta_i - b_j)]}{1 + \exp[1.7a_j(\theta_i - b_j)]}, \end{aligned} \tag{2}$$

where $\gamma_j = 1 - d_j$.

One identification restriction is that the upper asymptote must exceed the lower asymptote: $d_j > c_j$. Equivalently, the restriction $0 < c_j + \gamma_j < 1$ must be satisfied for the reparameterized 4PL model, Meanwhile, either the scale of latent abilities or the scale of item parameters must be restricted to identify the two0parameter IRT models. Three methods are widely used to identify two-parameter IRT models.

1. Fx the mean population level of ability to zero and the variance population level of ability to one (Lord and Novick, 1968; Bock and Aitkin, 1981; Fox and Glas, 2001; Fox, 2010), i.e., $\theta \sim N(0, 1)$.
2. Restrict the sum of item difficulty parameters to zero and the product of item discrimination parameters to one (Fox, 2001; Fox, 2005, 2010), i.e., $\sum_{j=1}^{J} b_j = 0$ and $\prod_{j=1}^{J} a_j = 1$.
3. Fix the item difficulty parameter at a specific value, most often zero, and restrict the discrimination parameter to a specific value, most often one (Fox, 2001; Fox, 2010), i.e., $b_1 = 0$ and $a_1 = 1$. The basic idea here is to identify the two-parameter logistic model by anchoring an item discrimination parameter to an arbitrary constant, typically $a_1 = 1$, for a given item. Meantime, a location identification constraint is imposed by restricting a difficulty parameter, typically $b_1 = 0$, for a given item. Based on the fixed anchoring values of the item parameters, other parameters are estimated on the same scale. The estimated difficulty or discrimination values of item parameters are interpreted based on their positions relative to the corresponding anchoring values. For details, see Fox (2010, p. 87).

In the present study, the main aim is to evaluate the accuracy of parameter estimation obtained by the slice sampling algorithm for different types of prior distributions. Therefore, the first of the above methods is used to eliminate the trade-offs between ability θ and the difficulty parameter b in location, and between ability θ (difficulty parameter b) and the discrimination parameter a in scale.

3. THEORETICAL FOUNDATION AND ANALYSIS OF THE ADVANTAGES OF THE SLICE SAMPLING ALGORITHM

3.1. Theoretical Foundation of the Slice Sampling Algorithm

The motivation for the slice sampling algorithm (Damien et al., 1999; Neal, 2003; Bishop, 2006; Lu et al., 2018) is that we can use the auxiliary variable approach to sample from posterior distributions arising from Bayesian non-conjugate models. The theoretical basis for this algorithm is as follows.

Suppose that the simulated values are generated from a target density function $t(x)$ given by $t(x) \propto \phi(x)\prod_{i=1}^{N} l_i(x)$ that cannot be sampled directly, where $\phi(x)$ is a known density from which samples can be easily drawn and $l_i(x)$ are non-negative invertible functions, which do not have to be density functions. We introduce the auxiliary variables represented by the vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)'$, each element of which is from $(0, +\infty)$ and where $\delta_1, \ldots, \delta_N$ are mutually independent. The inequalities $\delta_i < l_i(x)$ are established, and the joint density can be written as

$$t(x, \delta_1, \ldots, \delta_N) \propto \phi(x)\prod_{i=1}^{N} \mathrm{I}\{\delta_i < l_i(x)\}, \tag{3}$$

where the indicator function $\mathrm{I}(A)$ takes the value 1 if A is true and the value 0 if A is false. If the auxiliary variables are integrated out, the marginal distribution $t(x)$ is obtained as

$$t(x) = \int_0^{l_1(x)} \cdots \int_0^{l_N(x)} t(x, \delta_1, \ldots, \delta_N)\, \mathrm{d}\delta_N \cdots \mathrm{d}\delta_1,$$
$$\propto \phi(x) \int_0^{l_1(x)} \cdots \int_0^{l_N(x)} 1\, \mathrm{d}\delta_N \cdots \mathrm{d}\delta_1 = \phi(x) \prod_{i=1}^{N} l_i(x). \quad (4)$$

Using the invertibility of the function $l_i(x)$, we can then obtain the set $\Lambda_{\delta_i} = \{x \mid \delta_i < l_i(x)\}$. The simulated values are generated from the Gibbs sampler based on the auxiliary variables by repeatedly sampling from the full conditional distributions, proceeding as follows at iteration r:

- Sample $\delta_i^{(r)} \sim \mathrm{Uniform}\left(0, l_i(x^{(r-1)})\right), i = 1, \ldots, N.$
- Sample $x^{(r)} \sim \Lambda_{\delta_i} = \{x \mid \delta_i^{(r)} < l_i(x)\}.$

We thereby derive a horizontal "slice" under the density function. Thus, a Markov chain based on the new Gibbs sampler can be constructed by sampling points alternately from the uniform distribution under the density curve and only concerning the horizontal "slice" defined by the current sample points.

3.2. Advantages of the Slice Sampling Algorithm Compared With the Metropolis–Hastings Algorithm

In the Bayesian framework, we first consider the benefits of the slice sampling algorithm compared with the traditional Metropolis–Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970; Tierney, 1994; Chib and Greenberg, 1995; Chen et al., 2000). It is known that the MH algorithm relies heavily on the tuning parameters of the proposal distribution for different data sets. In addition, the MH algorithm is sensitive to step size. If the step size is too small, the chain will take longer to traverse the target density. If the step size is too large, there will be inefficiencies due to high rejection rate. More specifically, researchers should ensure that each parameter candidate is no more than 50% accepted by adjusting the tuning parameters of the MH algorithm. Further, for example, when we draw two-dimensional item parameters at the same time in the 2PL model, the probability of acceptance will be reduced to around 25% (Patz and Junker, 1999, p. 163). Thus, the sampling efficiency of the MH algorithm is greatly reduced. However, the slice sampling algorithm avoids the retrospective tuning that is needed in the MH algorithm if we do not know how to choose a proper tuning parameter or if no value for the tuning parameter is appropriate. It always keeps the drawn samples accepted, thus increasing the sampling efficiency. Next, we show that the slice sampling algorithm is more efficient than a particular independent MH chain.

Let us use the MH algorithm to obtain samples from the posterior distribution $t(x)$ given by $t(x) \propto \phi(x)l(x)$, where $\phi(x)$ is selected as a special proposal distribution. Let x^* be a candidate value from the proposal distribution $\phi(x)$ and let $x^{(r)}$ be the current point. The probability of the new candidate being accepted, $\min\{1, l(x^*)/l(x^{(r)})\}$, is determined by a random number u from $\mathrm{Uniform}(0, 1)$. Essentially, if $u < l(x^*)/l(x^{(r)})$, then $x^{(r+1)} = x^*$; otherwise, $x^{(r+1)} = x^{(r)}$. The process is to draw the candidate first and then determine whether or not to "move" or "stay" by using the random number u. The "stay" process will lead to a reduction in the sampling efficiency of the MH algorithm. By contrast, suppose we consider the inverse process of the above sampling to draw the random number u first. To achieve the purpose of moving, we need to draw the candidate x^* from $\phi(x)$ such that $u < l(x^*)/l(x^{(r)})$. Therefore, x^* can be regarded as a sample from $\phi(x)$ restricted to the set $\Theta_u(r) = \{x \mid l(x) > ul(x^{(r)})\}$. In this case, the chain will always be moved, thus improving the sampling efficiency.

In addition, with the MH algorithm, it is relatively difficult to sample parameters with monotonicity or truncated interval restrictions. Instead, it is possible to improve the accuracy of parameter estimation by employing strong informative prior distributions to avoid violating the restriction conditions (Culpepper, 2016). For example, the prior distributions of the lower asymptote and upper asymptote parameters used in Loken and Rulison (2010) are, respectively Beta(5, 17) and Beta(17, 5), and these two parameters are fairly concentrated in the range of 0.227–0.773. However, the advantage of the slice sampling algorithm is that it can easily draw the posterior samples from any prior distribution as long as these distributions have a reasonable value range of parameters. See the following sections for details.

3.3. Advantages of the Slice Sampling Algorithm Compared With the Gibbs Algorithm

The idea of the slice sampling algorithm is to draw the posterior samples from a truncated prior distribution by introducing auxiliary variables, where the truncated interval is deduced from the likelihood function. This differs from the approach of the Gibbs algorithm (Geman and Geman, 1984; Gelfand and Smith, 1990), which is to generate posterior samples by sweeping through each variable to sample from its conditional distribution, with the remaining variables fixed at their current values. However, slice sampling algorithm can be conceived of as extensions of the Gibbs algorithm. In particular, when the parameters in which we are interested are represented by a multidimensional vector $\boldsymbol{X}$, we cannot use the slice sampling algorithm directly to obtain the multivariate set $\boldsymbol{\Theta}_u = (\Theta_u^1, \ldots, \Theta_u^k, \ldots, \Theta_u^p)$, where p is the dimension of $\boldsymbol{X}$. Therefore, a Gibbs sampler is employed to draw the samples from the full conditional distribution $l(x_k \mid \boldsymbol{x}_{(-k)}, u)$ for $k = 1, \ldots, p$, which is a realization of $t(\boldsymbol{X})$. This involves sampling from $\phi(x_k \mid \boldsymbol{x}_{(-k)})$ restricted to the set $\Theta_u^k = \{x_k \mid l(x_k, \boldsymbol{x}_{(-k)}) > u\}$, where the premise must be satisfied that $l(x_k, \boldsymbol{x}_{(-k)})$ is invertible for all k given $\boldsymbol{x}_{(-k)}$.

It is well-known that the Gibbs algorithm can quickly and effectively draw samples from the posterior distribution owing to the fact that the full conditional posterior distribution is

easy to sample using the conjugate prior distribution. However, the Gibbs algorithm is not valid for Bayesian non-conjugate models such as the 2PL model. By comparison, the slice sampling algorithm for estimating the 2PL model has the advantage of a flexible prior distribution being introduced to obtain samples from the full conditional posterior distributions rather than being restricted to using the conjugate distributions, which is required in Gibbs sampling and is limited to the use of the normal ogive framework (Tanner and Wong, 1987; Albert, 1992; Béguin and Glas, 2001; Fox and Glas, 2001; Fox, 2010; Culpepper, 2016). The slice sampling algorithm allows the use of informative prior distributions and non-informative prior distributions, and even if an inappropriate prior distribution is adopted, it can still obtain satisfactory results. That is, any prior distribution can be used as long as the values sampled from it are in a reasonable range of the parameter support set. For example, for the discrimination parameter, the following prior distributions can be considered: the informative prior $\log N(0,1)$, the non-informative priors $N(0,1000)\mathrm{I}(a > 0)$, and the inappropriate priors Exp(1) and Gamma(2, 3).

4. BAYESIAN INFERENCE

4.1. Bayesian Estimation

In the present study, an efficient Gibbs-slice sampling algorithm in a fully Bayesian framework is used to estimate the following 4PL model. The sampling process of Gibbs-slice sampling algorithm consists of two parts. One part is the Gibbs sampling algorithm, which is used to update the guessing and slipping parameters from the truncated Beta distributions by introducing auxiliary variables (Béguin and Glas, 2001; Fox, 2010; Culpepper, 2016). The efficiency of Gibbs sampling is greatly improved by the use of conjugate prior distributions (Tanner and Wong, 1987; Albert, 1992). The other part is the slice sampling algorithm, which samples the 2PL model from the truncated full conditional posterior distributions by introducing different auxiliary variables.

Next, the specific sampling process of the Gibbs-slice sampling algorithm is described.

Gibbs Steps

First, following Béguin and Glas (2001), we introduce an auxiliary variable η_{ij}, where $\eta_{ij} = 1$ indicates that examinee i has the ability to answer item j correctly and $\eta_{ij} = 0$ otherwise. The purpose of introducing this auxiliary variable is to separate the guessing and slipping parameters from the 4PL model and make it easier to implement Gibbs sampling for the guessing and slipping parameters through the conjugate Beta distributions. Letting $\Delta = (\theta_i, a_j, b_j, c_j, \gamma_j)$, we can obtain the full conditional distribution of η_{ij} based on Bayes' theorem:

$$\begin{aligned}
P(\eta_{ij} = 1 \mid y_{ij} = 1, \Delta) &= \frac{P(\eta_{ij} = 1, y_{ij} = 1, \Delta)}{P(y_{ij} = 1 \mid \Delta)} \\
&= \frac{(1-\gamma_j)P_{ij}^*}{c_j + (1-\gamma_j - c_j)P_{ij}^*}, \\
P(\eta_{ij} = 0 \mid y_{ij} = 1, \Delta) &= \frac{P(\eta_{ij} = 0, y_{ij} = 1, \Delta)}{P(y_{ij} = 1 \mid \Delta)} \\
&= \frac{c_j(1-P_{ij}^*)}{c_j + (1-\gamma_j - c_j)P_{ij}^*}, \\
P(\eta_{ij} = 1 \mid y_{ij} = 0, \Delta) &= \frac{P(\eta_{ij} = 1, y_{ij} = 0, \Delta)}{P(y_{ij} = 0 \mid \Delta)} \\
&= \frac{\gamma_j P_{ij}^*}{1 - c_j - (1-\gamma_j - c_j)P_{ij}^*}, \\
P(\eta_{ij} = 0 \mid y_{ij} = 0, \Delta) &= \frac{P(\eta_{ij} = 0, y_{ij} = 0, \Delta)}{P(y_{ij} = 0 \mid \Delta)} \\
&= \frac{(1-c_j)(1-P_{ij}^*)}{1 - c_j - (1-\gamma_j - c_j)P_{ij}^*}.
\end{aligned} \tag{5}$$

where

$$P_{ij}^* = \frac{\exp[1.7a_j(\theta_i - b_j)]}{1 + \exp[1.7a_j(\theta_i - b_j)]}.$$

The priors of the guessing and slipping parameters follow the Beta distributions, i.e., $c_j \sim \mathrm{Beta}(\nu_0, u_0), \gamma_j \sim \mathrm{Beta}(\nu_1, u_1)$. However, the guessing and slipping parameters themselves satisfy the following truncated restrictions owing to model identification (Junker and Sijtsma, 2001; Culpepper, 2016):

$$\Xi = \{(c_j, \gamma_j) \mid 0 \le c_j < 1, 0 \le \gamma_j < 1, 0 \le c_j < 1 - \gamma_j\}. \tag{6}$$

The joint posterior distribution of the guessing and slipping parameters can be written as

$$\begin{aligned}
p(c_j, \gamma_j \mid \boldsymbol{y}_j, \boldsymbol{\eta}_j) \propto \prod_{i=1}^{N} \left[(1-\gamma_j)^{\eta_{ij}} c_j^{(1-\eta_{ij})}\right]^{y_{ij}} \left[\gamma_j^{\eta_{ij}}(1-c_j)^{(1-\eta_{ij})}\right]^{(1-y_{ij})} \\
p(c_j, \gamma_j)\mathrm{I}\big((c_j, \gamma_j) \in \Xi\big) \propto c_j^{\widehat{\kappa}_{00}+\nu_0-1}(1-c_j)^{\widehat{\kappa}_{01}+u_0-1} \\
\gamma_j^{\widehat{\kappa}_{10}+\nu_1-1}(1-\gamma_j)^{\widehat{\kappa}_{11}+u_1-1}\mathrm{I}\big((c_j, \gamma_j) \in \Xi\big).
\end{aligned} \tag{7}$$

Let $\boldsymbol{y}_j' = (y_{1j}, \ldots, y_{Nj})$, $\boldsymbol{\eta}_j' = (\eta_{1j}, \ldots, \eta_{Nj})$, and

$$\begin{aligned}
\widehat{\kappa}_{00} &= (\mathbf{1}_N - \boldsymbol{\eta}_j)'\boldsymbol{y}_j, & \widehat{\kappa}_{01} &= (\mathbf{1}_N - \boldsymbol{\eta}_j)'(\mathbf{1}_N - \boldsymbol{y}_j), \\
\widehat{\kappa}_{10} &= \boldsymbol{\eta}_j'(\mathbf{1}_N - \boldsymbol{y}_j), & \widehat{\kappa}_{11} &= \boldsymbol{\eta}_j'\boldsymbol{y}_j.
\end{aligned}$$

The full conditional posterior distributions of $\big(c_j, \gamma_j\big)$ can be written as

$$\begin{aligned}
c_j^{(r)} \mid \gamma_j^{(r-1)} &\sim \mathrm{Beta}(\widehat{\kappa}_{00} + \nu_0, \widehat{\kappa}_{01} + u_0)\mathrm{I}(0 \le c_j^{(r)} < 1 - \gamma_j^{(r-1)}), \\
\gamma_j^{(r)} \mid c_j^{(r)} &\sim \mathrm{Beta}(\widehat{\kappa}_{10} + \nu_1, \widehat{\kappa}_{11} + u_1)\mathrm{I}(0 \le \gamma_j^{(r)} < 1 - c_j^{(r)}).
\end{aligned} \tag{8}$$

Slice Steps

Supposing that the guessing and slipping parameters have been updated by the Gibbs algorithm, we update the parameters in the 2PL model using the slice sampling algorithm. Two

additional independent auxiliary variables λ_{ij} and φ_{ij}, defined on the intervals

$$\left(0, \frac{P_{ij}^{(r)} - c_j^{(r)}}{1 - \gamma_j^{(r)} - c_j^{(r)}}\right) \quad \text{and} \quad \left(0, \frac{1 - \gamma_j^{(r)} - P_{ij}^{(r)}}{1 - \gamma_j^{(r)} - c_j^{(r)}}\right),$$

are introduced to facilitate sampling, where r is the number of iterations. In fact, $(P_{ij} - c_j)/(1 - \gamma_j - c_j)$ is the correct response probability of the 2PL model, while $(1 - \gamma_j - P_{ij})/(1 - \gamma_j - c_j)$ is the corresponding incorrect response probability. Therefore, the joint likelihood of a, b, c, γ, θ based on the auxiliary variables λ and φ can be written as

$$\begin{aligned} &p(\boldsymbol{y} \mid \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\varphi}) \\ &\quad \propto \prod_{i=1}^{N} \prod_{j=1}^{J} \left[\mathrm{I}(y_{ij} = 1) \mathrm{I}\left(0 < \lambda_{ij} \leq \frac{P_{ij} - c_j}{1 - \gamma_j - c_j}\right) \right. \\ &\quad \left. + \mathrm{I}(y_{ij} = 0) \mathrm{I}\left(0 < \varphi_{ij} \leq \frac{1 - \gamma_j - P_{ij}}{1 - \gamma_j - c_j}\right) \right]. \end{aligned} \tag{9}$$

Equivalently,

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\varphi}) \propto \prod_{i=1}^{N} \prod_{j=1}^{J} &\mathrm{I}(y_{ij} = 1) \mathrm{I}(0 < \lambda_{ij} \leq P_{ij}^*) \\ &+ \mathrm{I}(y_{ij} = 0) \mathrm{I}(0 < \varphi_{ij} \leq Q_{ij}^*), \end{aligned} \tag{10}$$

where

$$P_{ij}^* = 1 - Q_{ij}^* = \frac{\exp[1.7a_j(\theta_i - b_j)]}{1 + \exp[1.7a_j(\theta_i - b_j)]} = \frac{P_{ij} - c_j}{1 - \gamma_j - c_j},$$
$$Q_{ij}^* = \frac{1}{1 + \exp[1.7a_j(\theta_i - b_j)]} = \frac{1 - \gamma_j - P_{ij}}{1 - \gamma_j - c_j}.$$

Integrating out the two random variables λ and φ in (10), the joint likelihood based on responses can be obtained:

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\varphi}) &\propto \prod_{i=1}^{N} \prod_{j=1}^{J} \mathrm{I}(y_{ij} = 1) E_{\lambda}[\mathrm{I}(0 < \lambda_{ij} \leq P_{ij}^*)] \\ &\quad + \mathrm{I}(y_{ij} = 0) E_{\varphi}[\mathrm{I}(0 < \varphi_{ij} \leq Q_{ij}^*)] \\ &\propto \prod_{i=1}^{N} \prod_{j=1}^{J} (P_{ij}^*)^{(y_{ij}=1)} (Q_{ij}^*)^{(y_{ij}=0)}, \end{aligned} \tag{11}$$

where E_λ is an expectation operation for the random variable λ. We know that $\boldsymbol{\eta}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\varphi}$ are independent of each other. Therefore, the joint posterior distribution based on the auxiliary variables can be written as

$$\begin{aligned} p(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\varphi} \mid \boldsymbol{y}) \propto &\, p(\boldsymbol{\eta} \mid \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{y}) p(\boldsymbol{\lambda}, \boldsymbol{\varphi} \mid \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{y}) \\ &\times p(\boldsymbol{\theta}) p(\boldsymbol{a}) p(\boldsymbol{b}) p(\boldsymbol{c}, \boldsymbol{\gamma}) \mathrm{I}\big((\boldsymbol{c}, \boldsymbol{\gamma}) \in \Xi\big). \end{aligned} \tag{12}$$

The specific form can be represented as

$$\begin{aligned} p(\boldsymbol{\eta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\varphi} \mid \boldsymbol{y}) \propto &\prod_{i=1}^{N} \prod_{j=1}^{J} \left[(1 - \gamma_j)^{\eta_{ij}} c_j^{(1-\eta_{ij})} \right]^{y_{ij}} \\ &\left[\gamma_j^{\eta_{ij}} (1 - c_j)^{(1-\eta_{ij})} \right]^{(1-y_{ij})} \\ &\times \left[\mathrm{I}(y_{ij} = 1) \mathrm{I}(0 < \lambda_{ij} \leq P_{ij}^*) \right. \\ &\left. + \mathrm{I}(y_{ij} = 0) \mathrm{I}(0 < \varphi_{ij} \leq Q_{ij}^*) \right] \\ &\times \prod_{j=1}^{J} p(a_j) p(b_j) p(c_j, \gamma_j) \mathrm{I}\big((c_j, \gamma_j) \in \Xi\big) \\ &\prod_{i=1}^{N} p(\theta_i). \end{aligned} \tag{13}$$

The detailed slice sampling algorithm is given below.

First, we update the auxiliary variables λ_{ij} and φ_{ij} when given θ_i, a_j, b_j, c_j, γ_j, and y_{ij}. According to (13), the auxiliary variables λ_{ij} and φ_{ij} have the following interval constraints:

$$0 < \lambda_{ij} \leq P_{ij}^* = \frac{P_{ij} - c_j}{1 - \gamma_j - c_j} \qquad \text{when} \quad y_{ij} = 1,$$
$$0 < \varphi_{ij} \leq Q_{ij}^* = \frac{1 - \gamma_j - P_{ij}}{1 - \gamma_j - c_j} \qquad \text{when} \quad y_{ij} = 0.$$

Therefore, the full conditional posterior distributions of λ_{ij} and φ_{ij} can be written as

$$\lambda_{ij} \mid \theta_i, a_j, b_j, c_j, \gamma_j, y_{ij} \sim \mathrm{Uniform}\left(0, \frac{P_{ij} - c_j}{1 - \gamma_j - c_j}\right) \quad \text{when} \quad y_{ij} = 1, \tag{14}$$

$$\varphi_{ij} \mid \theta_i, a_j, b_j, c_j, \gamma_j, y_{ij} \sim \mathrm{Uniform}\left(0, \frac{1 - \gamma_j - P_{ij}}{1 - \gamma_j - c_j}\right) \quad \text{when} \quad y_{ij} = 0. \tag{15}$$

Next, we update the difficulty parameter b_j. The prior of the difficulty parameter is assumed to follow a normal distribution with mean μ_b and variance σ_b^2. According to (10), $\forall i$, when $y_{ij} = 1$, we have $0 < \lambda_{ij} \leq P_{ij}^*$, and the following inequality can be established:

$$a_j(\theta_i - b_j) \geq \frac{1}{1.7} \log\left(\frac{\lambda_{ij}}{1 - \lambda_{ij}}\right), \text{ or equivalently}$$
$$b_j \leq \theta_i - \frac{1}{1.7a_j} \log\left(\frac{\lambda_{ij}}{1 - \lambda_{ij}}\right).$$

In fact, this inequality is obtained through the following calculation process:

$$0 < \lambda_{ij} \leq P_{ij}^*, \text{ or equivalently } 0 < \lambda_{ij} \leq \frac{\exp[1.7a_j(\theta_i - b_j)]}{1 + \exp[1.7a_j(\theta_i - b_j)]},$$

from which

$$\lambda_{ij} + \lambda_{ij} \exp[1.7a_j(\theta_i - b_j)] \leq \exp[1.7a_j(\theta_i - b_j)], \text{ or equivalently}$$
$$\frac{\lambda_{ij}}{1 - \lambda_{ij}} \leq \exp[1.7a_j(\theta_i - b_j)].$$

Therefore, we have

$$\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right) \leq [1.7a_j(\theta_i - b_j)], \text{ or equivalently}$$
$$a_j(\theta_i - b_j) \geq \frac{1}{1.7}\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right).$$

Finally, we obtain the following inequality:

$$b_j \leq \theta_i - \frac{1}{1.7a_j}\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right).$$

In the same way, $\forall i$, when $y_{ij} = 0$, we have $0 < \varphi_{ij} \leq Q^*_{ij}$. Therefore, the following inequality can be obtained:

$$a_j(\theta_i - b_j) \leq \frac{1}{1.7}\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right), \text{ or equivalently}$$
$$b_j \geq \theta_i - \frac{1}{1.7a_j}\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right).$$

Using the above inequalities $0 < \lambda_{ij} \leq P^*_{ij}$ and $0 < \varphi_{ij} \leq Q^*_{ij}$), we can obtain a truncated interval about the difficulty parameter b_j:

$$\theta_i - \frac{1}{1.7a_j}\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right) \leq b_j \leq \theta_i - \frac{1}{1.7a_j}\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right).$$

If this truncated interval is narrow, the sampling efficiency is improved and the parameter can converge fast. Therefore, we need to limit the upper and lower bounds of the truncated interval. In fact, we can obtain a maximum of $\theta_i - (1/1.7a_j)\log[(1-\varphi_{ij})/\varphi_{ij}]$ among all the examinees who correctly answer the jth item. Similarly, we can obtain a minimum of $\theta_i - (1/1.7a_j)\log[\lambda_{ij}/(1-\lambda_{ij})]$ among all the examinees who mistakenly answer the jth item. Finally, the full conditional posterior distribution of b_j can be obtained as a truncated prior distribution, with the truncated interval between maximum and minimum. The specific mathematical expressions are as follows.

Let $D_j = \{i \mid y_{ij} = 1, 0 < \lambda_{ij} \leq P^*_{ij}\}$ and $F_j = \{i \mid y_{ij} = 0, 0 < \varphi_{ij} \leq Q^*_{ij}\}$. Then, given a_j, c_j, γ_j, $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$, $\boldsymbol{\varphi}$, and $\boldsymbol{y}$, the full conditional posterior distribution of b_j is

$$b_j \mid a_j, c_j, \gamma_j, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\varphi}, \boldsymbol{y} \sim N(\mu_b, \sigma_b^2)\mathrm{I}(b_j^L \leq b_j \leq b_j^U), \quad (16)$$

where

$$b_j^L = \max_{i\in F_j}\left\{\theta_i - \frac{1}{1.7a_j}\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right)\right\} \quad \text{and}$$
$$b_j^U = \min_{i\in D_j}\left\{\theta_i - \frac{1}{1.7a_j}\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right)\right\}.$$

Subsequently, we update the discrimination parameter a_j. To ensure that this parameter is greater than zero, we use a truncated normal distribution with mean μ_a and variance σ_a^2 as a prior distribution, $N(\mu_a, \sigma_a^2)\mathrm{I}(a_j > 0)$. Under the condition $y_{ij} = 1, \forall i$, $\theta_i - b_j > 0$, we have $0 < \lambda_{ij} \leq P^*_{ij}$, while under the condition $y_{ij} = 0, \forall i$, $\theta_i - b_j < 0$, we have $0 < \varphi_{ij} \leq Q^*_{ij}$. The following inequalities concerning the discrimination parameter a_j can be established using a procedure similar to that used above to derive the truncated interval for the difficulty parameter b_j:

$$a_j \geq \frac{1}{1.7(\theta_i - b_j)}\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right),$$
$$a_j \geq \frac{1}{1.7(\theta_i - b_j)}\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right).$$

Similarly, when $y_{ij} = 1, \forall i$, $\theta_i - b_j < 0$, we have $0 < \lambda_{ij} \leq P^*_{ij}$, and when $y_{ij} = 0, \forall i$, $\theta_i - b_j > 0$, we have $0 < \varphi_{ij} \leq Q^*_{ij}$, from which we obtain

$$a_j \leq \frac{1}{1.7(\theta_i - b_j)}\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right)$$
$$a_j \leq \frac{1}{1.7(\theta_i - b_j)}\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right).$$

Let

$$\Delta_j = \{i \mid y_{ij} = 1, \theta_i - b_j > 0, 0 < \lambda_{ij} \leq P^*_{ij}\},$$
$$H_j = \{i \mid y_{ij} = 0, \theta_i - b_j < 0, 0 < \varphi_{ij} \leq Q^*_{ij}\},$$
$$\nabla_j = \{i \mid y_{ij} = 1, \theta_i - b_j < 0, 0 < \lambda_{ij} \leq P^*_{ij}\},$$
$$\Lambda_j = \{i \mid y_{ij} = 0, \theta_i - b_j > 0, 0 < \varphi_{ij} \leq Q^*_{ij}\}.$$

Given b_j, c_j, γ_j, $\boldsymbol{\lambda}$, $\boldsymbol{\varphi}, \boldsymbol{\theta}$, and $\boldsymbol{y}$, the full conditional posterior distribution of a_j is given by

$$a_j \mid b_j, c_j, \gamma_j, \boldsymbol{\lambda}, \boldsymbol{\varphi}, \boldsymbol{\theta}, \boldsymbol{y} \sim N(\mu_a, \sigma_a^2)\mathrm{I}(0 < a_j^L \leq a_j \leq a_j^U), \quad (17)$$

where

$$a_j^L = \max\left\{0, \max_{i\in\Delta_j}\left\{\frac{1}{1.7(\theta_i - b_j)}\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right)\right\}, \max_{i\in H_j}\left\{\frac{1}{1.7(\theta_i - b_j)}\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right)\right\}\right\},$$
$$a_j^U = \min\left\{\min_{i\in\nabla_j}\left\{\frac{1}{1.7(\theta_i - b_j)}\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right)\right\}, \min_{i\in\Lambda_j}\left\{\frac{1}{1.7(\theta_i - b_j)}\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right)\right\}\right\}.$$

In fact, the discrimination parameter is set to be greater than zero in the item response theory. Therefore, the prior distribution for the discrimination parameter is assumed to be a normal distribution truncated at 0. Based on the likelihood information, we can obtain the truncation interval of the discrimination parameter. However, the left endpoint of the truncation interval may be <0. In this case, we need to add 0 to the truncation interval to restrict the left endpoint in 17.

Finally, we update the latent ability θ_i. The prior of θ_i is assumed to follow a normal distribution, $\theta_i \sim N(\mu_\theta, \sigma_\theta^2)$. The latent ability θ_i is sampled from the following normal distribution with truncated interval between θ_i^L and θ_i^U:

$$\theta_i \mid \boldsymbol{\lambda}, \boldsymbol{\varphi}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{y} \sim N(\mu_\theta, \sigma_\theta^2)\mathrm{I}(\theta_i^L \leq \theta_i \leq \theta_i^U), \quad (18)$$

where

$$\theta_i^L = \max_{j\in D_i}\left\{\frac{1}{1.7a_j}\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right)+b_j\right\},$$
$$\theta_i^U = \min_{j\in F_i}\left\{\frac{1}{1.7a_j}\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right)+b_j\right\}.$$

4.2. Bayesian Model Assessment

In this paper, two Bayesian model assessment methods are considered to fit three different models (the 2PL, 3PL, and 4PL models), namely, the deviance information criterion (DIC; Spiegelhalter et al., 2002) and the logarithm of the pseudomarginal likelihood (LPML; Geisser and Eddy, 1979; Ibrahim et al., 2001). These two criteria are based on the log-likelihood functions evaluated at the posterior samples of the model parameters. Therefore, the DIC and LPML of the 4PL model can be easily computed. Write $\mathbf{\Omega} = (\mathbf{\Omega}_{ij}, i = 1,\ldots,N,\ j = 1,\ldots,J)$, where $\mathbf{\Omega}_{ij} = (\theta_i, a_j, b_j, c_j, \gamma_j)'$. Let $\{\mathbf{\Omega}^{(1)},\ldots,\mathbf{\Omega}^{(R)}\}$ denote an MCMC sample from the full conditional posterior distribution in (8) and (16)–(18), where $\mathbf{\Omega}^{(r)} = (\mathbf{\Omega}_{ij}^{(r)}, i = 1,\ldots,N,\ j = 1,\ldots,J)$ and $\mathbf{\Omega}_{ij}^{(r)} = (\theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, \gamma_j^{(r)})'$ for $i = 1,\ldots,N, j = 1,\ldots,J$, and $r = 1,\ldots,R$. The joint likelihood function of the responses can be written as

$$L(\boldsymbol{Y} \mid \mathbf{\Omega}) = \prod_{i=1}^{N}\prod_{j=1}^{J} f(y_{ij} \mid \theta_i, a_j, b_j, c_j, \gamma_j), \tag{19}$$

where $f(y_{ij} \mid \theta_i, a_j, b_j, c_j, \gamma_j)$ is the probability of response. The logarithm of the joint likelihood function in (19) evaluated at $\mathbf{\Omega}^{(r)}$ is given by

$$\log L(\boldsymbol{Y} \mid \mathbf{\Omega}^{(r)}) = \sum_{i=1}^{N}\sum_{j=1}^{J}\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, \gamma_j^{(r)}). \tag{20}$$

Since the joint log-likelihoods for the responses, $\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, \gamma_j^{(r)})$, $i = 1,\ldots,N$ and $j = 1,\ldots,J$, are readily available from MCMC sampling outputs, $\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, \gamma_j^{(r)})$ in (20) is easy to compute. We now calculate DIC as follows:

$$\text{DIC} = \widehat{\text{Dev}(\mathbf{\Omega})} + 2P_D = \widehat{\text{Dev}(\mathbf{\Omega})} + 2\left[\overline{\text{Dev}(\mathbf{\Omega})} - \widehat{\text{Dev}(\mathbf{\Omega})}\right], \tag{21}$$

where

$$\overline{\text{Dev}(\mathbf{\Omega})} = -\frac{2}{R}\sum_{r=1}^{R}\log L(\boldsymbol{Y} \mid \mathbf{\Omega}^{(r)}) \quad \text{and}$$
$$\widehat{\text{Dev}(\mathbf{\Omega})} = -2\max_{1\le r\le R}\log L(\boldsymbol{Y} \mid \mathbf{\Omega}^{(r)}).$$

In (21), $\overline{\text{Dev}(\mathbf{\Omega})}$ is a Monte Carlo estimate of the posterior expectation of the deviance function $\text{Dev}(\mathbf{\Omega}) = -2\log L(\boldsymbol{Y} \mid \mathbf{\Omega})$, $\widehat{\text{Dev}(\mathbf{\Omega})}$ is an approximation of $\text{Dev}(\widehat{\mathbf{\Omega}})$, where $\widehat{\mathbf{\Omega}}$ is the posterior mode, when the prior is relatively non-informative, and $P_D = \overline{\text{Dev}(\mathbf{\Omega})} - \widehat{\text{Dev}(\mathbf{\Omega})}$ is the effective number of parameters. Based on our construction, both DIC and P_D given in (21) are always non-negative. The model with a smaller DIC value fits the data better.

Letting $U_{ij,\max} = \max 1 \le r \le R\{-\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, \gamma_j^{(r)})\}$, we obtain a Monte Carlo estimate of the conditional predictive ordinate (CPO; Gelfand et al., 1992; Chen et al., 2000) as

$$\log\widehat{(\text{CPO}_{ij})} = -U_{ij,\max} - \log\left\{\frac{1}{R}\sum_{r=1}^{R}\exp[- \log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, \gamma_j^{(r)}) - U_{ij,\max}]\right\}. \tag{22}$$

Note that the maximum value adjustment used in $\log\widehat{(\text{CPO}_{ij})}$ plays an important role in numerical stabilization in computing $\exp[-\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, \gamma_j^{(r)}) - U_{ij,\max}]$ in (22). A summary statistic of the $\widehat{\text{CPO}_{ij}}$ is the sum of their logarithms, which is called the LPML and given by

$$\text{LPML} = \sum_{i=1}^{N}\sum_{j=1}^{J}\log\widehat{(\text{CPO}_{ij})}. \tag{23}$$

The model with a larger LPML has a better fit to the data.

5. SIMULATION STUDIES

5.1. Simulation Study 1

This simulation study is conducted to evaluate the recovery performance of the Gibbs-slice sampling algorithm based on different simulation conditions.

5.1.1. Simulation Design

The following manipulated conditions are considered: (a) test length $J = 20$ or 40 and (b) number of examinees $N = 500$, 1,000, or 2,000. Fully crossing different levels of these two factors yield six conditions (two test lengths × three sample sizes). Next, the true values of the parameters are given. True values of the item discrimination parameters a_j are generated from a uniform distribution, i.e., $a_j \sim U(0.5, 2.5)$, $j = 1, 2,\ldots,J$. The item difficulty parameters b_j are generated from a standardized normal distribution. The item guessing and slipping parameters (c_j, γ_j) are generated from $c_j \sim U(0, 0.25)$ and $\gamma_j \sim U(0, 0.25)\text{I}(\gamma_j < 1 - c_j)$. The ability parameters of examinees θ_i are also generated from a standardized normal distribution. In addition, we adopt non-informative prior distributions for the item parameters, i.e., $a_j \sim N(0, 10^5)\text{I}(0, +\infty)$, $b_j \sim N(0, 10^5)$, $g_j \sim \text{Beta}(1, 1)$, and $\gamma_j \sim \text{Beta}(1, 1)$, $j = 1, 2,\ldots,J$. The prior for the ability parameters is assumed to follow a standardized normal distribution owing to the model identification restrictions. One hundred replications are considered for each simulation condition.

5.1.2. Convergence Diagnostics

To evaluate the convergence of parameter estimation, we only consider convergence in the case of minimum sample sizes owing

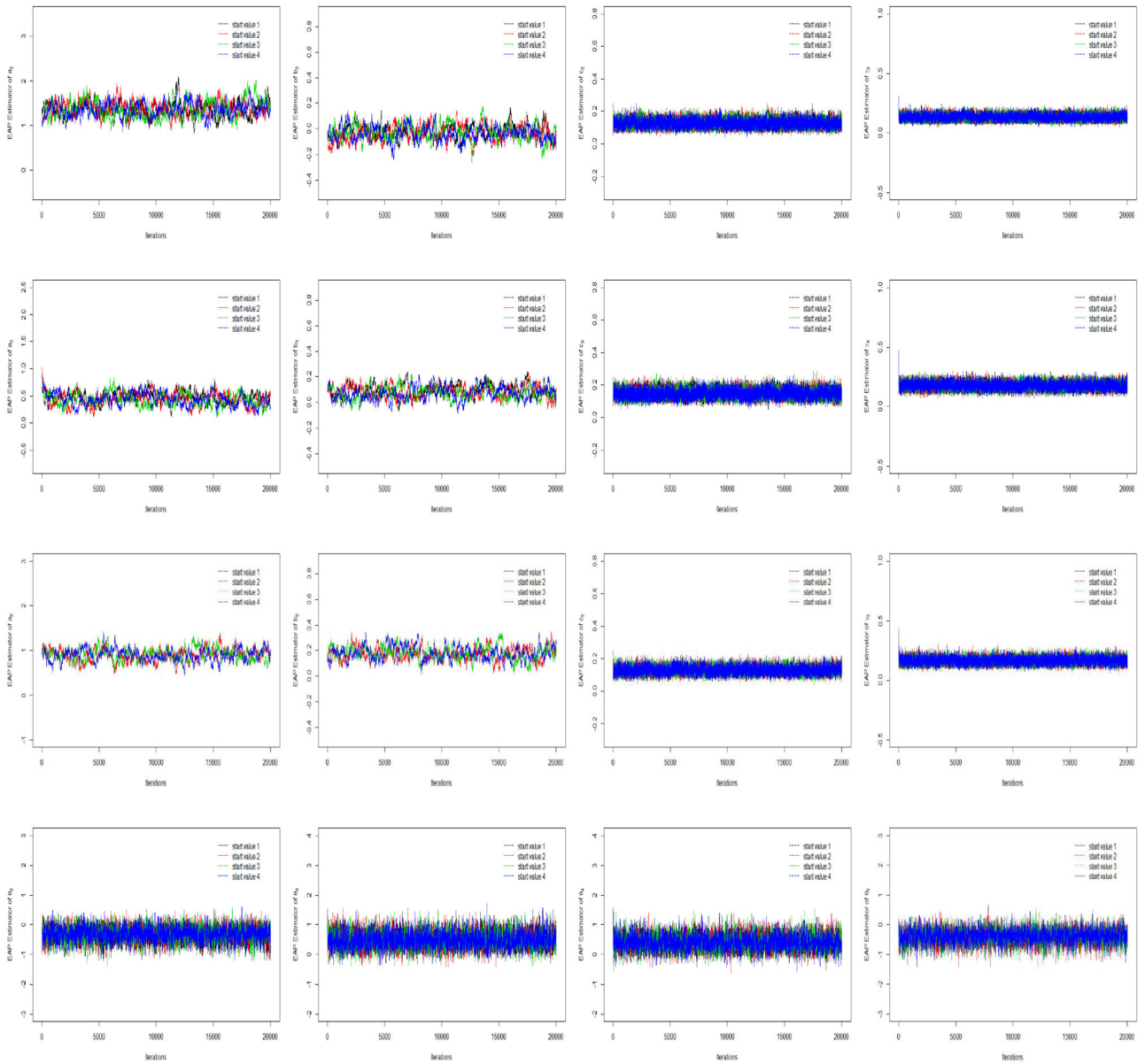

FIGURE 1 | The trace plots of three randomly selected items and persons for the Simulation Study 1.

to space limitations. That is, the test length is fixed at 20, and the number of examinees is 500. Two methods are used to check the convergence of our algorithm: the "eyeball" method to monitor convergence by visually inspecting the history plots of the generated sequences (Zhang et al., 2007), and the Gelman–Rubin method (Gelman and Rubin, 1992; Brooks and Gelman, 1998) to check the convergence of the parameters.

The convergence of the Gibbs-slice sampling algorithm is checked by monitoring the trace plots of the parameters for consecutive sequences of 20,000 iterations. The first 10,000 iterations are set as the burn-in period. As an illustration, four chains started at overdispersed starting values are run for each replication. The trace plots of three randomly selected items and persons are shown in **Figure 1**. In addition, the potential scale reduction factor (PSRF) ($\widehat{R}$; Brooks and Gelman, 1998) values of all item and person parameters are shown in **Figure 2**. We find that the PSRF values of all parameters are <1.2, which ensures that all chains converge as expected.

5.1.3. Item Parameter Recovery

The accuracy of the parameter estimates is measured by four evaluation criteria, namely, the Bias, mean squared error (MSE), standard deviation (SD), and coverage probability (CP) of the 95% highest probability density interval (HPDI) statistics. Let η

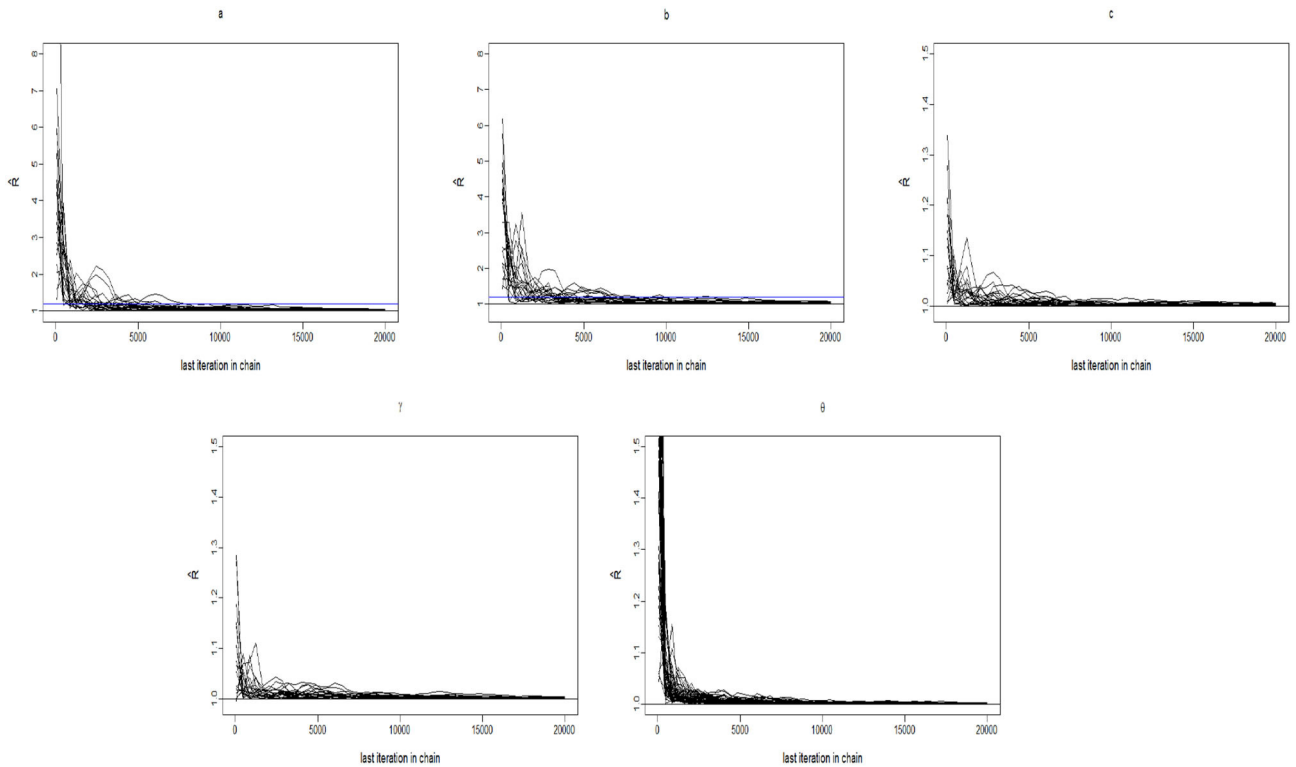

FIGURE 2 | The trace plots of $\widehat{R}$ for the Simulation Study 1.

be the parameter of interest. Assume that $M = 100$ data sets are generated. Also, let $\widehat{\eta}^{(m)}$ and $\mathrm{SD}^{(m)}(\eta)$ denote 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 η is defined as

$$\mathrm{Bias}(\eta) = \frac{1}{M}\sum_{m=1}^{M}(\widehat{\eta}^{(m)} - \eta), \tag{24}$$

the MSE for η is defined as

$$\mathrm{MSE}(\eta) = \frac{1}{M}\sum_{m=1}^{M}(\widehat{\eta}^{(m)} - \eta)^2, \tag{25}$$

and the average of the posterior standard deviation is defined as

$$\mathrm{SD}(\eta) = \frac{1}{M}\sum_{m=1}^{M}\mathrm{SD}^{(m)}(\eta). \tag{26}$$

Bias and MSE are important criteria used to evaluate the accuracy of parameter estimation in a simulation study. These criteria are used to investigate the relative distance between the parameter estimator and the true value. The greater the distance between the parameter estimator and the true value, the lower is the accuracy of parameter estimation and the poorer is the performance of the algorithm. However, for real data analysis, it is impossible to calculate Bias and MSE. The SD, on the other hand, can be calculated from the posterior samples of a Markov chain in simulation studies and real data analysis. In our simulation study, we calculate the average SD through repeated experiments to eliminate the error caused by randomness in a single simulation experiment.

The coverage probability is defined as

$$\mathrm{CP}(\eta) = \frac{\#\text{ of 95\% HPDIs containing } \eta \text{ in } M \text{ simulated data sets}}{M}. \tag{27}$$

The average Bias, MSE, SD, and CP for item parameters based on six different simulation conditions are shown in **Table 1**. The following conclusions can be drawn.

1. Given the total test length, when the number of individuals increases from 500 to 2,000, the average MSE and SD for discrimination, difficulty guessing, and slipping parameters show a decreasing trend. For example, let us consider a total test length of 20 items. When the number of examinees increases from 500 to 2,000, the average MSE and the average SD of all discrimination parameters decrease from 0.0625 to 0.0474 and from 0.1460 to 0.0759, respectively. The average MSE and the average SD of all difficulty parameters decrease from 0.0505 to 0.0263 and

TABLE 1 | Evaluating the accuracy of item parameters based on six different simulated conditions in Simulation Study 1.

	No. of examinees 500				No. of examinees 1,000				No. of examinees 2,000			
Item parameter	**Bias**	**MSE**	**SD**	**CP**	**Bias**	**MSE**	**SD**	**CP**	**Bias**	**MSE**	**SD**	**CP**
NO. OF ITEMS = 20												
Discrimination[a]	−0.0087	0.0625	0.1460	0.9514	−0.0217	0.0513	0.1037	0.9504	0.0005	0.0474	0.0759	0.9486
Difficulty[b]	−0.0000	0.0505	0.0559	0.9385	0.0000	0.0389	0.0390	0.9412	−0.0000	0.0263	0.0260	0.9285
Guessing[c]	−0.0215	0.0092	0.0247	0.9325	−0.0453	0.0045	0.0193	0.9378	−0.0830	0.0023	0.0156	0.9515
Slipping[γ]	0.0132	0.0060	0.0260	0.9342	−0.0176	0.0038	0.0217	0.9628	−0.0558	0.0025	0.0166	0.9548
NO. OF ITEMS = 40												
Discrimination[a]	−0.0029	0.0842	0.1482	0.9546	−0.0035	0.0705	0.0962	0.9390	−0.0129	0.0594	0.0638	0.9781
Difficulty[b]	−0.0000	0.0443	0.0561	0.9543	−0.0000	0.0325	0.0389	0.9495	0.0000	0.0224	0.0267	0.9652
Guessing[c]	−0.0238	0.0075	0.0250	0.9385	−0.0625	0.0059	0.0201	0.9322	−0.0677	0.0033	0.0154	0.9418
Slipping[γ]	−0.0061	0.0035	0.0264	0.9310	−0.0169	0.0025	0.0209	0.9438	−0.0407	0.0024	0.0152	0.9422

Note that the Bias, MSE, SD, and CP denote the average Bias, MSE, SD, and CP for the parameters. [a]Discrimination parameters, [b]Difficulty parameters, [c]Guessing parameters, [γ]Slipping parameters.

from 0.0559 to 0.0260, respectively. The average MSE and the average SD of all guessing parameters decrease from 0.0092 to 0.0023 and from 0.0247 to 0.0156, respectively. The average MSE and the average SD of all slipping parameters decrease from 0.0060 to 0.0025 and from 0.0260 to 0.0166, respectively.

2. Under the six simulated conditions, the average CPs of the discrimination, difficulty guessing, and slipping parameters are about 0.9500.
3. When the number of examinees is fixed at 500, 1,000, or 2,000, and the number of items is fixed at 40, the average MSE and SD show that the recovery results of the discrimination, difficulty, guessing, and slipping parameters are close to those in the case where the total test length is 20, which indicates that the Gibbs-slice sampling algorithm is stable and there is no reduction in accuracy owing to an increase in the number of items.

In summary, the Gibbs-slice sampling algorithm provides accurate estimates of the item parameters in term of various numbers of examinees and items. Next, we will explain why the Bias criterion is useful, and why it seems irrelevant in the simulation study.

If we want to determine whether our algorithm estimates the parameter accurately, we need more information to infer the parameter, which requires a large sample size. Here, Bias is an important criterion to evaluate the accuracy of parameter estimation. Let us give an example to illustrate the role of Bias. In Simulation Study 1, suppose that we investigate the accuracy of the algorithm in estimating a discrimination parameter. When the number of examinees increases from 500 to 2,000, the Bias of the discrimination parameter should show a decreasing trend. The result of Bias reduction further verifies that a greater number of samples are needed to improve the accuracy of parameter estimation.

In Simulation Study 1, we cannot enumerate the Bias of each item parameter one by one because there are too many simulation conditions and we are subject to space limitations. Therefore, we choose to calculate the average Bias of the parameter of interest. Next, we take the discrimination parameters as an example to further explain why Bias seems irrelevant in Simulation Study 1. Suppose that we have obtained 40 Biases of discrimination parameters, that the Bias values of these 40 discrimination parameters are either positive or negative, and that the average Bias of all 40 items is close to 0. However, the near-zero value of the average Bias does not show whether the parameter estimation is accurate or the result is caused by the positive and negative superposition of the 40 Biases. In fact, we find that the Bias for each item discrimination parameter show a decreasing trend with increasing number of examinees. To sum up, we do not analyze the results of the average Bias in the simulation studies, but Bias is indeed an important criterion to evaluate the accuracy of each parameter estimation.

5.1.4. Ability Parameter Recovery

Next, we evaluate the recovery of the latent ability using four accuracy evaluation criteria. The following conclusions can be obtained from **Table 2**.

1. Given a fixed number of examinees (500, 1,000, or 2,000), when the number of items increases from 20 to 40, the average MSE and SD for the ability parameters also show a decreasing trend.
2. Under the six simulated conditions, the average CP of the ability is also about 0.9500.
3. Given a fixed number of examinees (500, 1,000, or 2,000), when the number of items increases from 20 to 40, the correlation between the estimates and the true values tends to increase. For example, for 500 examinees, when the number of items increases from 20 to 40, the correlation between the estimates and the true values increases from 0.8631 to 0.9102.
4. Given a fixed number of items (20 or 40), when the number of examinees increases from 500 to 2,000, the correlation

TABLE 2 | Evaluating the accuracy of person parameters based on six different simulated conditions in Simulation Study 1.

No. of items	No. of examinees	Bias	MSE	SD	CP	Correlation with true value
20	500	0.0545	0.2783	0.2523	0.9428	0.8631
	1,000	0.0149	0.2923	0.2636	0.9675	0.8764
	2,000	0.0052	0.3341	0.2961	0.9322	0.8599
40	500	0.0315	0.2346	0.2180	0.9274	0.9102
	1,000	0.0764	0.2553	0.2343	0.9626	0.9182
	2,000	0.0439	0.3042	0.2866	0.9542	0.9225

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

between the estimates and the true values remains basically the same.

In summary, it is shown again that the Gibbs-slice sampling algorithm is effective and that the estimated results are accurate under various simulation conditions.

5.2. Simulation Study 2

Culpepper (2016) conducted an additional simulation study to confirm that the guessing and slipping parameters could give good recovery results in the process of Gibbs sampling regardless of whether informative or non-informative priors were used. Therefore, in this simulation study, we also adopt non-informative prior distributions for the guessing and slipping parameters in the Gibbs step to eliminate biased estimation of parameters due to wrong choices of the prior distributions, i.e., $c \sim \text{Beta}(1,1)$ and $\gamma \sim \text{Beta}(1,1)$, and we focus on the influence of different prior distributions on the accuracy of parameter estimation in the process of implementing the slice sampling algorithm. Note that in this simulation study, we do not focus on the accuracy of the guessing and slipping parameters, since Culpepper (2016) has already verified the accuracy of these two parameters in the case of the Gibbs algorithm under different types of prior distributions.

This simulation study is designed to show that the slice sampling algorithm is sufficiently flexible to recover various prior distributions of the item (discrimination and difficulty) and person parameters, and to address the sensitivity of the algorithm with different priors. Three types of prior distributions are examined: informative priors, non-informative priors, and inappropriate priors.

5.2.1. Simulation Design

The number of the examinees $N = 1,000$, and the test length $J = 20$. The true values for the items and persons are the same as in Simulation Study 1. One hundred replications are considered for each simulation condition. The following three kinds of prior distributions are considered in implementing the slice sampling algorithm:

(i) informative prior: $a \sim \log N(0,1)$, $b \sim N(0,1)$, and $\theta \sim N(0,1)$;
(ii) non-informative prior: $a \sim N(0,1000)\text{I}(0,+\infty)$, $b \sim \text{Uniform}(-1000,1000)$, and $\theta \sim N(0,1000)$;
(iii) inappropriate prior: (1) $a \sim \text{Exp}(1)$, $b \sim t(1)$, and $\theta \sim t(1)$; (2) $a \sim \text{Gamma}(3,2)$, $b \sim \text{Cauchy}(1,3)$, and $\theta \sim \text{Cauchy}(1,3)$.

The Gibbs-slice sampling algorithm is iterated 20,000 times. The first 10,000 iterations are discarded as burn-in. The PSRF values of all parameters are <1.2. The Bias, MSE, and SD of $\boldsymbol{a}$ and $\boldsymbol{b}$ based on the three kinds of prior distribution are shown in **Figure 3**.

5.2.2. Item Parameter Recovery

From **Figure 3**, we can see that the Bias, MSE, and SD of $\boldsymbol{a}$ and $\boldsymbol{b}$ are almost the same under different prior distributions. This shows that accuracy of parameter estimation can be guaranteed by the slice sampling algorithm, no matter what prior distribution is chosen, as long as the values sampled from this distribution belong to a reasonable parameter support set. In addition, the Bias, MSE and SD of $\boldsymbol{a}$ and $\boldsymbol{b}$ fluctuate around 0, which shows that the slice sampling algorithm is accurate and effective in estimating the item parameters.

5.2.3. Ability Parameter Recovery

Next, we evaluate the recovery of the latent ability based on different prior distributions in **Table 3**. We find that the MSE of ability parameters is between 0.2676 and 0.3014, and the corresponding SD is between 0.2436 and 0.3026 for all three kinds of prior distribution, which indicates that the choice of prior distribution has little impact on the accuracy of the ability parameters. In summary, the slice sampling algorithm is accurate and effective in estimating the person parameters. It is not sensitive to the specification of priors.

5.3. Simulation Study 3

In this simulation study, we use two Bayesian model assessment criteria to evaluate the model fittings. Two issues warrant further study. The first is whether the two criteria can accurately identify the true models under different design conditions. The second is that we study the phenomena of over-fitting and under-fitting between the true model and the fitting models.

5.3.1. Simulation Design

In this simulation, a number of individuals $N = 1,000$ is considered and the test length is fixed at 40. Three item response models are considered: the 2PL, 3PL, and 4PL models. Thus, we evaluate model fitting in the following three cases:

- Case 1: 2PL model (true model) vs. 2PL model, 3PL model, or 4PL model (fitted model).
- Case 2: 3PL model (true model) vs. 2PL model, 3PL model, or 4PL model (fitted model).

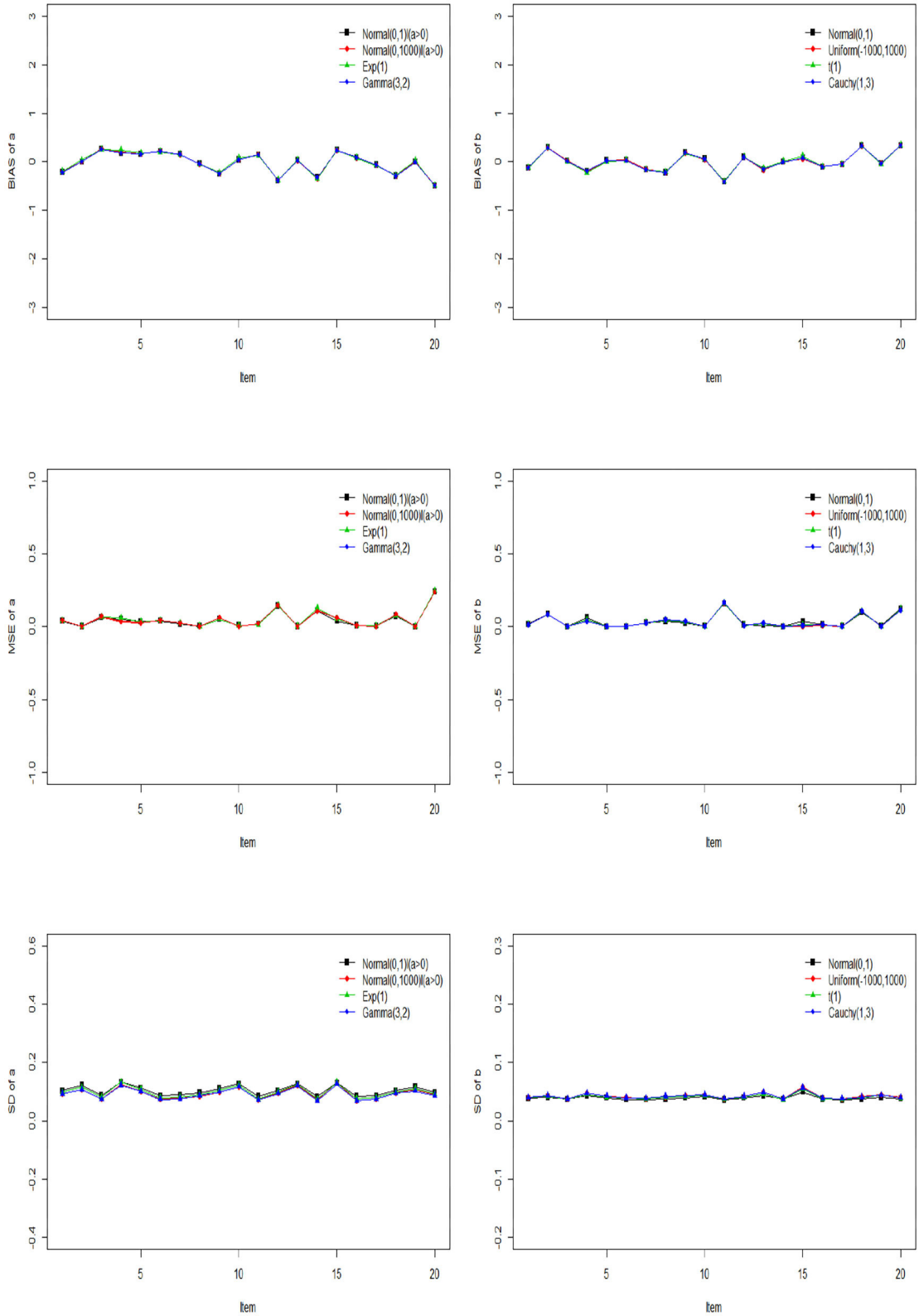

FIGURE 3 | The Bias, MSE, and SD of discrimination and difficulty parameters based on different priors.

TABLE 3 | Evaluating the accuracy of person parameters based on different prior distributions in Simulation Study 2.

Parameter	Accuracy evaluation index	Prior distribution			
		$N(0, 1)$	$N(0, 1,000)$	$t(1)$	$Cauchy(1, 3)$
θ	Bias	0.0064	0.0149	0.0087	0.0238
	MSE	0.2676	0.2923	0.3014	0.2713
	SD	0.2436	0.3026	0.2810	0.2983

Note that the Bias, MSE, and SD denote the average Bias, MSE, and SD for the ability parameters.

- Case 3: 4PL model (true model) vs. 2PL model, 3PL model, or 4PL model (fitted model).

The true values and prior distributions for the parameters are specified in the same way as in Simulation Study 1. To implement the MCMC sampling algorithm, chains of length 20,000 with an initial burn-in period of 10,000 are chosen. There are 100 replications for each simulation condition. The potential scale reduction factor (PSRF; Brooks and Gelman, 1998) values of all item and person parameters for each simulation condition are <1.2. The results of Bayesian model assessment based on the 100 replications are shown in **Table 4**.

From **Table 4**, we find that when the 2PL model is the true model, the 2PL model is chosen as the best-fitting model according to the results of DIC and LPML, which is what we expect to see. The medians of DIC and LPML are, respectively 29,333.1917 and −14,881.2617. The second best-fitting model is the 3PL model. The differences between the 2PL and 3PL models in the medians of DIC and LPML are −1234.1551 and 650.9820, respectively. The 4PL model is the worst model to fit the data. This is because the data are generated from a simple 2PL model, and the complex 4PL model is used to fit this data, which leads to over-fitting. The differences between the 2PL and 4PL models in the medians of DIC and LPML are −5369.4761 and 2805.5087, respectively. When the 3PL model is the true model, the DIC and LPML consistently choose the 3PL model as the best-fitting model, with the corresponding median values being 24,866.9338 and −12,523.6985, respectively. The second best-fitting model is the 2PL model. The differences between the 3PL and 4PL models in the medians of DIC and LPML are −7786.6968 and 3934.9003, respectively, while the corresponding differences between the 3PL and 2PL models are −7569.1249 and 3886.7071. This shows that when the data are generated from the 3PL model, the simple 2PL model is more appropriate to fit the data compared with the complex 4PL model. When the 4PL model is the true model, the two criteria consistently select the 4PL model as the best-fitting model. The other two models suffer from serious under-fitting. The differences between the 4PL and 2PL models in the medians of DIC and LPML are −7807.8880 and 4339.4735, respectively, while the corresponding differences between the 4PL and 3PL models are −1104.4156 and 634.0753. The failure to select the 2PL (3PL) model is attributed to the under-fitting caused by a few parameters. That is, the guessing and slipping parameters in the 4PL model play an important role in adjusting the probability of the tail of the item characteristic curve. In summary, the Bayesian assessment criteria are effective for identifying the true models and can be used in the following empirical example.

TABLE 4 | The results of Bayesian model assessment in the Simulation Study 3.

True model				2PL	3PL	4PL
Fitted model	2PL	DIC	Q_1	**29319.0702**	30539.6070	34676.5622
			Median	**29333.1917**	30567.3468	34702.6678
			Q_3	**29341.0284**	30591.9937	34722.2367
			IQR	21.9582	52.3867	45.6745
		LPML	Q_1	**−14888.3688**	−15543.2057	−17701.0943
			Median	**−14881.2617**	−15532.2437	−17686.7704
			Q_3	**−14875.4347**	−15515.0444	−17670.9324
			IQR	12.9341	28.1613	30.1319
	3PL	DIC	Q_1	32431.0873	**24857.3160**	32648.0788
			Median	32436.0587	**24866.9338**	32653.6306
			Q_3	32442.8955	**24878.2528**	32660.3940
			IQR	11.8082	20.9368	12.3152
		LPML	Q_1	−16413.9390	**−12528.9444**	−16462.3200
			Median	−16410.4056	**−12523.6985**	−16458.5988
			Q_3	−16406.8835	**−12517.8991**	−16453.9725
			IQR	7.0555	11.0453	8.3427
	4PL	DIC	Q_1	35560.7897	28870.1192	**27768.0166**
			Median	35583.7535	28880.2811	**27775.8655**
			Q_3	35611.7761	28890.8003	**27780.0024**
			IQR	50.9863	20.6810	11.9857
		LPML	Q_1	−18320.2375	−14603.6126	**−13965.3888**
			Median	−18302.6986	−14597.3004	**−13963.2251**
			Q_3	−18288.7386	−14593.5979	**−13958.0409**
			IQR	31.4988	10.0147	7.3479

Note that the boldface values indicate that the corresponding model is the best fitted model with the smallest DIC and largest LPML values.

6. EMPIRICAL EXAMPLE

In this example, the 2015 computer-based PISA (Program for International Student Assessment) science data are used. Among the many countries that have participated in this computer-based assessment of sciences, we choose students from the USA as the object of analysis. The original sample size of students is 658, and 110 students with Not Reached (original code 6) or Not Response (original code 9) are removed, with Not Reached and Not Response (omitted) being treated as missing data. The final 548 students answer 16 items. All 16 items are scored using a dichotomous scale. The descriptive statistics for these PISA data are shown in **Table 5**. We find that three items, DR442Q05C, DR442Q06C, and CR442Q07S, have lower correct rates than the other items, with the corresponding values being 25.7, 23.2, and 28.5%, respectively. The correct rate represents the proportion at which all examinees answer each item correctly. Moreover, the four items with the highest correct rates are

TABLE 5 | The descriptive statistics for PISA 2015 released computer-based sciences items.

Item	Item code	Correct rate (%)	Item	Item code	Correct rate (%)
1	CR083Q01S	54.2	9	CR442Q07S	28.5
2	CR083Q02S	83.6	10	CR245Q01S	53.8
3	CR083Q03S	75.2	11	CR245Q02S	60.0
4	CR083Q04S	66.6	12	CR101Q01S	43.6
5	DR442Q02C	80.1	13	CR101Q02S	87.6
6	DR442Q03C	76.5	14	CR101Q03S	57.7
7	DR442Q05C	25.7	15	CR101Q04S	80.1
8	DR442Q06C	23.1	16	CR101Q05S	48.7

Note that the correct rate represents the percentage of all examinees who correctly answer each item.

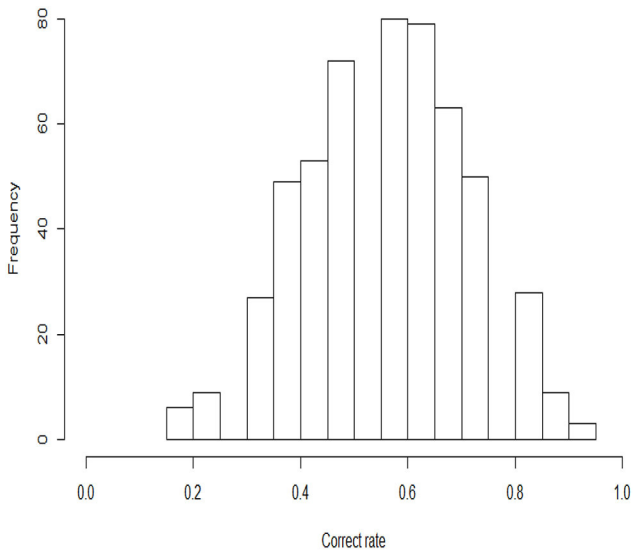

FIGURE 4 | Frequency histograms of the correct rates for 548 examinees.

CR101Q02S (87.6%), CR083Q02S (83.6%) DR442Q02C (80.1%), and CR101Q04S (80.1%). The frequency histogram of the correct rates for the 548 examinees is shown in **Figure 4**.

6.1. Bayesian Model Assessment

We consider three models to fit the PISA data: the 2PL, 3PL, and 4PL models. In the estimation procedure, the same non-informative priors as in Simulation Study 1 are utilized for the unknown parameters. In all of the Bayesian computations, we use 20,000 MCMC samples after a burn-in of 10,000 iterations for each model to compute all posterior estimates. The convergence of the chains is checked by PSRF. The PSRF values of all item and ability parameters for each model are <1.2. On this basis, the results of Bayesian model assessment for the PISA data are shown in **Table 6**.

According to DIC and LPML in **Table 6**, we find that the 4PL model is the best-fitting model compared with the 2PL and 3PL models. The values of DIC and LPML are 10,854.2075 and −5494.4088, respectively. The second best-fitting model is the 3PL model. The differences between the 4PL and 3PL models in DIC and LPML are −1376.1744 and 674.5340, respectively. This shows that the introduction of slipping parameters in the 3PL model is sufficient to fit these PISA data. The worst-fitting model is the 2PL model. This is attributed to the relatively simple structure of this model, which makes it unable to describe changes in probability at the end of the item characteristic curve caused by guessing or slipping. The differences between the 4PL and 2PL models in DIC and LPML are −3353.7433 and 1796.3457, respectively.

TABLE 6 | The results of Bayesian model assessment for the PISA data.

Model	DIC	LPML
2PL model	14206.9508	−7290.7545
3PL model	12230.3819	−6168.9428
4PL model	**10854.2075**	**− 5494.4088**

Note that the boldface values indicate that the corresponding model is the best fitted model with the smallest DIC and largest LPML values.

Next, we will use the 4PL model to analyze the PISA data in detail based on the results of the model assessment.

6.2. Analysis of Item Parameters

The estimated results for the item parameters are shown in **Table 7**, from which we find that the expected a posteriori (EAP) estimations of the 11 item discrimination parameters are greater than one. This indicates that these items can distinguish the differences between abilities well. The five items with the lowest discrimination are items 16 (CR101Q05S), 10 (CR245Q01S), 12 (CR101Q01S), 2 (CR083Q02S), and 5 (DR442Q02C) in turn. The EAP estimates of the discrimination parameters for these five items are 0.6681, 0.6792, 0.7348, 0.8083, and 0.8901. In addition, the EAP estimates of seven of the difficulty parameters are less than zero, which indicates that these seven items are easier than the other nine items. The five most difficult items are items 8 (DR442Q06C), 7 (DR442Q05C), 9 (CR442Q07S), 12 (CR101Q01S), and 16 (CR101Q05S) in turn. The EAP estimates of the difficulty parameters for these five items are 1.2528, 1.2203, 1.0804, 0.4521, and 0.3102. The corresponding correct rates in **Table 5** for these five items are 23.1, 25.7, 28.5, 43.63 and 48.7%, respectively. The most difficult five items have low correct rates, which is consistent with our intuition. The EAP estimates of the guessing parameters for the 16 items range from 0.0737 to 0.1840. The five items with the highest guessing parameters are items 2 (CR083Q02S), 5 (DR442Q02C), 13 (CR101Q02S), 15 (CR101Q04S), and 3 (CR083Q03S) in turn. The EAP estimates of the guessing parameters for these five items are 0.1840, 0.1791, 0.1790, 0.1673, and 0.3102. We find that the five items with high guessing parameters also have high correct rates. The corresponding correct rates for these five items are 83.6, 80.1, 87.6, 80.1, and 75.2%. This shows that these five items are more likely to be guessed correctly than the other 11 items. In addition, the five easiest slipping items are items 8 (DR442Q06C), 7 (DR442Q05C), 9 (CR442Q07S), 12 (CR101Q01S), and 16 (CR101Q05S) in turn. The EAP estimates of the slipping parameters for these five items are 1.785, 1.619, 1.581, 0.1481, and 0.1431. We find that the more difficult an

TABLE 7 | The estimation results of item parameter for the PISA data.

PARM	EAP	SD	HPDI	PARM	EAP	SD	HPDI
a_1	1.0416	0.1227	[0.8215, 1.2856]	b_1	0.1939	0.0615	[0.0861, 0.3222]
a_2	0.8083	0.1316	[0.5715, 1.0665]	b_2	−0.8496	0.0815	[−0.9936, −0.6793]
a_3	1.1171	0.1513	[0.8327, 1.4101]	b_3	−0.5071	0.0625	[−0.6214, −0.3699]
a_4	1.1119	0.1308	[0.8813, 1.3996]	b_4	−0.1947	0.0563	[−0.3030, −0.0876]
a_5	0.8901	0.1034	[0.6847, 1.0933]	b_5	−0.6969	0.0623	[−0.8230, −0.5741]
a_6	1.2772	0.1719	[0.9642, 1.6355]	b_6	−0.5966	0.0700	[−0.7525, −0.4675]
a_7	1.3404	0.1348	[1.0800, 1.5839]	b_7	1.2203	0.0778	[1.0635, 1.3738]
a_8	1.1202	0.1608	[0.7827, 1.4713]	b_8	1.2528	0.0966	[1.0313, 1.4246]
a_9	1.2377	0.1475	[0.9338, 1.5149]	b_9	1.0804	0.0819	[0.9117, 1.2155]
a_{10}	0.6792	0.1125	[0.4780, 0.9079]	b_{10}	0.1669	0.0640	[0.0423, 0.2832]
a_{11}	1.0720	0.1214	[0.8432, 1.3184]	b_{11}	0.0258	0.0512	[−0.0617, 0.1330]
a_{12}	0.7348	0.0897	[0.5528, 0.9035]	b_{12}	0.4521	0.0548	[0.3448, 0.5506]
a_{13}	1.1994	0.1706	[0.8682, 1.5513]	b_{13}	−1.1843	0.0841	[−1.3510, −1.0305]
a_{14}	1.0083	0.1219	[0.7666, 1.2336]	b_{14}	0.0985	0.0525	[0.0029, 0.2053]
a_{15}	1.2047	0.1707	[0.8618, 1.5329]	b_{15}	−0.7719	0.0667	[−0.9095, −0.6543]
a_{16}	0.6681	0.0924	[0.4999, 0.8482]	b_{16}	0.3102	0.0584	[0.2012, 0.4321]
c_1	0.1344	0.0254	[0.0870, 0.1853]	γ_1	0.1170	0.0225	[0.0738, 0.1616]
c_2	0.1840	0.0363	[0.1137, 0.2545]	γ_2	0.0736	0.0142	[0.0466, 0.1023]
c_3	0.1650	0.0315	[0.1065, 0.2285]	γ_3	0.0804	0.0155	[0.0506, 0.1106]
c_4	0.1532	0.0292	[0.1006, 0.2137]	γ_4	0.0950	0.0182	[0.0605, 0.1306]
c_5	0.1791	0.0343	[0.1131, 0.2461]	γ_5	0.0781	0.0149	[0.0495, 0.1077]
c_6	0.1607	0.0309	[0.1014, 0.2210]	γ_6	0.0749	0.0148	[0.0458, 0.1032]
c_7	0.0737	0.0147	[0.0459, 0.1023]	γ_7	0.1619	0.0314	[0.1034, 0.2261]
c_8	0.0805	0.0152	[0.0507, 0.1096]	γ_8	0.1785	0.0339	[0.1145, 0.2470]
c_9	0.0842	0.0159	[0.0549, 0.1165]	γ_9	0.1581	0.0307	[0.0983, 0.2178]
c_{10}	0.1561	0.0279	[0.1024, 0.2115]	γ_{10}	0.1313	0.0248	[0.0832, 0.1786]
c_{11}	0.1485	0.0268	[0.0996, 0.2035]	γ_{11}	0.1028	0.0197	[0.0646, 0.1408]
c_{12}	0.1361	0.0243	[0.0897, 0.1842]	γ_{12}	0.1481	0.0275	[0.0967, 0.2040]
c_{13}	0.1790	0.0354	[0.1118, 0.2484]	γ_{13}	0.0607	0.0118	[0.0368, 0.0827]
c_{14}	0.1469	0.0268	[0.0952, 0.1991]	γ_{14}	0.1100	0.0211	[0.0697, 0.1523]
c_{15}	0.1673	0.0322	[0.1057, 0.2299]	γ_{15}	0.0716	0.0143	[0.0444, 0.1006]
c_{16}	0.1505	0.0266	[0.0991, 0.2028]	γ_{16}	0.1431	0.0268	[0.0931, 0.1960]

PARM denotes parameter, EAP is the expected a posteriori estimation, SD denotes the standard deviation, and HPDI denotes the highest probability density interval.

item is, the more likely is the examinee to slip in answering it, which leads to a reduction in the correct rate. The SDs of the discrimination parameters range from 0.0897 to 0.1719, those of the difficulty parameters from 0.0512 to 0.0966, those of the guessing parameters from 0.0147 to 0.0363, and those of the slipping parameters from 0.0118 to 0.0339.

6.3. Analysis of Person Parameters

The histograms of the posterior estimates of the ability parameters are shown in **Figure 5**. Most of the estimated abilities of the examinees are near zero. The number of examinees with high ability (the estimates are between 0 and 1.2) is more than the number with low ability (the estimates are between −1.2 and 0). The ability parameter posterior histogram is consistent with the frequency histogram of the correct rate (**Figure 4**). That is, the trend of change in the correct rate in the histogram is same as that in the ability posterior histogram. The number of examinees with high correct rate is more than the number with low correct rate. It is once again verified that the results of the estimation are accurate.

7. DISCUSSION

In this paper, an efficient Gibbs-slice sampling algorithm in a fully Bayesian framework has been proposed to estimate the 4PL model. This algorithm, as its name suggests, can be conceived of as an extension of the Gibbs algorithm. The sampling process consists of two parts. One part is the Gibbs algorithm, which is used to update the guessing and slipping parameters when non-informative uniform priors are employed for cases that are prototypical of educational and psychopathology items. This part implements sampling by using a conjugate prior and greatly increases efficiency. The other part is the slice sampling algorithm, which samples the 2PL IRT model from the truncated

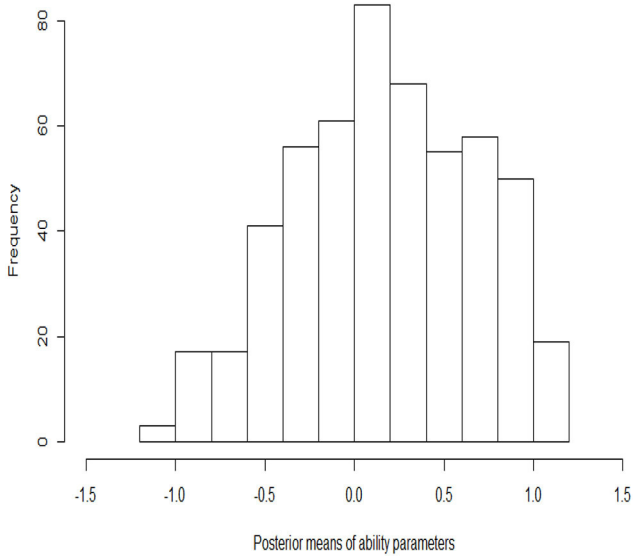

FIGURE 5 | The histograms of the posterior estimates of ability parameters.

full conditional posterior distribution by introducing auxiliary variables. The motivations for the slice sampling algorithm are manifold. First, this algorithm has the advantage of flexibility in the choice of prior distribution to obtain samples from the full conditional posterior distributions, rather than being restricted to using the conjugate distributions as in the Gibbs sampling process, which is also limited to the normal ogive framework. This allows the use of informative priors, non-informative priors, and inappropriate priors for the item parameters. Second, the Metropolis–Hastings algorithm depends on the proposal distributions and variances (tuning parameters) and is sensitive to step size. If the step size is too small, the chain will take longer to traverse the target density. If the step size is too large, there will be inefficiencies due to a high rejection rate. However, the slice sampling algorithm can automatically tune the step size to match the local shape of the target density and draw samples with acceptance probability equal to one. Thus, it is easier and more efficient to implement.

However, the computational burden of the Gibbs-slice sampling algorithm becomes intensive, especially when a large numbers of examinees or items are considered, or a large MCMC sample size is used. Therefore, it is desirable to develop a standalone R package associated with C++ or Fortran software for more a extensive large-scale assessment program. In fact, the new algorithm based on auxiliary variables can be extended to estimate some more complex item response and response time models, for example, the graded response model or the Weibull response time model. Only DIC and LPML have been considered in this study, but other Bayesian model selection criteria such as marginal likelihoods may also be potentially useful to compare different IRT models. These extensions are beyond the scope of this paper but are currently under further investigation.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.oecd.org/pisa/data/.

AUTHOR CONTRIBUTIONS

JZ completed the writing of the article and provided article revisions. JL provided original thoughts. JZ, JL, HD, and ZZ provided key technical support. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Fundamental Research Funds for the Central Universities of china (Grant No. 2412020QD025), and the "Youth Development Project" of School of Mathematics and Statistics of Northeast Normal University and the Foundation for Postdoctoral of Yunnan University (Grant No. CI76220200).

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibb ssampling. *J. Educ. Stat.* 17, 251–269.

Baker, F. B., and Kim, S. H. (2004). *Item Response Theory: Parameter Estimation Techniques*. New York, NY: Marcel Dekker.

Barton, M. A., and Lord, F. M. (1981). *An Upper Asymptote for the Three-Parameter Logistic Item Response Model*. Princeton, NJ: Educational Testing Service.

Béguin, A. A., and Glas, C. A. W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika* 66, 541–561. doi: 10.1007/BF02296195

Birnbaum, A. (1957). *Efficient Design and Use of Tests of a Mental Ability for Various Decision-Making Problems. Series Report No. 58-16*. Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores* , eds F. M. Lord and M. R. Novick (Reading, MA: MIT Press), 397–479.

Bishop, C. M. (2006). "Slice sampling," in *Pattern Recognition and Machine Learning*, eds M. Jordan, J. Kleinberg, B. Schölkopf (New York, NY: Springer), 523–558.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46, 443–459. doi: 10.1007/BF02293801

Brooks, S. P., and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the Renvironment. *J. Stat. Softw.* 48, 1–29. doi: 10.18637/jss.v048.i06

Chang, H.-H., and Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika* 73, 441–450. doi: 10.1007/s11336-007-9047-7

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

Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *Am. Stat.* 49, 327–335. doi: 10.1080/00031305.1995.10476177

Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika* 81, 1142–1163. doi: 10.1007/s11336-015-9477-6

Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by auxiliary variables. *J. R. Stat. Soc. Ser. B* 61, 331–344. doi: 10.1111/1467-9868.00179

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

Ferrando, P. J. (1994). Fitting item response models to the EPI-A impulsivity subscale. *Educ. Psychol. Measure.* 54, 118–127. doi: 10.1177/0013164494054001016

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York, NY: Springer.

Fox, J. P. (2005). Multilevel IRT using dichotomous and polytomous items. *Br. J. Math. Stat. Psychol.* 58, 145–172. doi: 10.1348/000711005X38951

Fox, J. P., and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 66, 271–288. doi: 10.1007/BF02294839

Fraley, R. C., Waller, N. G., and Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *J. Pers. Soc. Psychol.* 78, 350–365. doi: 10.1037/0022-3514.78.2.350

Geisser, S., and Eddy, W. F. (1979). A predictive approach to model selection. *J. Am. Stat. Assoc.* 74, 153–160. doi: 10.1080/01621459.1979.10481632

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). "Model determination using predictive distributions with implementation via sampling-based methods (with discussion)," in *Bayesian Statistics, Vol. 4*, eds J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford: Oxford University Press), 147–167.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* 85, 398–409. doi: 10.1080/01621459.1990.10476213

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* 6, 721–741. doi: 10.1109/TPAMI.1984.4767596

Gray-Little, B., Williams, V. S. L., and Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Pers. Soc. Psychol. Bull.* 23, 443–451. doi: 10.1177/0146167297235001

Green, B. F. (2011). A comment on early student blunders on computer-based adaptive tests. *Appl. Psychol. Measure.* 35, 165–174. doi: 10.1177/0146621610377080

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109. doi: 10.1093/biomet/57.1.97

Hessen, D. J. (2005). Constant latent odds- ratios models and the Mantel-Haenszel null hypothesis. *Psychometrika* 70, 497–516. doi: 10.1007/s11336-002-1040-6

Hockemeyer, C. (2002). A comparison of non-deterministic procedures for the adaptive assessment of knowledge. *Psychol. Beiträge* 44, 495–503.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York, NY: Springer.

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Measure.* 25, 258–272. doi: 10.1177/01466210122032064

Lanza, S. T., Foster, M., Taylor, T. K., and Burns, L. (2005). *Assessing the impact of measurement specificity in a behavior problems checklist: An IRT analysis.* Technical Report 05-75. The Pennsylvania State University; The Methodology Center, University Park, PA.

Liao, W.-W., Ho, R.-G., Yen, Y.-C., and Cheng, H.-C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Soc. Behav. Pers.* 40, 1679–1694. doi: 10.2224/sbp.2012.40.10.1679

Loken, E., and Rulison, K. (2010). Estimation of a four-parameter item response theory model. *Br. J. Math. Stat. Psychol.* 63, 509–525. doi: 10.1348/000711009X474502

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Lu, J., Zhang, J. W., and Tao, J. (2018). Slice-Gibbs sampling algorithm for estimating the parameters of a multilevel item response model. *J. Math. Psychol.* 82, 12–25. doi: 10.1016/j.jmp.2017.10.005

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Appl. Psychol. Measure.* 37, 304–315. doi: 10.1177/0146621613475471

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092. doi: 10.1063/1.1699114

Neal, R. (2003). Slice sampling. *Ann. Stat.* 31, 705–767. doi: 10.1214/aos/1056562461

Ogasawara, H. (2012). Asymptotic expansions for the ability estimator in item response theory. *Comput. Stat.* 27, 661–683. doi: 10.1007/s00180-011-0282-0

Osgood, D. W., McMorris, B. J., and Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: item response theory scaling. *J. Quant. Criminol.* 18, 267–296. doi: 10.1023/A:1016008004010

Patz, R. J., and Junker, B. W. (1999). A straight forward approach to Markov chain Monte Carlo methods for item response models. *J. Educ. Behav. Stat.* 24, 146–178. doi: 10.3102/10769986024002146

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Reise, S. P., and Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychol. Methods* 8, 164–184. doi: 10.1037/1082-989X.8.2.164

Rouse, S. V., Finger, M. S., and Butcher, J. N. (1999). Advances in clinical personality measurement: an item response theory analysis of the MMPI-2 PSY-5 scales. *J. Pers. Assess.* 72, 282–307. doi: 10.1207/S15327752JP720212

Rulison, K. L., and Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Appl. Psychol. Measure.* 33, 83–101. doi: 10.1177/0146621608324023

Spiegelhalter, D. J, Thomas, A., Best, N. G., and Lunn, D. (2003). *WinBUGS Version 1.4 User Manual*. Cambridge: MRC Biostatistics Unit. Available online at: http://www.mrc-bsu.cam.ac.uk/wp-content/uploads/manual14.pdf

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64, 583–639. doi: 10.1111/1467-9868.00353

Steinberg, L., and Thissen, D. (1995). "Item response theory in personality research," in *Personality Research, Methods, and Theory: A Festschrift Honoring Donald W. Fiske*, eds P. E. Shrout and S. T. Fiske (Hillsdale, NJ: Erlbaum), 161–181.

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82, 528–550. doi: 10.1080/01621459.1987.10478458

Tavares, H. R., de Andrade, D. F., and Pereira, C. A. (2004). Detection of determinant genes and diagnostic via item response theory. *Genet. Mol. Biol.* 27, 679–685. doi: 10.1590/S1415-47572004000400033

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussions). *Ann. Stat.* 22, 1701–1762. doi: 10.1214/aos/1176325750

Van der Linden, W. J., and Hambleton, R. K. (eds.). (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer.

Waller, N. G., and Feuerstahler, L. M. (2017). Bayesian modal estimation of the four-parameter item response model in real, realistic, and idealized data sets. *Multivar. Behav. Res.* 52, 350–370. doi: 10.1080/00273171.2017.1292893

Waller, N. G., and Reise, S. P. (2010). "Measuring psychopathology with non-standard IRT models: fitting the four-parameter model to the MMPI," in *Measuring Psychological Constructs: Advances in Modelbased Approaches*, eds S. Embretson and J. S. Roberts (Washington, DC: American Psychological Association), 147–173.

Yen, Y.-C., Ho, R.-G., Laio, W.-W., Chen, L.-J., and Kuo, C.-C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Appl. Psychol. Measure.* 36, 75–78. doi: 10.1177/0146621611432862

Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., and Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *Int. J. Behav. Dev.* 31, 374–383. doi: 10.1177/0165025407077764

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.