MATHEMATICAL MODELS FOR INTERTEMPORAL CHOICE

EDITED BY: Salvador Cruz Rambaud and Taiki Takahashi

PUBLISHED IN: Frontiers in Psychology and
Frontiers in Applied Mathematics and Statistics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-409-4
DOI 10.3389/978-2-88971-409-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

MATHEMATICAL MODELS FOR INTERTEMPORAL CHOICE

Topic Editors:
Salvador Cruz Rambaud, University of Almeria, Spain
Taiki Takahashi, Hokkaido University, Japan

Citation: Rambaud, S. C., Takahashi, T., eds. (2021). Mathematical Models for Intertemporal Choice. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88971-409-4

Table of Contents

EDITORIAL
published: 04 August 2021
doi: 10.3389/fams.2021.743337

Editorial: Mathematical Models for Intertemporal Choice

*Salvador Cruz Rambaud**

Department of Economics and Business, University of Almeria, Almería, Spain

Keywords: intertemporal choice, anomalies, behavioral finance, mathematical model, decision making

OPEN ACCESS

Edited and reviewed by:
Young Shin Aaron Kim,
Stony Brook University, United States

****Correspondence:***
Salvador Cruz Rambaud
scruz@ual.es

Specialty section:
This article was submitted to Mathematical Finance, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *18 July 2021*
Accepted: *26 July 2021*
Published: *04 August 2021*

Citation:
Cruz Rambaud S (2021) Editorial: Mathematical Models for Intertemporal Choice. Front. Appl. Math. Stat. 7:743337. doi: 10.3389/fams.2021.743337

Editorial on the Research Topic

Mathematical Models for Intertemporal Choice

In general, a wide variety of approaches are allowed in Mathematical Finance: one of them involves the implementation of mathematical models able to explain the complexity of real situations in Finance. In particular, intertemporal choice is gaining the attention of researchers because of its increasing application to other fields –such as psychology or health. Obviously, every model presents logical mistakes (or gaps), and intertemporal choice is not an exception. This was shown in the recent Research Topic labeled as “Intertemporal Choice and Its Anomalies”.

The objective of this Research Topic was to describe intertemporal choices as mathematical models, as general as possible, with the aim to cover all possible situations and analyze the properties which can be useful for decision makers. Since most financial decisions include decision making over time, this Research Topic is aimed also at mathematical modeling of important anomalies such as Allais’ paradox (violation of von Neumann and Morgenstern’s independence axiom), mental accounting, and myopic loss aversion in behavioral finance (discovered by Nobel laureate Professor Richard H. Thaler and colleagues).

Also, in behavioral finance, Nobel laureate Professor Robert J. Schiller observed excessive volatility in comparison to streams of future dividends in the United States stock markets, which reflects inefficiency in the market and irrationality in people who trade stocks. Mathematical models which have implications for these anomalies in the markets are also within a scope of this collection. Furthermore, recent advances in neuroeconomics revealed the important roles of emotion in a decision over time and under uncertainty.

Jiwei Zhang, Jing Lu, Hang Du and Zhaoyuan Zhang introduce a new Gibbs slice sampling algorithm for estimating the four-parameter logistic model which has reached lot of interest in educational testing and psychological measurement. The sampling process was divided into two parts. The first part is the Gibbs algorithm, which was used to update the guessing and slipping parameters when non-informative uniform priors are employed for cases which are prototypical of educational and psychopathology items. The second part is the slice algorithm, which samples the 2PL IRT model from the truncated full conditional posterior distribution by using auxiliary variables.

Yang Lu, Jian Wang, Chenyang Li, Haoya Huang and Xintian Zhuang provide an extension of the paper by Hoelzl et al. (2011) and Cruz Rambaud et al. (2019) as they deal with the improving sequence effect in loans contexts. Traditionally, this anomaly of the intertemporal choice was reduced to choices between rising earnings and other increasing/decreasing sequences. In a beginning, previous studies have shown a consistent preference for falling sequences in the context of loan repayment plans. However, the results show that consumers follow a comparison-based decision making process rather than optimization when evaluating temporally reframed loan offerings. Individuals preferred the falling over the constant profile only if the interest rate was 10% and the loan profiles were described in a per-year form.

On the other hand, Jiwei Zhang, Jing Lu, Jing Yang, Zhaoyuan Zhang, and Shanshan Sun introduce the so-called "Mixture Multiple Strategy-Deterministic, Inputs, Noisy and Gate (MMS-DINA) model in order to investigate individual differences in the choice of responses categories. The simulation indicates that the Markov chain Monte Carlo (MCMC) algorithm can be used to obtain accurate parameter estimates. Additionally, two Bayesian model assessment criterions are considered to evaluate the model fitting among DINA model, MS-DINA model and MMS-DINA model. Thus, it is shown that, when the data are generated from the simple single-strategy DINA model, the MMS-DINA model fits the data better than the MS-DINA model.

Finally, Xuemei Xue, Jing Lu, and Jiwei Zhang introduce a multidimensional Rasch model for measuring learning and change (MRMLC) and its dichotomous and polytomous extensions is used in longitudinal study. Two simulation studies have been carried out to further illustrate the advantages of this item-weighted likelihood estimation method compared to the traditional Maximum a Posteriori (MAP) estimation method, Maximum Likelihood Estimation (MLE) method, Warm's Weighted Likelihood Estimation (WLE) method, and Type-Weighted maximum Likelihood Estimation (TWLE) method, resulting in a better recover examinees' true ability level for both complex longitudinal IRT models and unidimensional IRT models compared to the existing likelihood estimation methods.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

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 $\delta_i^{(r)} \sim \mathrm{Uniform}\Big(0, l_i(x^{(r-1)})\Big)$, $i = 1, \ldots, N$.
- 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 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 $\mathrm{Exp}(1)$ and $\mathrm{Gamma}(2,3)$.

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:

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

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):

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

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

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

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

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

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

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

Equivalently,

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

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:

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

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

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

The specific form can be represented as

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

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:

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

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

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

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

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 $0 < \lambda_{ij} \le P_{ij}^{*}$, and the following inequality can be established:

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

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

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

from which

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

Therefore, we have

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

Finally, we obtain the following inequality:

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

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

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

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

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

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

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

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 $0 < \lambda_{ij} \le P^*_{ij}$, while under the condition $y_{ij} = 0, \forall i, \theta_i - b_j < 0$, we have $0 < \varphi_{ij} \le Q^*_{ij}$. 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:

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

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

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

Let

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

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

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

where

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

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:

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

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 $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_{ij}, i = 1, \ldots, N, j = 1, \ldots, J)$, where $\boldsymbol{\Omega}_{ij} = (\theta_i, a_j, b_j, c_j, \gamma_j)'$. Let $\{\boldsymbol{\Omega}^{(1)}, \ldots, \boldsymbol{\Omega}^{(R)}\}$ denote an MCMC sample from the full conditional posterior distribution in (8) and (16)–(18), where $\boldsymbol{\Omega}^{(r)} = (\boldsymbol{\Omega}_{ij}^{(r)}, i = 1, \ldots, N, j = 1, \ldots, J)$ and $\boldsymbol{\Omega}_{ij}^{(r)} = (\theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, \gamma_j^{(r)})'$ 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

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

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 $\boldsymbol{\Omega}^{(r)}$ is given by

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

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:

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

where

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

In (21), $\overline{\text{Dev}(\boldsymbol{\Omega})}$ is a Monte Carlo estimate of the posterior expectation of the deviance function $\text{Dev}(\boldsymbol{\Omega}) = -2 \log L(\boldsymbol{Y} \mid \boldsymbol{\Omega})$, $\widehat{\text{Dev}(\boldsymbol{\Omega})}$ is an approximation of $\text{Dev}(\widehat{\boldsymbol{\Omega}})$, where $\widehat{\boldsymbol{\Omega}}$ is the posterior mode, when the prior is relatively non-informative, and $P_D = \overline{\text{Dev}(\boldsymbol{\Omega})} - \widehat{\text{Dev}(\boldsymbol{\Omega})}$ 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

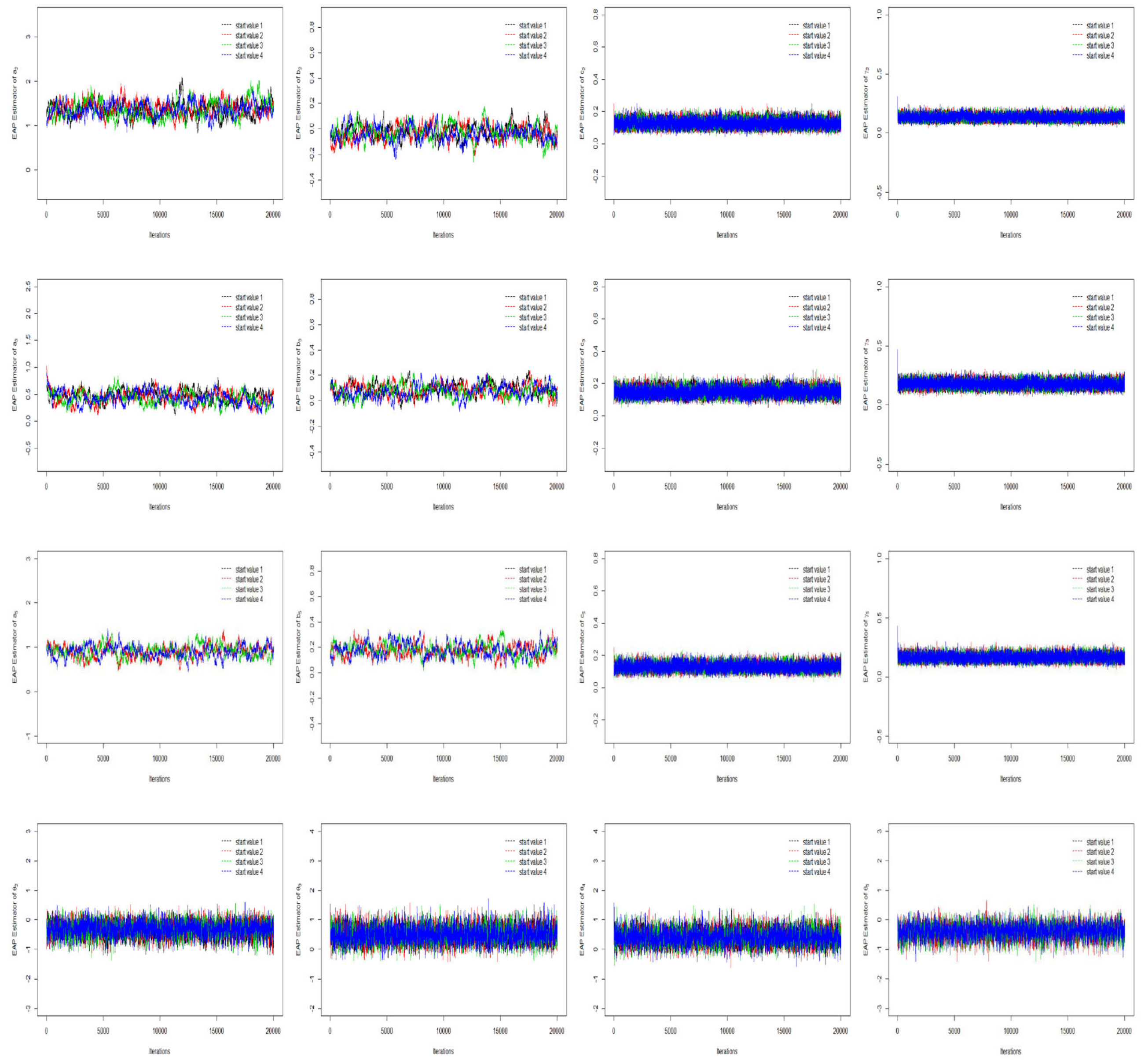

FIGURE 1 | The trace plots of three randomly selected items and persons for the Simulation Study 1.

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

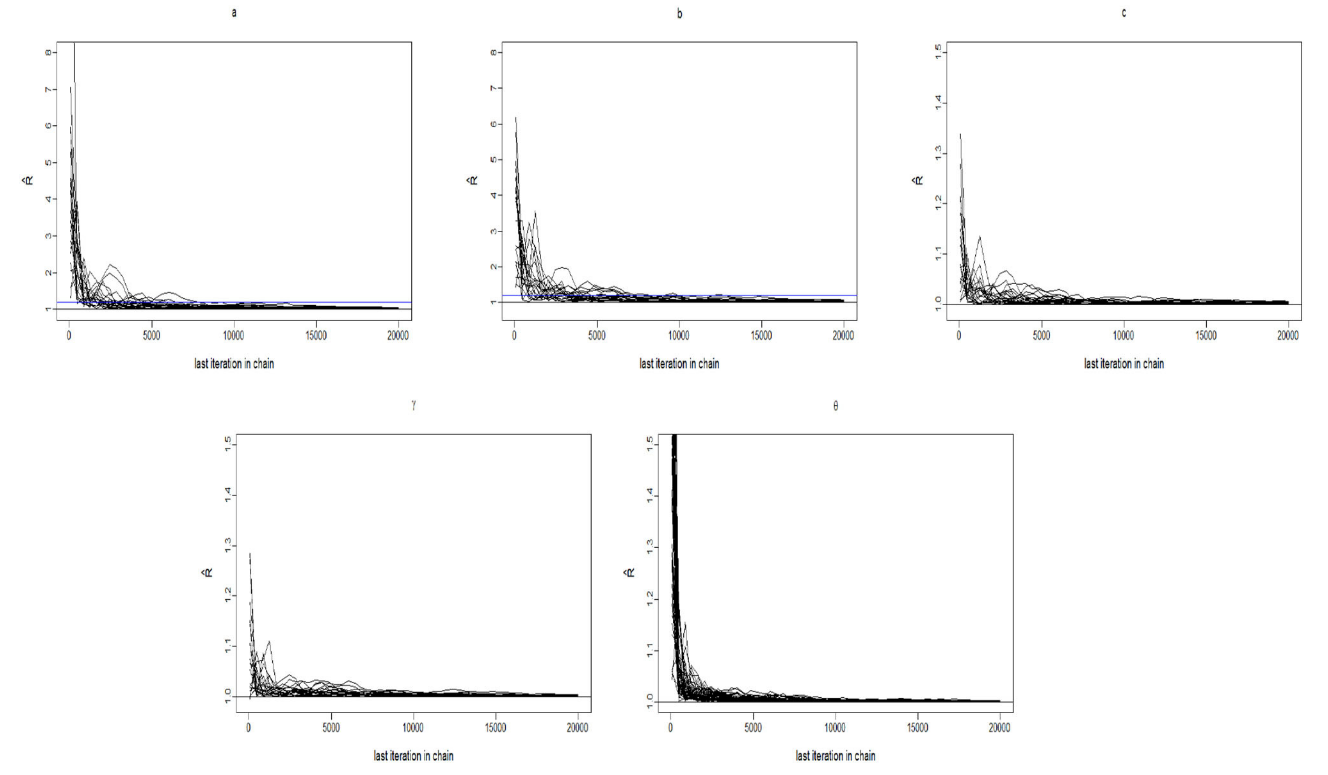

FIGURE 2 | The trace plots of $\widehat{R}$ for the Simulation Study 1.

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).

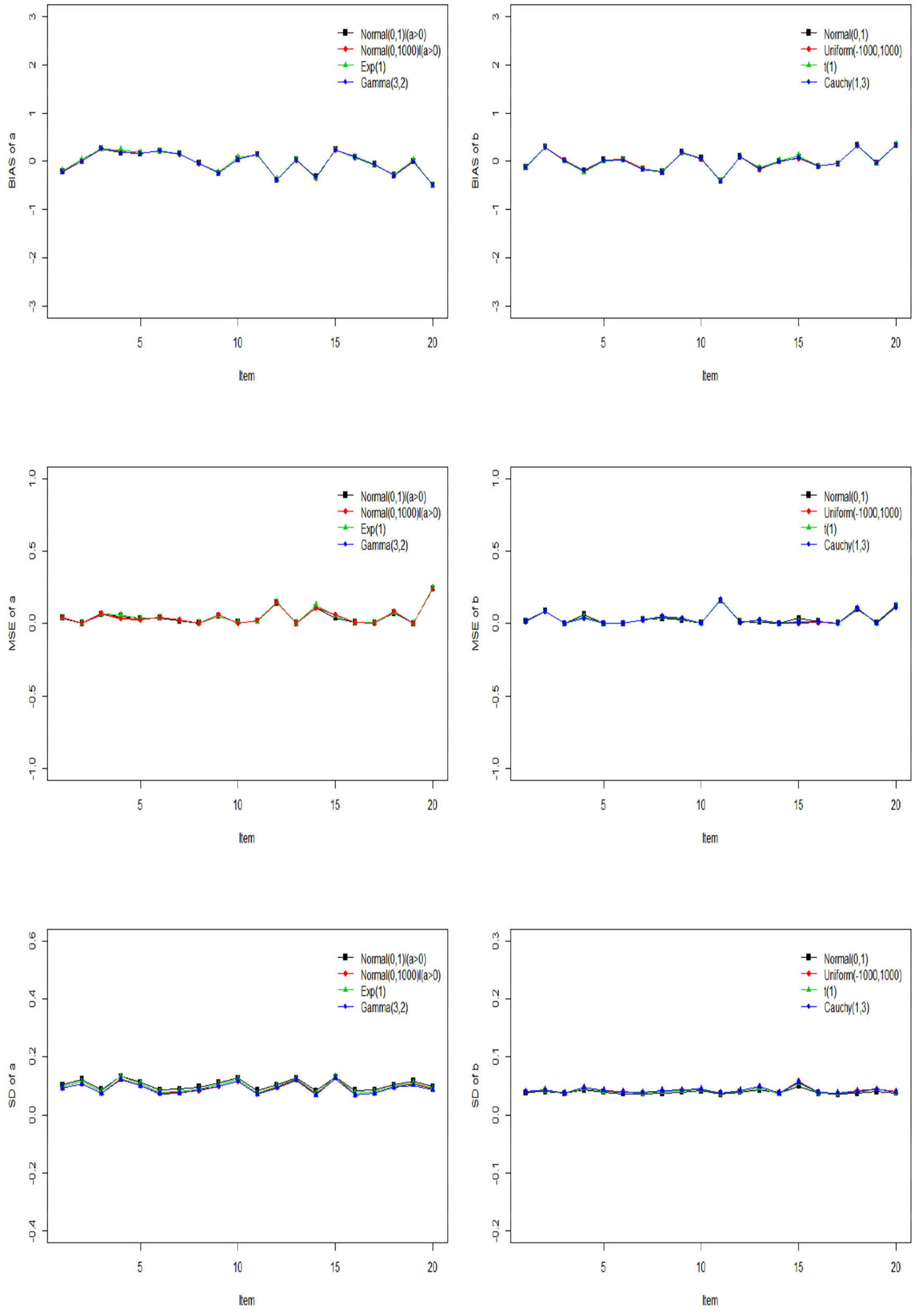

FIGURE 3 | The Bias, MSE, and SD of discrimination and difficulty parameters based on different priors.

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.

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.

- 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.

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.

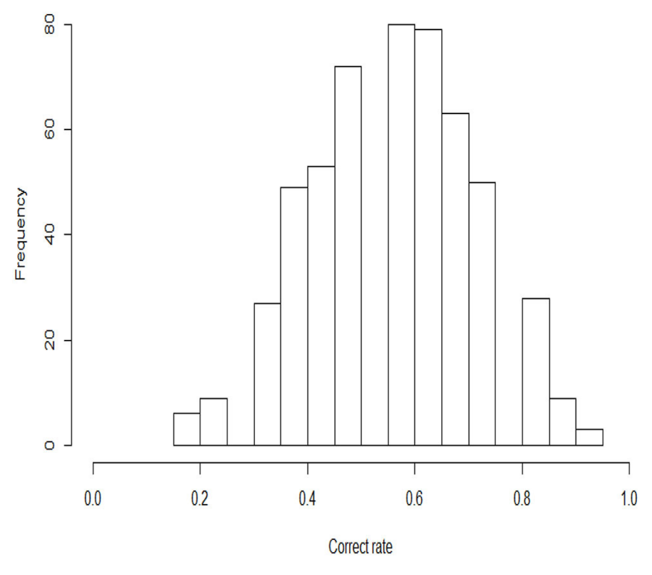

FIGURE 4 | Frequency histograms of the correct rates for 548 examinees.

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

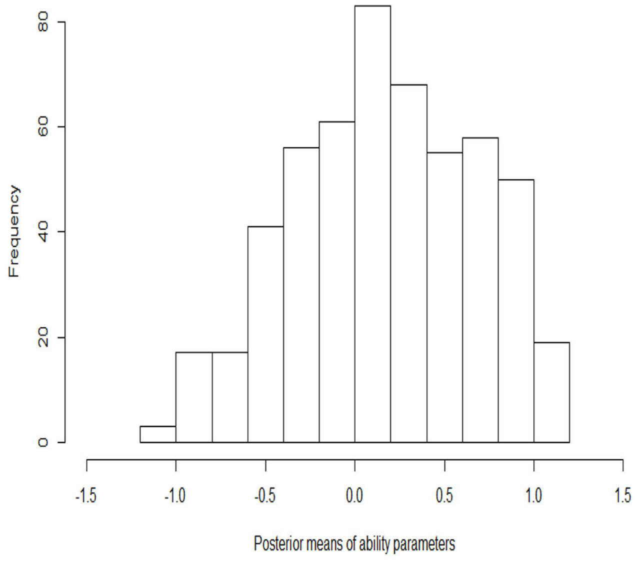

FIGURE 5 | The histograms of the posterior estimates of ability parameters.

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.

ORIGINAL RESEARCH
published: 03 June 2021
doi: 10.3389/fpsyg.2021.568348

Exploring Multiple Strategic Problem Solving Behaviors in Educational Psychology Research by Using Mixture Cognitive Diagnosis Model

Jiwei Zhang[1], *Jing Lu*[2]*, *Jing Yang*[3]*, *Zhaoyuan Zhang*[4] *and Shanshan Sun*[5]

[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] College of Mathematics, Taiyuan University of Technology, Jinzhong, China, [4] School of Mathematics and Statistics, Yili Normal University, Yili, China, [5] Government of Jilin Province, Changchun, China

A mixture cognitive diagnosis model (CDM), which is called mixture multiple strategy-Deterministic, Inputs, Noisy "and" Gate (MMS-DINA) model, is proposed to investigate individual differences in the selection of response categories in multiple-strategy items. The MMS-DINA model system is an effective psychometric and statistical approach consisting of multiple strategies for practical skills diagnostic testing, which not only allows for multiple strategies of problem solving, but also allows for different strategies to be associated with different levels of difficulty. A Markov chain Monte Carlo (MCMC) algorithm for parameter estimation is given to estimate model, and four simulation studies are presented to evaluate the performance of the MCMC algorithm. Based on the available MCMC outputs, two Bayesian model selection criteria are computed for guiding the choice of the single strategy DINA model and multiple strategy DINA models. An analysis of fraction subtraction data is provided as an illustration example.

Keywords: Bayesian inference, cognitive diagnosis, classification, Markov chain Monte Carlo, multiple-strategy models

OPEN ACCESS

Edited by:
Salvador Cruz Rambaud,
University of Almeria, Spain

Reviewed by:
Jung Yeon Park,
George Mason University,
United States
Steffen Zitzmann,
University of Tübingen, Germany

****Correspondence:***
Jing Lu
luj282@nenu.edu.cn
Jing Yang
yangj014@nenu.edu.cn

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

Received: *01 June 2020*
Accepted: *30 April 2021*
Published: *03 June 2021*

Citation:
Zhang J, Lu J, Yang J, Zhang Z and Sun S (2021) Exploring Multiple Strategic Problem Solving Behaviors in Educational Psychology Research by Using Mixture Cognitive Diagnosis Model. Front. Psychol. 12:568348. doi: 10.3389/fpsyg.2021.568348

1. INTRODUCTION

Multiple classification latent class models, namely cognitive diagnosis models (CDMs), have been developed specifically to diagnose the presence or absence of multiple fine-grained skills required for solving problems in an examination (Doignon and Falmagne, 1999; Junker and Sijtsma, 2001; Tatsuoka, 2002; de la Torre and Douglas, 2004; Templin and Henson, 2006; DiBello et al., 2007; Haberman and von Davier, 2007; de la Torre, 2009, 2011; Henson et al., 2009; von Davier, 2014; Chen et al., 2015). Compared with the traditional item response theory models, one of the advantages of multiple classification latent class models is that they can provide effective measurement of student learning and progression, design better teaching instruction, and conduct possibly intervention guidance for different individual and group needs.

However, most CDMs only consider the probability that examinees solve a problem in one way. In fact, examinees may solve a problem in different ways. Fuson et al. (1997) found that the children at elementary schools used more than one strategy to solve the problem of multi-digit addition and subtraction. Moreover, in eye-movement studies, Gorin (2007) expounded that the subjects often used very different cognitive strategies when solving similar reading tasks. More specifically, an example of a multiple-strategy used by de la Torre and Douglas (2008) in educational research is on the analysis of fraction subtraction data including responses of 2,144 examinees to 15 fraction subtraction items. The attributes required for the fraction subtraction are as follows: (a) performing basic fraction subtraction operation; (b) simplifying/reducing; (c) separating whole number from fraction; (d) borrowing one from whole number to fraction; (e) converting whole number to fraction; (f) converting mixed number to fraction; (g) column borrowing in subtraction (de la Torre and Douglas, 2008). As an illustration, they use two strategies to solve $4\frac{4}{12} - 2\frac{7}{12}$. Strategy 1 requires attributes a, b, c, and d. Strategy 2 requires attributes a, b, and f. The detailed calculation processes were shown in de la Torre and Douglas (2008).

de la Torre and Douglas (2008) proposed a multiple strategy-Deterministic, Inputs, Noisy "and" Gate (MS-DINA) model to address the problem of fraction subtraction, where the DINA model (Haertel, 1989; Doignon and Falmagne, 1999; Junker and Sijtsma, 2001; de la Torre and Douglas, 2004; de la Torre, 2009) was the most popular and widely used model among various CDMs which assumed that examinees were expected to answer an item correctly only when they possessed all the required attributes. The MS-DINA model is a straightforward extension of the DINA model that allows incorporating multiple strategies for cognitive diagnosis based on competing assumptions. However, as de la Torre and Douglas (2008) indicated, although the simplicity of the MS-DINA model was appealing, it made a restrictive assumption that the item parameters were same for different strategies, which implied that the application of each strategy was equally difficulty. Another limitation of MS-DINA model is that the joint distribution attributes is expressed as a function of a higher-order continuous ability. The joint distribution of the attributes as the most special form of the saturated model may not be applied to all cases (Huo and de la Torre, 2014). Moreover, the MS-DINA model cannot provide the information of the strategies selected by the examinees, that is, in the case that multiple strategies are available, the probability of each strategy being used cannot be obtained, and the strategy diagnosis for examinees is an important part in the multiple strategies cognitive diagnosis.

To maximize the diagnostic results of multiple-strategy (MS) assessment and overcome the limitation that assumes identical item parameters across strategies, in this paper, we propose a cognitive diagnosis framework for analyzing the MS data. Specifically, the framework describes a psychometric model that can exploit multiple-strategy information. The psychometric model is a multiple-strategy model called the mixture multiple-strategy DINA (MMS-DINA) model. The details of the framework are laid out in section 2. In section 3, MCMC algorithm is employed to estimate model parameters. In section 4, four simulation studies are used to evaluate the viability of the proposed framework and to simulate true testing conditions to evaluate the performance of the MCMC algorithm based on several different criteria. According to the available MCMC outputs, two Bayesian model selection criteria are computed to guide the choice of the single strategy DINA model and multiple strategy DINA models. An empirical example of fraction subtraction is used to illustrate the application of the proposed MMS-DINA model in section 5. The final section concludes the article with discussion and some directions for further research.

2. MODELS

2.1. Multiple-Strategy DINA Model

The MS-DINA model (de la Torre and Douglas, 2008; Huo and de la Torre, 2014) is a straightforward extension of the DINA model, which allows several different strategies of solution for each item. Let u_{ij} denote the observed item response for the ith examinee to response jth item, where $i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, J$, $u_{ij} = 1$, if the ith examinee correct answer the jth item, 0 otherwise. The ith examinee mastery attribute profile, $\boldsymbol{\alpha}_i$, can be represented by a vector of length K, that is, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}, \ldots, \alpha_{iK})^{'}$, where

$$\alpha_{ik} = \begin{cases} 1, & \text{the } i\text{th examinee masters the } k\text{th attribute;} \\ 0, & \text{otherwise.} \end{cases}$$

Suppose each item has as many as M distinct strategies that would suffice to solve it. A strategy is defined as a subset of the K attributes which could be used together to solve the item. This may be coded by constructing M different matrices, $Q_1, \ldots, Q_M$, and the element in the jth row and kth column of Q_m $(m = 1, 2, \ldots, M)$ is denoted as

$$q_{jkm} = \begin{cases} 1, & \text{if item } j \text{ requires skill } k \text{ of } m\text{th strategy} \\ 0, & \text{otherwise} \end{cases}$$

Let

$$\eta_{ijm} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jkm}}, m = 1, 2, \ldots, M.$$

The latent variable η_{ijm} denotes whether the examinee i has the all the required attributes to apply the mth strategy to the jth item. Let

$$\eta_{ij} = \max\left\{\eta_{ij1}, \eta_{ij2}, \ldots, \eta_{ijm}, \ldots, \eta_{ijM}\right\}.$$

The variable η_{ij} is 1 if examinee i satisfies the attribute requirements of at least one of the M strategies. Therefore, the item response function of the MS-DINA model is given as

$$p\left(u_{ij} = 1 \,|\boldsymbol{\alpha}_i\right) = \left(1 - s_j\right)^{\eta_{ij}} g_j^{1-\eta_{ij}}, \tag{1}$$

where the parameter s_j is the slipping parameter, which indicates the probability of slipping on the jth item when an examinee has mastered all the required attributes for at least one of the strategies. The parameter g_j is the guessing parameter, which denotes the probability of correctly answering the jth item when an examinee does not master all the required attributes for at least one of the strategies.

2.2. Mixture Multiple-Strategy DINA Model

We can see that the MS-DINA model assumes that the slipping and guessing parameters are the same for different strategies. The assumption that the application of each strategy has equally difficulty is too restrictive, as indicated by de la Torre and Douglas (2004). Then, de la Torre and Douglas (2008) tried and suggested a variant of the multiple-strategy model in order to break the limitation mentioned above. However, one of the issues they discussed is a feasible approach for estimating the parameters in their model can not be provided due to the necessary identifiability issues. Inspired by their thoughts, we propose a multiple-strategy model to overcome the limitation that assumes identical item parameters across strategies. One way to solve the problem is to use a discrete mixture model. Discrete mixture models assume that a data set is composed of distinct subpopulations of observations that are described by different parametric distributions (Titterington et al., 1985). Thus, a mixture multiple-strategy-DINA (MMS-DINA) model is proposed to allow for different strategies to be associated with different levels of difficulty. The item response function of the MMS-DINA model is given by

$$p\left(u_{ij}=1\left|\boldsymbol{\alpha}_i\right.\right)=\sum_{m=1}^{M}\pi_m p_{ijm}=\sum_{m=1}^{M}\pi_m\left(1-s_{jm}\right)^{\eta_{ijm}}g_{jm}^{1-\eta_{ijm}}, \tag{2}$$

swhere M is the number of strategy, p_{ijm} indicates the correct response probability that the ith examinee adopts the mth strategy to answer the jth item, and $\pi_m\ (m=1,2,\ldots,M)$ is a mixing proportion satisfying $\sum_{m=1}^{M}\pi_m=1$. In addition to the specific strategy, mixing proportion parameters are related to the distribution of $\boldsymbol{\alpha}$. The average value of latent attributes for all examinees ($\boldsymbol{\alpha}$) using strategy m is μ_m. The parameters s_{jm} and g_{jm} denote the slipping and guessing parameters for the mth strategy to the jth item, respectively. When the number of strategies is one (i.e., $M=1$), it is apparent that the MMS-DINA model in Equation (2) reduces to the DINA model.

3. BAYESIAN INFERENCES

3.1. Bayesian Estimation

Within a fully Bayesian framework, the Metropolis-Hastings within the Gibbs sampling algorithm (Geman and Geman, 1984; Casella and George, 1992; Chib and Greenberg, 1995; Gilks, 1996; Patz and Junker, 1999a,b) is used to estimate the model parameters. In fact, MCMC methods have been found to be particularly useful in estimating mixture distributions (Diebold and Robert, 1994), including mixtures that involve random effects within classes (Lenk and DeSarbo, 2000). A common MCMC strategy is to sample a class membership parameter for each observation at each stage of the Markov chain (Robert, 1996). For the current model, a strategy membership parameter, $c_i=1,2,\ldots,M$, is sampled for each examinee i along with a latent attribute parameter $\boldsymbol{\alpha}_i$. Then, the item response function of the MMS-DINA model in Equation (2) can be expressed as

$$p\left(u_{ij}=1\left|\boldsymbol{\alpha}_i,s_j,g_j\right.\right)=\sum_{m=1}^{M}p\left(c_i=m\right)\left(1-s_{jm}\right)^{\eta_{ijm}}g_{jm}^{1-\eta_{ijm}}, \tag{3}$$

where the latent variable c_i takes a value in the set $\{1,2,\ldots,M\}$ for the ith examinee, indicating which type of strategies the ith examinee uses.

The following prior distributions for $\boldsymbol{\pi}$, $\boldsymbol{c}$, $\boldsymbol{\alpha}$, $\boldsymbol{s}$, and $\boldsymbol{g}$ are used in conjunction with the MMS-DINA model, where $\boldsymbol{c}=\left(c_1,c_2,\ldots,c_N\right)$, $\boldsymbol{s}=\left(s_1,s_2,\ldots,s_J\right)$ and $\boldsymbol{g}=\left(g_1,g_2,\ldots,g_J\right)$,

$$\boldsymbol{\pi}=\left(\pi_1,\pi_2,\ldots,\pi_M\right)\sim Dirichlet\left(\beta_1,\beta_2,\ldots,\beta_M\right),$$

$$c_i\sim Multinominal\left(1\left|\pi_1,\pi_2,\ldots,\pi_M\right.\right),$$

$$\mu_m\sim Beta\left(\lambda_1,\lambda_2\right),$$

$$\left[\alpha_{ik}\left|c_i=m\right.\right]\sim Bernoulli\left(\mu_m\right),$$

$$s_{jm}\sim 4\text{-}Beta\left(v_s,t_s,a_s,b_s\right),$$

$$g_{jm}\sim 4\text{-}Beta\left(v_g,t_g,a_g,b_g\right).$$

Based on the results of de la Torre and Douglas (2004)'s research, we use the four-parameter *Beta* distribution as the prior distribution of slipping and guessing parameters. The four parameter *Beta* distribution, 4-*Beta* $\left(v,t,a,b\right)$, is a generalization of the *Beta* (v,t) distribution, and it has the interval $\left(a,b\right)$ rather than $(0,1)$ as its support set. Then, the joint posterior distribution can be written as

$$\begin{aligned}p\left(\boldsymbol{\alpha},\boldsymbol{s},\boldsymbol{g},\boldsymbol{\pi}\left|\boldsymbol{u}\right.\right)\propto&\left[\prod_{i=1}^{N}\prod_{j=1}^{J}\sum_{m=1}^{M}p\left(c_i=m\right)f\left(u_{ij}\left|\boldsymbol{\alpha}_i,s_{jm},g_{jm}\right.\right)\right]\\&\left[\prod_{i=1}^{N}p\left(\boldsymbol{\alpha}_i\left|\mu_m\right.\right)^{\mathrm{I}(c_i=m)}\right]\\&\times f_{prior}\left(\mu_m\right)\left[\prod_{m=1}^{M}\prod_{j=1}^{J}f_{prior}\left(s_{jm}\right)f_{prior}\left(g_{jm}\right)\right]\\&\prod_{m=1}^{M}f_{prior}\left(\pi_m\right),\end{aligned} \tag{4}$$

where $\boldsymbol{u}=\left(\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots,\boldsymbol{u}_i,\ldots,\boldsymbol{u}_N\right)'$ and $\boldsymbol{u}_i=\left(u_{i1},u_{i2},\ldots,u_{iJ}\right)$.

The MCMC sampling procedure is composed of the following steps:

Step 1: Sample the mixing proportions $\boldsymbol{\pi}=\left(\pi_1,s\pi_2,\ldots,\pi_M\right)'$. Assuming conditional independence between the mixing proportions and all parameters except the strategy memberships of examinees, the mixing proportions have a full condition posterior distribution of the form:

$$p\left(\boldsymbol{\pi}\left|\text{all other parameters}\right.\right)\propto p\left(c\left|\boldsymbol{\pi}\right.\right)f_{prior}\left(\boldsymbol{\pi}\right), \tag{5}$$

where n_m is the number of examinees using strategy m. This full conditional distribution is

$$Dirichlet\left(\beta_1 + n_1, \beta_2 + n_2, \ldots, \beta_M + n_M\right).$$

Step 2: Sample a strategy membership c_i for each examinee, where $i = 1, \ldots, M$. Assuming independence of examinees, the full condition posterior distribution of c_i can be written as

$$\begin{aligned} p\left(c_i = m \left| \text{all other parameters}\right.\right) &\propto p\left(\boldsymbol{u}_i \left| c_i = m, \boldsymbol{\alpha}_i, s_m, g_m\right.\right) \\ &\quad p\left(\boldsymbol{\alpha}_i \left| \mu_m, c_i = m\right.\right) \\ &\propto \left[\prod_{j=1}^{J} p_{ijm}^{u_{ij}} \left(1 - p_{ijm}\right)^{1-u_{ij}}\right] \\ &\quad \prod_{k=1}^{K} Bernoulli\left(\alpha_{ik}; \mu_m\right) \pi_m, \end{aligned} \tag{6}$$

where $\boldsymbol{u}_i = \left(u_{i1}, \ldots, u_{iJ}\right)'$ is the item response vector for examinee i across items, J and K are respectively the numbers of item and attribute, and $Bernoulli\left(\alpha_{ik}; \mu_m\right)$ is the *Bernoulli* density evaluated at α_{ik} with parameter μ_m.

Step 3: Sample attribute mean μ_m for each strategy. Assuming the attribute distribution parameters are independent of all parameters expect the attribute vectors for examinees in mth strategy, the full conditional distribution of μ_m can be written as

$$p\left(\mu_m \left| \text{all other parameters}\right.\right) \propto \left[\prod_{i=1}^{N} p\left(\boldsymbol{\alpha}_i \left| \mu_m\right.\right)^{\mathrm{I}(c_i=m)}\right] f_{prior}\left(\mu_m\right), \tag{7}$$

which results in the following full conditional distribution for μ_m:

$$\begin{aligned} \mu_m \sim Beta\Bigg(&\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_{ik} \mathrm{I}\left(c_i = m\right) + \lambda_1, (N \times K) \\ &- \sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_{ik} \mathrm{I}\left(c_i = m\right) + \lambda_2\Bigg). \end{aligned} \tag{8}$$

where $\mathrm{I}(\cdot)$ denotes the indicator function. $\mathrm{I}\left(c_i = m\right) = 1$ if the ith examinee choose the mth strategy to answer the item, 0 otherwise.

Step 4: Sample a latent variable $\boldsymbol{\alpha}_i$ for each examinee, where $i = 1, \ldots, N$. Assuming independence of examinees, the full conditional distribution of $\boldsymbol{\alpha}_i$ can be written as

$$\begin{aligned} p\left(\boldsymbol{\alpha}_i \left| \text{all other parameters}\right.\right) &\propto p\left(\boldsymbol{u}_i \left| c_i = m, \boldsymbol{\alpha}_i, \boldsymbol{s}_m, \boldsymbol{g}_m\right.\right) \\ &\quad p\left(\boldsymbol{\alpha}_i \left| \mu_m, c_i = m\right.\right) \\ &\propto \left[\prod_{j=1}^{J} p_{ijm}^{u_{ij}} \left(1 - p_{ijm}\right)^{1-u_{ij}}\right] \\ &\quad \prod_{k=1}^{K} Bernoulli\left(\alpha_{ik}; \mu_m\right). \end{aligned} \tag{9}$$

Step 5: Sample item parameters s_{jm} and g_{jm} for each strategy and each item. Assuming conditional independence across items, the full conditional distribution of s_{jm} and g_{jm} can be written as

$$\begin{aligned} &p\left(s_{jm}, g_{jm} \left| \text{all other parameters}\right.\right) \\ &\quad \propto \left[\prod_{i=1}^{N} p\left(\boldsymbol{u}_j \left| c_i = m, \boldsymbol{\alpha}_i, s_{jm}, g_{jm}\right.\right)\right] f_{prior}\left(s_{jm}\right) f_{prior}\left(g_{jm}\right) \\ &\quad \propto \left\{\prod_{i=1}^{N} \left[p_{ijm}^{u_{ij}} \left(1 - p_{ijm}\right)^{1-u_{ij}}\right]^{\mathrm{I}(c_i=m)}\right\} \left[Beta\left(s_{jm}; \nu_s, t_s, a_s, b_s\right)\right] \\ &\quad \times \left[Beta\left(g_{jm}; \nu_g, t_g, a_g, b_g\right)\right], \end{aligned} \tag{10}$$

where $\boldsymbol{u}_j = \left(u_{1j}, \ldots, u_{Nj}\right)'$ is the item response vector for item j across examinees, N is the number of examinees.

3.2. Bayesian Model Assessment

Within the Bayesian framework, the deviance information criterion (DIC; Spiegelhalter et al., 2002) and the logarithm of the pseudo-marignal likelihood (LPML; Geisser and Eddy, 1979; Ibrahim et al., 2001) are considered to compare three different models (the DINA model, the MS-DINA model, and the MMS-DINA model). As an explanation, we only provide the most complicated calculation process of DIC and LPML in the MMS-DINA model, and the calculation formulas of DIC and LPML for the DINA model and MS-DINA model are similar. These two criteria are based on the log-likelihood functions evaluated at the posterior samples of model parameters. Therefore, the DIC and LPML of the MMS-DINA model can be easily computed. Let $\boldsymbol{\Omega} = \left(\boldsymbol{\Omega}_{ij}, i = 1, \ldots, N, j = 1, \ldots, J, m = 1, \ldots, M\right)$, where $\boldsymbol{\Omega}_{ijm} = \left(\boldsymbol{\alpha}_i, s_{jm}, g_{jm}, \pi_m\right)'$. Let $\left\{\boldsymbol{\Omega}^{(1)}, \ldots, \boldsymbol{\Omega}^{(R)}\right\}$, where $\boldsymbol{\Omega}^{(r)} = \left(\boldsymbol{\Omega}_{ijm}^{(r)}, i = 1, \ldots, N, j = 1, \ldots, J, m = 1, \ldots, M\right)$, $\boldsymbol{\Omega}_{ijm}^{(r)} = \left(\boldsymbol{\alpha}_i^{(r)}, s_{jm}^{(r)}, g_{jm}^{(r)}, \pi_m^{(r)}\right)'$ for $i = 1, \ldots, N$, $j = 1, \ldots, J$, $m = 1, \ldots, M$ and $r = 1, \ldots, R$, which denotes rth MCMC sample from the posterior distribution in (4). The joint likelihood function of the responses can be written as

$$L\left(\boldsymbol{u} \left| \boldsymbol{\Omega}\right.\right) = \prod_{i=1}^{N}\prod_{j=1}^{J}\sum_{m=1}^{M} \pi_m p\left(u_{ij} \left| \boldsymbol{\alpha}_i, s_{jm}, g_{jm}\right.\right), \tag{11}$$

where $p\left(u_{ij} \left| \boldsymbol{\alpha}_i, s_{jm}, g_{jm}\right.\right)$ is the response probability. The logarithm of the joint likelihood function in (11) evaluated at $\boldsymbol{\Omega}^{(r)}$ is given by

$$\log L\left(\boldsymbol{u} \left| \boldsymbol{\Omega}^{(r)}\right.\right) = \sum_{i=1}^{N}\sum_{j=1}^{J} \log \sum_{m=1}^{M} \pi_m^{(r)} p\left(u_{ij} \left| \boldsymbol{\alpha}_i^{(r)}, s_{jm}^{(r)}, g_{jm}^{(r)}\right.\right). \tag{12}$$

Since the joint log-likelihoods for the responses, $\log \sum_{m=1}^{M} \pi_m^{(r)} p\left(u_{ij} \left| \boldsymbol{\alpha}_i^{(r)}, s_{jm}^{(r)}, g_{jm}^{(r)}\right.\right)$, $i = 1, \ldots, N, j = 1, \ldots, J$, and $m = 1, \ldots, M$ are readily available from MCMC sampling outputs,

$\log \sum_{m=1}^{M} \pi_m^{(r)} p\left(u_{ij} \middle| \boldsymbol{\alpha}_i^{(r)}, s_{jm}^{(r)}, g_{jm}^{(r)}\right)$ in (12) is easy to compute. Now, we calculate DIC as follows

$$\text{DIC} = \widehat{\text{Dev}(\boldsymbol{\Omega})} + 2P_D = \widehat{\text{Dev}(\boldsymbol{\Omega})} + 2\left[\overline{\text{Dev}(\boldsymbol{\Omega})} - \widehat{\text{Dev}(\boldsymbol{\Omega})}\right], \quad (13)$$

where

$$\overline{\text{Dev}(\boldsymbol{\Omega})} = -\frac{2}{R}\sum_{r=1}^{R} \log L\left(\boldsymbol{u} \middle| \boldsymbol{\Omega}^{(r)}\right) \text{ and } \widehat{\text{Dev}(\boldsymbol{\Omega})}$$
$$= -2 \max_{1 \le r \le R} \log L\left(\boldsymbol{u} \middle| \boldsymbol{\Omega}^{(r)}\right).$$

In (13), $\overline{\text{Dev}(\boldsymbol{\Omega})}$ is a Monte Carlo estimate of the posterior expectation of the deviance function $\text{Dev}(\boldsymbol{\Omega}) = -2\log L(\boldsymbol{u}\,|\boldsymbol{\Omega})$, $\widehat{\text{Dev}(\boldsymbol{\Omega})}$ is an approximation of $\text{Dev}(\widehat{\boldsymbol{\Omega}})$, where $\widehat{\boldsymbol{\Omega}}$ is the posterior mode, when the prior is relatively non-informative, and $P_D = \overline{\text{Dev}(\boldsymbol{\Omega})} - \widehat{\text{Dev}(\boldsymbol{\Omega})}$ is the effective number of parameters. Based on our construction, both DIC and P_D given in (13) are always non-negative. The model with a smaller DIC value fits the data better.

Letting $G_{ij,\max} = \max_{1 \le r \le R}\left[-\log \sum_{m=1}^{M} \pi_m^{(r)} p\left(u_{ij} \middle| \boldsymbol{\alpha}_i^{(r)}, s_{jm}^{(r)}, g_{jm}^{(r)}\right)\right]$, a Monte Carlo estimate of the conditional predictive ordinate (CPO; Gelfand et al., 1992; Chen et al., 2000) is given by

$$\log(\widehat{\text{CPO}_{ij}}) = -G_{ij,\max} - \log\left[\frac{1}{R}\sum_{r=1}^{R} \exp\left\{-\log \sum_{m=1}^{M} \pi_m^{(r)} p\left(u_{ij} \middle| \boldsymbol{\alpha}_i^{(r)}, s_{jm}^{(r)}, g_{jm}^{(r)}\right) - U_{ij,\max}\right\}\right]. \quad (14)$$

Note that the maximum value adjustment used in $\log(\widehat{\text{CPO}_{ij}})$ plays an important role in numerical stabilization in computing $\exp\left\{-\log \sum_{m=1}^{M} \pi_m^{(r)} p\left(u_{ij} \middle| \boldsymbol{\alpha}_i^{(r)}, s_{jm}^{(r)}, g_{jm}^{(r)}\right) - G_{ij,\max}\right\}$ in (14). A summary statistic of the $\widehat{\text{CPO}_{ij}}$ is the sum of their logarithms, which is called the LPML and given by

$$\text{LPML} = \sum_{i=1}^{N}\sum_{j=1}^{J} \log(\widehat{\text{CPO}_{ij}}). \quad (15)$$

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

3.3. The Accuracy Evaluation of Parameter Estimation

To implement the MCMC sampling algorithm, chains of length 10,000 with an initial burn-in period 5,000 are chosen. Fifty replications are used in the following simulation studies. Three indices are used to assess the accuracy of the parameter estimates. Let ϑ be the parameter of interest. Assume that $M = 50$ data sets are generated. Also, let $\widehat{\vartheta}^{(m)}$ and $\text{SD}^{(m)}(\vartheta)$ denote the posterior mean and the posterior standard deviation of ϑ obtained from the mth simulated data set for $m = 1, \ldots, M$.

The Bias for parameter ϑ is defined as

$$\text{Bias}(\vartheta) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\vartheta}^{(m)} - \vartheta\right), \quad (16)$$

and the mean squared error (MSE) for parameter ϑ is defined as

$$\text{MSE}(\vartheta) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\vartheta}^{(m)} - \vartheta\right)^2, \quad (17)$$

and the average of posterior standard deviation can be defined as

$$\text{SD}(\vartheta) = \frac{1}{M}\sum_{m=1}^{M} \text{SD}^{(m)}(\vartheta). \quad (18)$$

In addition, four criteria are used to assess the accuracy of the examinee classification methods. These criteria include the following: (h) the marginal correct classification rate for each attribute; (t) the proportion of examinees classified correctly for all K attributes; (v) the proportion of examinees classified correctly for at least $K - 1$ attributes; (z) the proportion of examinees classified incorrectly for $K - 1$ or K attributes.

4. SIMULATION

4.1. Simulation 1

This simulation study is conducted to evaluate the parameter recoveries of the proposed model using the MCMC algorithm as the number of examinees increases. Here, we fix the test length and the numbers of attributes.

4.1.1. Simulation Designs

The following manipulated conditions are considered. Test length is fixed at 20, and 2 strategies with 5 attributes are used in this simulation. The corresponding Q matrix of the 20 items is the same as de la Torre (2008, p. 605); and the number of examinees, $N = 500$, 1,000, and 2,000. Fully crossing different levels produce 3 simulation conditions (1 test length × 3 sample sizes). The true values of slipping and guessing parameters are set to be 0.3 and 0.1, respectively. Assuming independence among examinees and independence among attributes, the true value of α_{ik} is generated from *Bernoulli* (0.5). We can obtain a $N \times 5$ matrix $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_i, \ldots, \boldsymbol{\alpha}_N)'$, and the ith row vector $\boldsymbol{\alpha}_i$ denotes the ith examinee's true cognitive state. The hyper-parameters of the prior distributions are fixed as follows: $\beta_1 = \beta_2 = 0.01$, and $\lambda_1 = \lambda_2 = 0.5$. We assume the priors of the slipping and guessing parameters to follow a 4-*Beta* $(1, 2, 0.1, 0.5)$ based on de la Torre and Douglas (2004)'s paper. Response data are simulated using the MMS-DINA model. About 50 replications are considered to evaluate the parameters recovery in this simulation.

To evaluate the convergence of parameter estimations, we only consider the convergence in the case of minimum sample sizes. That is, the number of examinees is 500. Two methods are used to check the convergence of our algorithm.

TABLE 1 | Evaluating the accuracy of the item parameters based on different sample sizes in simulation study 1.

Sample	Strategy 1						Strategy 2					
	s_1			g_1			s_2			g_2		
size	ABias	AMSE	ASD	ABias	AMSE	ASD	ABias	AMSE	ASD	ABias	AMSE	ASD
500	0.101	0.010	0.057	0.078	0.011	0.021	0.100	0.010	0.057	0.101	0.014	0.022
1,000	0.086	0.007	0.048	0.063	0.009	0.018	0.097	0.009	0.049	0.091	0.010	0.018
2,000	0.079	0.006	0.044	0.058	0.008	0.016	0.089	0.008	0.046	0.083	0.006	0.015

Note that the ABias, AMSE, and ASD denote the average Bias, average MSE, and average SD for all item parameters.

One is the "eyeball" method to monitor the convergence by visually inspecting the history plots of the generated sequences (Hung and Wang, 2012), and the other method is to use the Gelman-Rubin method (Gelman and Rubin, 1992; Brooks and Gelman, 1998) to check the convergence of the parameters. The convergence of Bayesian algorithm is checked by monitoring the trace plots of the parameters for consecutive sequences of 10,000 iterations. The trace plots show that all parameter estimates converge quickly. We set the first 5,000 iterations as the burn-in period. In addition, the values of the potential scale reduction factor $\widehat{R}$ (PSRF; Brooks and Gelman, 1998) are calculated. We find the PSRF (Brooks and Gelman, 1998) values of all parameters are less than 1.2, which ensures that all chains converge as expected.

4.1.2. Recovery Results Based on Minimum Sample Sizes

As an illustration, we only show the Bias, MSE, and SD for all of the slipping and guessing parameters based on 500 examinees. In the case of the strategy 1, the Bias is between 0.083 and 0.110 for the slipping parameters and between 0.053 and 0.096 for the guessing parameters. The MSE is between 0.007 and 0.019 for the slipping parameters and between 0.004 and 0.013 for the guessing parameters. The SD are about 0.057 and 0.020 for the slipping and guessing parameters. In the case of the strategy 2, the Bias is between 0.087 and 0.107 for the slipping parameters and between 0.069 and 0.114 for the guessing parameters. The MSE is between 0.007 and 0.011 for the slipping parameters, between 0.006 and 0.018 for the guessing parameters. The SDs are about 0.057 and 0.022 for the slipping and guessing parameters.

We consider the criteria (h) in this simulation study, and the results show that the marginal correct classification rates are consistently high for the MMS-DINA model. Based on the criteria (t) through (z), we find that the MMS-DINA model consistently classifies examinees correctly high at least $K - 1$ attributes and produces few severe misclassifications. Thus, the classification method on the MMS-DINA model is effective.

4.1.3. Item Parameters Recovery Based on Different Sample Sizes

Given the total test length, when the number of individuals increases from 500 to 2,000, the average Bias, MSE, and SD for slipping and guessing parameters decrease. For example, under the first strategy, the average Bias of all slipping parameters decreases from 0.101 to 0.079, the average MSE of all slipping parameters decreases from 0.010 to 0.006, and the average SD of all slipping parameters decreases from 0.057 to 0.044. The average Bias of all guessing parameters decreases from 0.078 to 0.058, the average MSE of all guessing parameters decreases from 0.011 to 0.008, and the average SD of all guessing parameters decreases from 0.021 to 0.016. The evaluation results of the accuracy of item parameter estimation for different numbers of examinees are given in **Table 1**. We find that as the number of individuals increases, the estimates of item parameters become more accurate. In summary, the estimation of this algorithm is effective and accurate under the condition of simulation study 1.

4.2. Simulation 2

This simulation study is conducted to assess the parameter recoveries of the proposed model using the MCMC algorithm as the number of items increases. Here, we fix the sample size and the numbers of attributes.

4.2.1. Simulation Designs

The following manipulated conditions are considered. The number of examinees is fixed at 1,000, and the number of items, $J = 20$ or 30. Two strategies with five attributes are considered in this simulation. The corresponding Q matrix of the 20 items is the same as de la Torre (2008, p. 605), and the Q matrix of the 30 items is shown in **Table 2**. Fully crossing different levels have two conditions (2 test lengths × 1 sample size).

The true values and prior distributions for the parameters are the same as the simulation 1. To implement the MCMC sampling algorithm, chains of length 10,000 with an initial burn-in period 5,000 are chosen. Fifty replications are considered in this simulation. The following conclusions can be obtained. Given the total number of examinees, when the number of items increases from 20 to 30, the average Bias, MSE, and SD for slipping and guessing parameters increase. For example, for the first strategy, the average Bias of all slipping parameters increases from 0.086 to 0.093, the average MSE of all slipping parameters increases from 0.007 to 0.009, and the average SD of all slipping parameters increases from 0.048 to 0.051. The average Bias of all guessing parameters increases from 0.063 to 0.087, the average MSE of all guessing parameters increases from 0.009 to 0.014, and the average SD of all guessing parameters increases from 0.018 to 0.023. The evaluation results of the accuracy of item parameter estimation for different numbers of items are specified in **Table 3**.

TABLE 2 | The *Q* matrix design in simulation 2.

Item	Attribute									
	Strategy A					Strategy B				
1	1	1	0	0	0	0	1	0	1	1
2	1	0	1	0	0	0	0	1	1	1
3	1	0	0	1	0	0	1	1	0	1
4	1	0	0	0	1	0	1	1	1	0
5	0	1	1	0	0	1	0	0	1	1
6	0	1	0	1	0	1	1	0	0	1
7	0	1	0	0	1	1	1	0	1	0
8	0	0	1	1	0	1	0	1	0	1
9	0	0	1	0	1	1	0	1	1	0
10	0	0	0	1	1	1	1	1	0	0
11	1	1	1	0	0	0	0	0	1	1
12	1	1	0	1	0	0	1	0	0	1
13	1	1	0	0	1	0	1	0	1	0
14	1	0	1	1	0	0	0	1	0	1
15	1	0	1	0	1	0	0	1	1	0
16	1	0	0	1	1	0	1	1	0	0
17	0	1	1	1	0	1	0	0	0	1
18	0	1	1	0	1	1	0	0	1	0
19	0	1	0	1	1	1	1	0	0	0
20	0	0	1	1	1	1	0	1	0	0
21	1	1	0	0	0	1	0	0	0	0
22	1	0	1	0	0	0	1	0	0	0
23	1	0	0	1	0	0	0	1	0	0
24	1	0	0	0	1	0	0	0	1	0
25	0	1	1	0	0	0	0	0	0	1
26	1	0	0	0	0	0	1	1	0	0
27	0	1	0	0	0	1	0	0	0	1
28	0	0	1	0	0	1	0	0	1	0
29	0	0	0	1	0	1	1	0	0	0
30	0	0	0	0	1	1	0	1	0	0

TABLE 3 | Evaluating the accuracy of the item parameters based on different numbers of items in simulation study 2.

	Strategy 1					
	s_1			g_1		
Test length	ABias	AMSE	ASD	ABias	AMSE	ASD
20	0.086	0.007	0.048	0.063	0.009	0.018
30	0.093	0.009	0.051	0.087	0.014	0.023
	Strategy 2					
	s_2			g_2		
Test length	**ABias**	**AMSE**	**ASD**	**ABias**	**AMSE**	**ASD**
20	0.097	0.009	0.049	0.091	0.010	0.018
30	0.106	0.012	0.051	0.103	0.016	0.023

Note that the ABias, AMSE, and ASD denote the average Bias, average MSE, and average SD for all item parameters.

4.3. Simulation 3

This simulation study is conducted to evaluate the recoveries of the proposed model using the MCMC algorithm as the number of attributes increases. Here, the sample size and the test length are fixed.

4.3.1. Simulation Designs

The following manipulated conditions are considered. The number of examinees is fixed at 1,000, and the number of items is fixed at 40, that is, $J = 40$. Two strategies with seven attributes are considered in this simulation. The corresponding Q matrix of the 40 items is shown in **Table 4**. The true values and prior distributions for the parameters are the same as the simulation 1. To implement the MCMC sampling algorithm, chains of length 10,000 with an initial burn-in period 5,000 are chosen. Fifty replications are considered in this simulation. The recovery results of item parameters are shown in **Table 5**.

We find that when the number of attributes increases, the maximums of the average Bias, MSE, and SD for all of the slipping parameters are 0.085, 0.011, and 0.039, respectively, and the maximums of the average Bias, MSE, and SD for all of the guessing parameters are 0.097, 0.019, and 0.021, respectively. In summary, it is found that the MCMC algorithm can provide accurate parameters and can be used to guide practice through the three different simulation studies.

4.4. Simulation 4

In this simulation study, we use the DIC and LPML model assessment criteria to evaluate model fitting.

4.4.1. Simulation Designs

In this simulation, the number of examinees $N = 1,000$ is considered and the test length is fixed at 20. The Q matrix from de la Torre (2008, p. 605)'s paper is used in this simulation study. Three cognitive diagnosis models will be considered. That is, the DINA model, the MS-DINA model, and the MMS-DINA model. Therefore, we evaluate the model fitting in the following three cases.

Case 1: True model: DINA model vs. Fitted model: DINA model, MS-DINA model, and MMS-DINA model;
Case 2: True model: MS-DINA model vs. DINA model, MS-DINA model, and MMS-DINA model;
Case 3: True model: MMS-DINA model vs. Fitted model: DINA model, MS-DINA model, and MMS-DINA model.

The true values and prior distributions for the parameters are the same as the simulation 1. To implement the MCMC sampling algorithm, chains of length 10,000 with an initial burn-in period 5,000 are chosen. The results of the Bayesian model assessment based on the 50 replications are shown in **Table 6**. Note that the following results of DIC and LPML are based on the average of 50 replications.

From **Table 6**, we find that when the DINA model is the true model, the DINA model fits the data best as we expected. The average DIC and LPML for the DINA model are 17605.31

TABLE 4 | The Q matrix design in simulation 3.

Item	Attribute													
	Strategy A							Strategy B						
1	1	1	1	0	0	0	0	1	1	0	0	0	0	0
2	0	1	0	0	0	0	1	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0	1	0	0	1	0	0	0
4	1	0	0	1	0	0	0	1	0	0	0	1	0	0
5	0	0	1	0	1	0	0	0	0	0	1	0	0	0
6	0	0	0	0	0	1	0	1	0	0	0	0	0	1
7	1	0	0	0	0	0	1	0	0	0	1	1	1	0
8	0	1	0	1	0	0	0	1	0	0	0	0	0	0
9	1	0	1	0	0	0	0	1	0	1	0	0	0	0
10	0	0	0	1	0	0	0	0	0	1	0	0	0	0
11	1	0	0	0	1	0	0	1	0	0	0	0	1	0
12	0	0	1	0	0	1	0	0	0	0	0	1	0	0
13	0	0	0	0	0	0	1	0	0	0	0	0	1	0
14	0	1	1	0	0	0	0	0	0	0	0	0	0	1
15	1	1	0	0	0	0	0	1	1	1	0	0	0	0
16	0	1	0	0	1	0	0	1	1	0	1	0	0	0
17	0	1	0	0	0	0	0	1	1	0	0	1	0	0
18	0	0	1	1	0	0	0	1	1	0	0	0	1	0
19	0	0	0	0	1	0	0	0	0	1	0	0	1	1
20	1	0	0	0	0	1	0	0	1	1	1	0	0	0
21	0	0	1	0	0	0	1	0	0	1	0	1	0	1
22	0	0	0	1	1	0	0	1	1	0	0	0	0	1
23	0	0	0	1	0	1	0	0	1	1	0	0	0	1
24	0	0	0	1	0	1	0	0	0	1	1	1	0	0
25	0	0	0	1	0	0	1	0	0	0	1	0	1	1
26	0	0	0	0	1	1	0	0	1	1	0	0	0	0
27	0	0	0	0	1	0	1	0	0	1	0	0	1	0
28	0	0	0	0	0	1	1	0	1	0	0	1	0	0
29	1	0	0	0	0	0	0	0	1	0	0	0	0	1
30	1	1	0	1	0	0	0	0	0	1	1	0	0	0
31	1	1	0	0	1	0	0	0	0	1	0	1	0	0
32	1	1	0	0	0	1	0	0	1	0	1	0	0	0
33	1	1	0	0	0	0	1	0	0	0	1	0	1	0
34	0	1	1	1	0	0	0	0	0	0	1	1	0	0
35	0	0	1	0	1	0	1	0	0	0	1	0	1	0
36	0	0	1	0	0	1	1	0	0	1	0	0	0	1
37	0	1	1	0	0	0	1	0	0	0	1	0	0	1
38	0	0	1	1	1	0	0	0	0	0	0	1	1	0
39	0	0	0	1	0	1	1	0	0	0	0	1	0	1
40	0	0	0	1	1	1	0	0	0	0	0	0	1	1

and −9544.81. The second best fitting model is the MMS-DINA model. The differences between DINA model and MMS-DINA model in the average DIC and LPML are −2708.68 and 791.37, respectively. The differences between DINA model and MS-DINA model in the average DIC and LPML are −3316.56 and 2502.22, respectively. This indicates that the MMS-DINA model is more sufficient fitting compared with the MS-DINA model if the data are generated from a simple DINA model. When the MS-DINA model is the true model, the MS-DINA fitting the data generated from the MS-DINA is better than the DINA model and the MMS-DINA model. The DINA model is worst model. The differences between MS-DINA model and MMS-DINA model in the average DIC and LPML are −36.04 and 1474.36, respectively, and the differences between MS-DINA model and DINA model in the average DIC and LPML are −4452.13 and 2081.16, respectively. When the MMS-DINA is the true model, the average DIC difference between MMS-DINA model and MS-DINA (DINA) model is

TABLE 5 | Evaluating the accuracy of the item parameters when the examined attributes increase.

	Strategy 1					
	s_1			g_1		
Test length × Examinee	**ABias**	**AMSE**	**ASD**	**ABias**	**AMSE**	**ASD**
40 × 1000	0.077	0.008	0.033	0.095	0.015	0.017
	Strategy 2					
	s_2			g_2		
Test length × Examinee	**ABias**	**AMSE**	**ASD**	**ABias**	**AMSE**	**ASD**
40 × 1000	0.085	0.011	0.039	0.097	0.019	0.021

Note that the ABias, AMSE, and ASD denote the average Bias, average MSE, and average SD for all item parameters.

TABLE 6 | The results of Bayesian model assessment in simulation 4.

True Model			DINA	MS-DINA	MMS-DINA
Fitted model	DINA	DIC	**17605.31**	20921.87	20313.99
		LPML	**−9544.81**	−12047.03	−10336.18
	MS-DINA	DIC	26998.25	**22546.12**	22582.16
		LPML	−13805.85	**−11724.69**	− 13199.05
	MMS-DINA	DIC	21264.30	21023.73	**19944.07**
		LPML	−11851.35	−11393.62	**−10126.66**

The meaning of the bold values is the model with the best-fitting data among the three candidate models.

about −1079.66 (−1320.23), and the average LPML difference between MMS-DINA model and MS-DINA (DINA) model is about 1266.96 (1724.69). This shows that when the data come from the mixture multiple strategy model, the DINA model with a single strategy is obviously ineffective in fitting this batch of data. The MS-DINA model has better fitting than the DINA model. No matter which models (DINA and MS-DINA) generate data, the MMS-DINA model is better fitting model than the other not true models. The MMS-DINA model is effective under many conditions of model fitting. In summary, the Bayesian assessment criterion is effective for identifying the true models, and it can be used in the subsequent real data study.

5. EMPIRICAL EXAMPLE ANALYSIS

5.1. Data

To study the applicability of the mixture multiple-strategy DINA model, we consider a real data including responses by 528 middle school students to answer 15 fraction subtraction items, which is a subset of the data originally used and described by Tatsuoka (2002). The *Q*-matrix design is given in de la Torre and Douglas (2008) research. Two strategies are considered to solve the 15 items, where the attribute definition is the same as in the introduction. The prior distributions described in the simulation section are used for the relevant parameters of the MMS-DINA model. Parameter estimates are based on averaging the estimates from 5 parallel chains with randomly chosen starting values. The standard deviations are obtained by averaging the sample SDs of the parameters from the separate chains. Each of these parallel chains is run for 10,000 iterations with the first 5,000 iterations as burn-in.

TABLE 7 | MMS-DINA model parameter estimates for the fraction subtraction data.

	Strategy 1				Strategy 2			
	s_{j1}		g_{j1}		s_{j2}		g_{j2}	
Item	**Estimate**	**SD**	**Estimate**	**SD**	**Estimate**	**SD**	**Estimate**	**SD**
1	0.13	0.02	0.12	0.05	0.13	0.01	0.11	0.01
2	0.12	0.01	0.22	0.03	0.14	0.02	0.18	0.02
3	0.10	0.02	0.20	0.02	0.11	0.02	0.12	0.02
4	0.11	0.03	0.18	0.02	0.13	0.02	0.11	0.02
5	0.20	0.01	0.25	0.03	0.13	0.01	0.23	0.01
6	0.21	0.02	0.10	0.01	0.17	0.02	0.12	0.02
7	0.10	0.01	0.13	0.01	0.13	0.02	0.12	0.02
8	0.10	0.02	0.21	0.03	0.12	0.03	0.14	0.03
9	0.10	0.01	0.19	0.02	0.12	0.01	0.14	0.01
10	0.15	0.03	0.17	0.01	0.12	0.02	0.11	0.02
11	0.15	0.01	0.19	0.02	0.11	0.03	0.12	0.03
12	0.16	0.02	0.12	0.03	0.13	0.01	0.11	0.01
13	0.13	0.03	0.15	0.00	0.15	0.02	0.12	0.02
14	0.15	0.01	0.11	0.01	0.13	0.01	0.12	0.01
15	0.17	0.01	0.11	0.02	0.11	0.02	0.11	0.02

5.2. Bayesian Model Assessment

Three comparative models, the DINA model, the MS-DINA model, and the MMS-DINA model, are used to fit the fraction subtraction data. The deviance information criterion (DIC; Spiegelhalter et al., 2002) and the logarithm of the pseudo-marignal likelihood (LPML; Geisser and Eddy, 1979; Ibrahim et al., 2001) are computed on the "CODA" R package (Plummer et al., 2006). Based on the comparable values of the DIC, that is, 5941.12 for the DINA model vs. 6652.13 (7306.29) for the MMS-DINA model (MS-DINA model). The LPMLs for the DINA model, MS-DINA model, and the MMS-DINA model are −2970.56, −3653.14, and −3326.06, respectively. The second best fitting model is also the MMS-DINA model. Based on the above model assessment results, we find that the DINA model fits the data most appropriately. The two multiple strategy models may show the over-fitting phenomenon, which results in that the data fitting is not as good as the simple DINA model. In addition, the MMS-DINA model is preferred for this data set because its relatively flexible formulation do not lead to worse fit compared with the MS-DINA model.

5.3. Results

The estimated posterior means and the SDs for the MMS-DINA model are shown in **Table 7**. The estimates of the slipping parameters range from 0.10 to 0.23 and the estimates of the guessing parameters range from 0.10 to 0.25. For the item 2, the students choose two strategies to answer the item, in which the first strategy examines four attributes (attributes 1, 2, 3, and 4), and the second strategy examines two attributes (attributes 1 and 6). We know that the more attributes an item measures, the lower the probability that the specific examinee will answer correctly. This is because the examinee can answer the item correctly if they have mastered all the attributes. If the examinee answers correctly the item with more attributes, the examinee is more likely to guess correctly the item. Therefore, for item 2, the estimate of the guessing parameter under the first strategy is 0.22, which is higher than the estimate of the guessing parameter under the second strategy is 0.18. Similarly, for item 4, the first strategy examines five attributes (attributes 1, 2, 3, 4, and 5) and the second strategy examines three attributes (attributes 1, 5, and 6). The corresponding estimates of guessing parameters are 0.18 and 0.12, respectively. When the number of attributes examine under the two strategies is the same, the estimates of the guessing parameters of the two strategies are basically the same. For example, four attributes are examined under both strategies for item 15. The probability of guessing under both strategies is the same as 0.11. In addition, the three items with the easiest slipping are items 6, 5, and 15 when using the strategy 1, and the corresponding estimates of the slipping parameters are 0.21, 0.20, and 0.17, respectively. When using the strategy 2, the three items with the easiest slipping are items 6, 13, and 2. The corresponding estimates of the slipping parameters are 0.17, 0.15, and 0.14, respectively.

In order to depict individual tendency of which strategy the examinees used, we use the probability plots of examinees choosing different strategies to show the selection tendency of all 528 examinees. In **Figure 1**, We find that 432 examinees use the first strategy to answer all 15 items. Compared to the first strategy, the number of examinees who adopt the second strategy is relatively small, only 96 examinees.

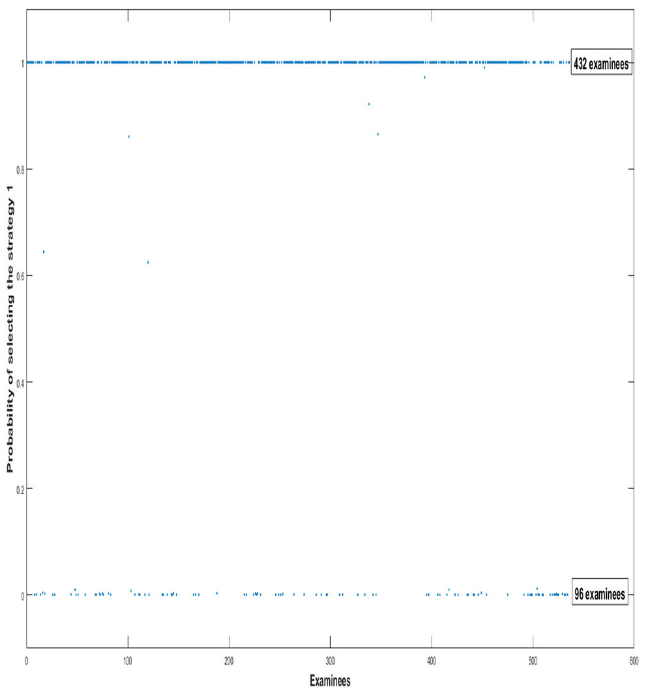

FIGURE 1 | The probabilities of examinees choosing different strategies. The y-axis indicates the probabilities of all examinees using the first strategy to answer items. 0 indicates that examinees use the first strategy to answer item with 0% probability, while 1 indicates that examinees use the first strategy to answer items with 100% probability.

6. CONCLUSIONS AND DISCUSSION

The goal of this article is to investigate a discrete mixture version of multiple-strategy model for cognitive diagnosis. A unique feature of the mixture model (MMS-DINA model) presented in this article is its capacity to break the limitation that assumes identical item parameters across strategies. The model-based approach presented in this article provides a natural generalization of the DINA model that allows it to account for the strategies to have different item parameters for each item. In the simulation studies, two simulation designs to examine the accuracy of the algorithm estimation from three different perspectives. The simulation results indicate that MCMC algorithm can be used to obtain accurate parameter estimates. Thus, this research provides researchers a tool that allows them to explore the practicability of the MMS-DINA model, which can in turn pave the way for the applications of CDMs in practical education settings to inform instruction and learning. In addition, two Bayesian model assessment criterion are considered to evaluate the model fitting among DINA model, MS-DINA model and MMS-DINA model. We find that when the data are generated from the simple single-strategy DINA model, the MMS-DINA model fits the data better than the MS-DINA model. This may be because each strategy is selected with a certain probability in the MMS-DINA model, unlike the MS-DINA model, which randomly chooses one strategy from multiple strategies. In this way, the Q matrix used in the MS-DINA model may be inconsistent with the Q matrix of the DINA model that generates data, resulting in the biased estimates and poor fitting. However, when the data are generated from MMS-DINA model, the DINA model is the worst fitting model. The worst fitting result is attributed to the relatively simple model structure, which leads to the phenomenon of under-fitting. Finally, we draw a valuable conclusion that no matter which models (DINA and MS-DINA) generate the data, the MMS-DINA is better fitting model than other not true models. However, in the real data analysis, the DINA model is preferred for this data set because its relatively simple formulation do not lead to worse fit compared with the MS-DINA model and MMS-DINA model.

Classification methods based on CDMs play an important role in cognitive diagnosis, because it is desired in some educational settings to classify examinees as masters or non-masters of multiple discrete latent attributes. In simulation study, as an illustration, we consider the MMS-DINA model is used in the situation that 500 examinees answer 20 items, indicating that it classifies few examinees correctly on all K skills but classifies high ability examinees almost or exactly correctly with few severe misclassification.

Because there are a large number of parameters in MMS-DINA model, we can only rely on MCMC algorithm to estimate the parameters. However, the computational burden of the MCMC algorithm becomes intensive especially when a large number of examinees or the items is considered, or a large number of the MCMC sample size is used. Therefore, it is desirable to develop a standing-alone R package associated with C++ or Fortran software for more extensive large-scale assessment program. In addition, the convergence of Bayesian algorithm need to be further investigated in the next studies. Firstly, for the PSRF value, we use a relatively relaxed 1.2 as a cutoff for determining the convergence of Bayesian estimation based on the previous literature (Brooks and Gelman, 1998; Fagua et al., 2019). In fact, we cannot decide whether 1.2 as a cutoff is really sufficient to determine the convergence. Educational psychologists have to be more careful when choosing 1.2 as a cutoff. This is because the effective sample size (ESS) can be only small, which can result in the summary statistics for the chain that provide only poor approximations of the Bayesian estimates. More specifically, the mean of the chain might not be very close to the expected value of the posterior distribution from the perspective of Bayesian point estimation. Therefore, in more substantive applications of the model, a more conservative PSRF cutoff (e.g., PSRF < 1.05) should ideally be used (Gelman et al., 2014; Vehtari et al., 2019; Zitzmann and Hecht, 2019). However, if we use a more conservative criterion for the PSRF, it is unknown how long it will take to achieve a PSRF of 1.05, and it will be a great challenge for our MMS-DINA model due to the large number of unknown parameters to be estimated. In order to achieve a cutoff of 1.05 for PSRF, we need to run a longer Markov chains to achieve the required number of ESS for convergence, but this process is very time-consuming and requires a large amount of computer memory. These require us to do a lot of simulation studies in later stages to give the definite results. Secondly, we also need to further investigate whether the obtained standard errors are accurate by using the coverage rate. However, these studies are beyond the purpose of this study to analyze the different solution strategies of the examinees by constructing a MMS-DINA model.

There are several avenues for further research on multiple-strategy models. In this paper, we focus on the comparison of multiple-strategy models under the most commonly used DINA model framework, and explore the cognitive process of solving items using different strategies among examinees, without focusing on the comparison of other multiple strategy cognitive diagnostic models, such as MS high-order DINA model, or some saturated type MS CDMs which are MS generalized DINA models, or MS loglinear cognitive model and so on. As Li et al. (2016) point out, it needs to be further explored to find the most appropriate model to fit data among the numerous cognitive diagnosis models. Therefore, in the later research, we will focus on the comparison of different MS CDMs to find out the advantages, disadvantages, and application scope of each model. In addition, the different classification methods may be helpful in both item selection and final examinee classification (Xu et al., 2003; Cheng, 2009). Also, note that a strategy is merely defined by the set of attributes required by a particular approach to solving a problem. One can imagine that a strategy might instead be determined by a set of attributes as well as a procedure and sequence for using them. So depending on how the attributes are defined, this will not always be the case, and one may consider different methods of using the same attributes. In addition, in this study, we only analyze two strategies. When the number of strategies increase, the performance of our MMS-DINA model needs to be further investigated. For example, we need to investigate that whether the identification conditions are satisfied as the number of strategies increases, as well as whether the parameter estimates are recovered well. In addition, the computational efficiency may be reduced due to the large number of parameters with the increased strategies.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://cran.r-project.org/web/packages/CDM/index.html.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 12001091). In addition, this work was also supported by the Fundamental Research Funds for the Central Universities of China (Grant No. 2412020QD025).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.568348/full#supplementary-material

REFERENCES

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

Casella, G., and George, E. I. (1992). Explaining the Gibbs sampler. *Am. Stat.* 46, 167–174. doi: 10.1080/00031305.1992.10475878

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York, NY: Springer. doi: 10.1007/978-1-4612-1276-8

Chen, Y., Liu, J., Xu, G., and Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *J. Am. Stat. Assoc.* 110, 850–866. doi: 10.1080/01621459.2014.934827

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika* 74, 619–632. doi: 10.1007/s11336-009-9123-2

Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *Am. Stat.* 49, 327–335. doi: 10.2307/2684568

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *J. Educ. Measur*. 45, 343–362. doi: 10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2009). DINA model and parameter estimation: a didactic. *J. Educ. Behav. Stat.* 34, 115–130. doi: 10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., and Douglas, J. (2004). Higher-order latent trait modelsss for cognitive diagnosis. *Psychometrika* 69, 333–353. doi: 10.1007/BF02295640

de la Torre, J., and Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: an analysis of fraction subtraction data. *Psychometrika* 73, 595–624. doi: 10.1007/s11336-008-9063-2

DiBello, L., Roussos, L., and Stout, W. (2007). 31A Review of cognitively diagnostic assessment and a summary of psychometric models 1,2. *Handb. Stat.* 26, 979–1030. doi: 10.1016/S0169-7161(06)26031-0

Diebold, J., and Robert, C. P. (1994). Estimation of finite mixture distribution through Bayesian sampling. *J. R. Stat. Soc. B* 56, 163–175. doi: 10.1111/j.2517-6161.1994.tb01985.x

Doignon, J. P., and Falmagne, J. C. (1999). *Knowledge Spaces*. New York, NY: Springer. doi: 10.1007/978-3-642-58625-5

Fagua, J. C., Baggio, J. A., and Ramsey, R. D. (2019). Drivers of forest cover changes in the Chocó-Darien Global Ecoregion of South America. *Ecosphere* 10:e02648. doi: 10.1002/ecs2.2648

Fuson, K. C., Smith, S., and Lo Cicero, A. (1997). Supporting Latino first graders' ten-structured thinking in urban classrooms. *J. Res. Math. Educ.* 28, 738–760. doi: 10.2307/749640

Geisser, S., and Eddy, W. F. (1979). A predictive approach to model selection. *J. Am. Stat. Assoc.* 74, 153–160. doi: 10.2307/2286745

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 4*, eds J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford: Oxford University Press), p. 147–167.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis, 3rd Edn*. Boca Raton, FL: CRC Press. doi: 10.1201/b16018

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.2307/2246093

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

Gilks, W. R. (1996). "Full conditional distributions," in *Markov Chain Monte Carlo in Practice*, eds W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Washington, DC: Chapman and Hall), 75–88. doi: 10.1007/978-1-4899-4485-6_5

Gorin, J. S. (2007). "Test cnstruction and diagnostic testing," in *Cognitive Diagnostic Assessment for Education: Theory and Applications*, eds J. P. Leighton and M. J. Gierl (New York, NY: Cambridge University Press), 173–201. doi: 10.1017/CBO9780511611186.007

Haberman, S. J., and von Davier, M. (2007). Some notes on models for cognitively based skill diagnosis. *Handb. Stat.* 26, 1031–1038. doi: 10.1016/S0169-7161(06)26040-1

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Meas.* 26:21. doi: 10.1111/j.1745-3984.1989.tb00336.x

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/s11336-008-9089-5

Hung, L.-F., and Wang, W.-C. (2012). The generalized multilevel facets model for longitudinal data. *J. Educ. Behav. Stat.* 37, 231–255. doi: 10.3102/1076998611402503

Huo, Y., and de la Torre, J. (2014). An EM algorithm for the multiple-strategy DINA model. *Appl. Psychol. Meas.* 38, 464–485. doi: 10.1177/0146621614533986

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York, NY: Springer. doi: 10.1002/0470011815.b2a11006

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

Lenk, P. J., and DeSarbo, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika* 65, 93–119. doi: 10.1007/BF02294188

Li, H., Hunter, C. V., and Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading test. *Lang. Test.* 33, 391–409. doi: 10.1177/0265532215590848

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

Patz, R. J., and Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *J. Educ. Behav. Stat.* 24, 342–366. doi: 10.3102/10769986024004342

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: convergence diagnosis and output analysis of MCMC. *R News* 6, 7–11. doi: 10.1159/000323281

Robert, C. P. (1996). "Mixtures of distributions: inference and estimation," in *Markov Chain Monte Carlo in Practice*, eds W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Washington, DC: Chapman & Hall), 75–88. doi: 10.1007/978-1-4899-4485-6_24

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 Stat Methodol.* 64, 583–639. doi: 10.1111/1467-9868.00353

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *J. R. Stat. Soc. Ser. C Appl. Stat.* 51, 337–350. doi: 10.1111/1467-9876.00272

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989x.11.3.287

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York, NY: Wiley. doi: 10.2307/2531224

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P. C. (2019). Rank-normalization, folding, and localization: an improved $\widehat{R}$ for assessing convergence of MCMC. *ArXiv[Preprint]* 1–27. doi: 10.1214/20-ba1221

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: two variants of a model equivalency. *Br. J. Math. Stat. Psychol.* 67, 49–71. doi: 10.1111/bmsp.12003

Xu, X., Chang, H., and Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. *Paper Presented at the Annual Meeting of the American Educational Research Association* (Chicago, IL).

Zitzmann, S., and Hecht, M. (2019). Going beyond convergence in Bayesian estimation: why precision matters too and how to assess it. *Struct. Equat. Model.* 26, 646–661. doi: 10.1080/10705511.2018.1545232

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 June 2021
doi: 10.3389/fpsyg.2021.532696

Price Attractiveness and Price Complexity: Why People Prefer Level-Payment Loans

Yang Lu, Jian Wang, Chenyang Li, Haoya Huang and Xintian Zhuang*

School of Business and Administration, Northeastern University, Shenyang, China

The improving sequence effect suggests that in choices between a rising earning and any other sequences, participants prefer the rising earning. Recent studies show that the improving sequence effect also exists in a loan context. As consumers have a strong preference for falling loan profiles, banks may consider to offer loans in which the loan repayments concentrate at the beginning of the loan term. In this paper, we examined the improving sequence effect in context of a car loan with three repayment plans expressed in temporally reframed prices (TRP). By regressing the evaluation of loan profiles on the perceived price attractiveness, price complexity, TRP and the interaction terms, we find that (1) the perceived price attractiveness and price complexity significantly predict the loan evaluation, and they also explain a significant proportion of variance in loan evaluation; (2) the TRP effect interacts with the improving sequence effect. Specifically, with the introduction of TRP, respondents prefer constant profiles over falling profiles. TRP may explain why level-payment loans are still popular in real world, though the improving sequence effect suggests otherwise.

Keywords: sequence effect, temporal reframing of price, q-exponential discount model, intertemporal choice, discounted utility model

OPEN ACCESS

Edited by:
Jason C. Immekus,
University of Louisville, United States

Reviewed by:
Keiko Ishii,
Nagoya University, Japan
Salvador Cruz Rambaud,
University of Almería, Spain

****Correspondence:***
Yang Lu
ylu@mail.neu.edu.cn

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

Received: *05 February 2020*
Accepted: *12 May 2021*
Published: *10 June 2021*

Citation:
Lu Y, Wang J, Li C, Huang H and Zhuang X (2021) Price Attractiveness and Price Complexity: Why People Prefer Level-Payment Loans. Front. Psychol. 12:532696. doi: 10.3389/fpsyg.2021.532696

INTRODUCTION

Firstly introduced by Samuelson (1937), the Discounted Utility Model (hereinafter, DUM) has been widely used to evaluate present utility of future rewards. This theory assumes that individuals evaluate future rewards based on the present value of the rewards by using an exponential discount function. According to the DUM, individuals would prefer falling sequences over rising sequences when evaluating positive future rewards, i.e., individuals prefer rewards received in an decreasing sequence rather than increasing, whilst the total amount of the rewards stays the same. This is because the rewards in a falling sequence concentrate at the beginning of the period, and thus have greater present value than that of a rising sequence of rewards with equal total amount. Similarly, by employing the DUM, we can also conclude that individuals prefer rising sequences over falling sequences if future outcomes are negative.

However, the preference for improvement contradicts the DUM. Loewenstein and Sicherman (1991) first found that when choosing between a falling sequence and a rising sequence of money, whilst the aggregate amount of money of the two sequences was the same, most people preferred the rising sequence. The preference for sequences of monetary rewards has been studied extensively. For positive series of future rewards such as incomes, restaurant visits, leisure activities or other gains, the preference for improvement means that individuals prefer to start with the least attractive outcome and end with the most attractive outcome than the opposite, i.e., they prefer the rising sequence over the falling sequence adding up to the same total amount (Loewenstein and Prelec, 1991, 1993; Loewenstein and Sicherman, 1991; Gigliotti and Sopher, 1997; Thaler, 1999; Matsumoto et al., 2000; Guyse et al., 2002; Duffy and Smith, 2013; Duxbury et al., 2013). Likewise, for negative series of outcomes such as pains, annoying noise, discomfort or other losses, individuals prefer the falling sequence over the rising sequence (Ariely and Loewenstein, 2000; Ariely and Zauberman, 2000; Langer et al., 2005; Rambaud et al., 2018; Garcia et al., 2020).

Some researchers examined human preferences for sequences with respect to loan repayment plans. Hassenzahl (2005) found a preference for decreasing loan profiles. Participants were requested to take out a loan for a vacation, and to choose between a profile starting with a large repayment followed by a series of small repayments, and a profile ending with the large repayment. The majority of respondents preferred an earlier large repayment. Hoelzl et al. (2011) viewed loan repayments as a sequence of installments that are either falling, rising or constant over time. The respondents preferred the falling repayment plan over other options, and they took out loans that contradicted their financial benefits. Rambaud et al. (2019) also found a strong preference for falling sequence in car loans, and used the q-exponential discounting to explain the improving sequence effect.

In real world, marketers continually tried to minimize the perceived cost of a product. A common practice is the temporal reframing of prices (hereinafter, TRP), in which the price is expressed by marketers according to a short period, such as car insurance for "$1 a day" as opposed to "$365 a year," despite of the fact that the physical cash flows of the payments remain the same. In an initial study, Gourville (1998) referred to this technique as "pennies-a-day." Gourville (1998, 1999) found that consumers' purchase intentions increased in domains such as charitable donations, cellular telephone services, and health clubs memberships, when the prices were expressed in a per-day form. Gourville (2003) examined the reframed prices of three periods, and found that both per-day and per-month forms were preferred to a per-year form. Bambauer-Sachse and Grewal (2011) examined the role of four moderating variables, and found that per-day reframed prices were more beneficial than aggregate prices for high-priced products, especially in combination with even price endings, a comparatively short time period, or customers with poor calculation affinity.

However, Bambauer-Sachse and Mangold (2009) showed the negative effects of TRP on product evaluations. They found that TRP has positive effects through higher price attractiveness but negative effects through higher complexity of the price structure and a stronger feeling of being manipulated by the marketer. Specifically, price attractiveness positively influences loan evaluations. Previous studies show that objective price presentation influences price perceptions, which affect perceived product quality, value, and willingness to buy (e.g., Dodds et al., 1991; Grewal et al., 1998; Gourville, 2003). If TRP has a positive effect on perceived price attractiveness, it then should result in better evaluations and purchase intentions. In contrast, price complexity negatively influences loan evaluations. According to equity theory (Adams, 1965; Martins and Monroe, 1994), the greater complexity of the temporally reframed price structure implies that more cognitive input is needed, relative to the output gained from the product. Thus, more complex price structures may cause consumers to suspect they are being manipulated by marketers, prompting comparatively negative product evaluations. Price complexity therefore captures both the complexity of price structure and a feeling of being misled (Bambauer-Sachse and Mangold, 2009; Bambauer-Sachse and Grewal, 2011).

The main objective of this paper is to examine the improving sequence effect in a loan context by employing TRP technique. The repayment plans of the loan are expressed in per-day forms and per-year forms. We use perceived price attractiveness to represent the positive effect of TRP, and perceived price complexity to represent the negative effect of TRP. As Bambauer-Sachse and Grewal (2011) stated, per-day reframed loan profiles are perceived as more attractive relative to per-year reframed loan profiles, and thus may result in better evaluation due to this positive effect of TRP. However, they are also perceived as more complex at the same time, and may as well be less preferred due to the negative effect of TRP. The overall evaluation of a loan profile depends on the joint role of price attractiveness and price complexity.

Temporally reframed prices may also interact with the improving sequence effect. According to the improving sequence effect, individuals prefer falling over rising and constant loan profiles. However, some research also detected a strong preference for constant sequences (e.g., Read and Powell, 2002; Hoelzl et al., 2011). Read and Powell (2002) related the preference for constant sequences to "the ease with which money can be managed." This explanation is closely related to price complexity in TRP. A logical deduction is that if the constant loan profile is considered as an easier way to manage money, it may also be perceived as less complex than other profiles. Particularly, marketers can express constant loan profiles using a per-day loan cost, but they have to use a series of falling or rising per-day costs when describing falling or rising profiles. A series of prices are usually considered as more complex than a single price, and then constant profiles will be preferred due to less price complexity. Thus, we hypothesize that the effect of TRP differs across profiles. Specifically, the introduction of the per-day framings affects price complexity of constant loan profiles differently than other profiles. This may explain the popularity of level-payment loans in real-life banking service, as they benefit from less price complexity. Hence, the main objective of this study is to explore the interaction effect between the improving sequence effect and the TRP effect. The foregoing discussion generates the following testable hypotheses:

H1. Ratings of loan profiles are positively correlated with perceived price attractiveness, and negatively correlated with perceived price complexity.

H2. Per-day reframed loan profiles are perceived to be more attractive than per-year reframed profiles.

H3. Per-day reframed loan profiles are perceived to be more complex than per-year reframed profiles for falling and rising profiles, but not for constant profiles.

H4. An interaction effect exists between the improving sequence effect and the TRP effect. When loan profiles are expressed in a per-day form, individuals prefer constant loan profiles over falling and rising loan profiles.

The organization of this paper is as follows. In Section "Methodology," we explain the empirical methodology. In Section "Results," we regress scores of loan profiles on price attractiveness, price complexity, TRP and the interaction terms. We present our conclusions in Section "Discussion."

METHODOLOGY

Material

We conducted this experiment in the same way as Hoelzl et al. (2011) and Rambaud et al. (2019). Participants read scenarios which described that they worked for a big company and earned 10,000 Yuan per month after taxes (1USD≈7 Yuan or $1≈¥7, ¥10,000≈$1,400). They will stay in this job for at least three years. They were asked to consider purchasing a new car that costs ¥120,000 (≈$17,000) on credit. Research shows that per-day framings are more beneficial for products consumed on an ongoing basis than on a lump sum basis (Gourville, 1999), and for high-priced products than low-priced products (Bambauer-Sachse and Grewal, 2011). As cars are expensive and consumed on a continuous basis, we expected that the respondents would prefer the per-day reframed car loans. The loan value was the same as the price of the car with three optional repayment plans (i.e., constant installments, falling installments or rising installments), and with regard to two annual interest rates (10 vs. 0%). The loan is three-year term. Both Hoelzl et al. (2011) and Rambaud et al. (2019) used 5-year loan term in their experiments, but 3-year term is more common in China's auto loan market. The loan was repaid in monthly installments. The monthly principal repayments of the falling plan were ¥5,000 (¥60,000/12) in year 1, ¥3,333.3 (¥40,000/12) in year 2, and ¥1,666.7 (¥20,000/12) in year 3. The monthly principal repayments of the rising plan were ¥1,666.7 in year 1, ¥3,333.3 in year 2, and ¥5,000 in year 3. We adopted similar amortization schedule as Rambaud et al. (2019) except for constant profiles. Both Hoelzl et al. (2011) and Rambaud et al. (2019) designed the constant profiles by fixing the monthly principal repayment. The monthly payments of such constant profiles are actually a falling sequence, as the monthly payment of interest falls over time. In contrast to these studies, our experiment defined the constant sequence as a level payment loan with identical monthly payments (principal + interest) over the term of the loan [see equation (1)].

$$\mathrm{MP_c} = \mathrm{L}\left(\frac{\mathrm{r_L}(1+\mathrm{r_L})^{\mathrm{t}}}{(1+\mathrm{r_L})^{\mathrm{t}}-1}\right) \tag{1}$$

where MP_c is the constant monthly payment, L is the loan principal, r_L is the loan rate, t is the number of installments of this loan, $t\in[1,2,...,n]$.

The loan profiles were presented with per-year repayments or per-day repayments. Although repayments are temporally reframed, the respondents still pay off the loan on a monthly basis. A per-year reframed repayment is the sum of the twelve actual monthly payments in that year, and the per-day reframed repayment is the per-year reframed repayment/365 (see **Supplementary Appendix A**).

Participants

144 MBA students (76 males and 68 females) from Northeastern University (China) with a mean age of 29.48 years took part in the experiment.

Measures

All items were measured on a seven-point rating scale from 1 to 7. At first, participants were asked to evaluate each loan plan, where "1" was the score for a loan they would never choose and "7" was the score for what they considered to be the best plan. Next, they were required to respond to two questions regarding the profiles: price attractiveness ("not at all attractive/extremely attractive"), and price complexity ("not at all complex/extremely complex"). These scales were derived from previous studies (e.g., Bambauer-Sachse and Grewal, 2011; Bornemann and Homburg, 2011; Hoelzl et al., 2011; Shirai, 2018; Rambaud et al., 2019).

Procedure

The questionnaires (see **Supplementary Appendix B**) were presented in a paper-pencil-version at Northeastern University (China), and were distributed in MBA classes. Participants were asked to assign scores to the three repayment plans at two interest rates and at per-day or per-year framings. They were randomly assigned to one of the four experimental groups via the questionnaires (2 rates × 2 temporal framings), which were also randomized. We decided the sample size according to the number of MBA students. Also, we designed our study to let each group have the same number (36) of participants for comparison's sake, thereby resulting in an analytic sample of 144 (36 × 4) participants.

Participants were allowed to assign the same score to the three plans. They were then requested to evaluate price attractiveness and price complexity of the profiles using a 1 to 7 scale. At the beginning of the experiment, the researcher explained the procedure. The experiment took approximately 15 min to complete. No monetary incentive was given for participation.

Finally, to offset the impact of stylized responses, the order of presentation of the profiles was counterbalanced across subjects. Therefore, for those 36 subjects in each group, 12 saw falling, constant and rising profile from left to right, 12 saw constant, rising, and falling profile from left to right, while 12 saw rising, falling and constant profile from left to right.

RESULTS

Interaction Effect Between the Improving Effect and TRP Effect

Means of Evaluations

Participants evaluated the rising profile as the least preferred option regardless of the loan rate and temporal framings.

TABLE 1 | Means of evaluations.

Groups	Number of subjects	Sequence Profiles								
		Falling			Constant			Rising		
		Score	PC	PA	Score	PC	PA	Score	PC	PA
Per-year, 0%	36	5.19 (1.62)	1.69 (1.43)	4.08 (1.68)	5.64 (1.10)	1.67 (1.76)	4.94 (1.37)	3.81 (1.6181)	1.61 (1.34)	3.14 (1.85)
Per-day, 0%	36	3.86 (1.74)	3.19 (1.97)	4.97 (1.25)	5.75 (1.32)	1.72 (1.45)	5.58 (1.32)	3.14 (1.81)	3.33 (2.11)	4.11 (1.47)
Per-year, 10%	36	5.19 (1.95)	1.56 (0.88)	4.56 (1.65)	4.92 (1.44)	1.64 (1.38)	4.69 (1.77)	3.50 (1.63)	1.81 (1.53)	2.39 (1.73)
Per-day, 10%	36	4.64 (1.62)	2.97 (1.08)	4.86 (1.46)	5.28 (1.26)	1.61 (0.77)	5.25 (1.23)	2.56 (2.08)	2.75 (1.52)	2.92 (2.26)

Score is the overall evaluation of loan profiles, PC is the perceived price complexity, and PA is the perceived price attractiveness (with standard deviations in parentheses).

This result provides additional support for the improving effect (Loewenstein and Prelec, 1993), and is consistent with the result of Hoelzl's (2011) study. The preference order of per-year reframed profiles at 10% discount rate (falling > constant > rising) was consistent with the order deduced from utilizing the DUM and exponential discounting. However, the preference orders of the other three groups contradicted the DUM. **Table 1** shows the group means of scores, the perceived price complexity and price attractiveness of the profiles.

ANOVA Results

We analyzed the means using 3(sequences) × 2(TRP) × 2(Interest) ANOVAs. Normality is not an issue for our large sample size. According to central limit theorem, for sufficiently large samples with size greater than 30 (144 in our study), the sampling distribution for means is always normally distributed regardless of a variable's original distribution. Because the loan profiles have roughly equal standard deviations, ranging from 1.3 to 1.9, the assumption of homoscedasticity is also met. We run the tests in SPSS version 20. The sequence score, perceived price complexity, and perceived price attractiveness were used as the dependent variables (a within-subject factor). The independent variables included the interest rate (10%, or 0%), and TRP (day-framing or year-framing), which are all between-subjects factors. A **Tables 2–4** show the results of the ANOVAs. **Figures 1A–C** show the estimated marginal means.

TABLE 2 | ANOVA results for evaluation score.

Factor	DF1	DF2	F	$MS_{between}$	MS_{within}	η_p^2
Sequence	2	264	54.936***	173.419	3.157	0.282
TRP	1	140	17.433***	27.502	1.578	0.111
Interest	1	140	3.241	5.113	1.578	0.023
Sequence × TRP	2	264	4.748**	14.988	3.157	0.033
Sequence × Interest	2	264	3.213*	10.141	3.157	0.022
TRP × Interest	1	140	1.070	1.688	1.578	0.008
Sequence × TRP × Interest	2	264	0.794	2.507	3.157	0.006

*Columns list the degrees of freedom for the numeration (DF1), and denominator (DF2), the F ratio (F), the mean-squared between ($MS_{between}$), the mean-squared within (MS_{within}), and the partial eta squared (η_p^2). *p < 0.05, **p < 0.01, ***p < 0.001.*

TABLE 3 | ANOVA results for price complexity.

Factor	DF1	DF2	F	$MS_{between}$	MS_{within}	η_p^2
Sequence	2	264	19.910***	23.863	1.199	0.125
TRP	1	140	22.467***	94.454	4.204	0.138
Interest	1	140	0.564	2.370	4.204	0.004
Sequence × TRP	2	264	19.238***	23.058	1.199	0.121
Sequence × Interest	2	264	0.141	0.169	1.199	0.001
TRP × Interest	1	140	0.637	2.676	4.204	0.005
Sequence × TRP × Interest	2	264	1.207	1.447	1.199	0.009

*Columns list the degrees of freedom for the numeration (DF1), and denominator (DF2), the F ratio (F), the mean-squared between ($MS_{between}$), the mean-squared within (MS_{within}), and the partial eta squared (η_p^2). ***p < 0.001.*

TABLE 4 | ANOVA results for price attractiveness.

Factor	DF1	DF2	F	$MS_{between}$	MS_{within}	η_p^2
Sequence	2	264	58.420***	152.521	2.611	0.294
TRP	1	140	17.639***	45.370	2.572	0.112
Interest	1	140	5.475*	14.083	2.572	0.038
Sequence × TRP	2	264	0.107	0.280	2.611	0.001
Sequence × Interest	2	264	4.631*	12.090	2.611	0.032
TRP × Interest	1	140	1.440	3.704	2.572	0.010
Sequence × TRP × Interest	2	264	0.230	0.600	2.611	0.002

*Columns list the degrees of freedom for the numeration (DF1), and denominator (DF2), the F ratio (F), the mean-squared between ($MS_{between}$), the mean-squared within (MS_{within}), and the partial eta squared (η_p^2). *p < 0.05 and ***p < 0.001.*

Figure 1 shows the estimated marginal means of falling, constant and rising profiles with regard to per-day and per-year framings. The lines in **Figures 1A,B** are far from parallel, suggesting an interaction effect between the improving sequence effect and the TRP effect, i.e., the improving sequence effect is different for per-year reframed and per-day reframed profiles.

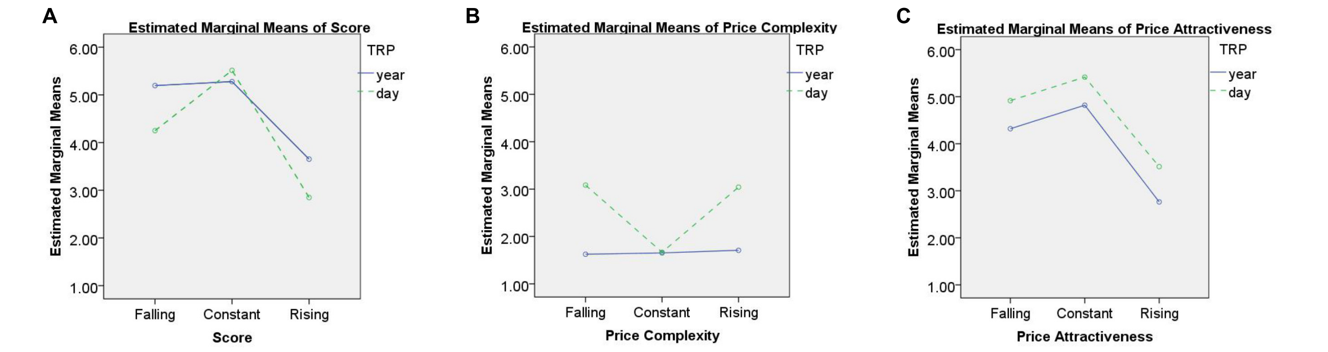

FIGURE 1 | Estimated marginal means (day vs. year). **(A)** Score, **(B)** price complexity, and **(C)** price attractiveness.

Table 2 shows that the main effects of Sequence and TRP are significant, suggesting the existence of the improving sequence effect and TRP effect. The results also show a significant Sequence × TRP interaction, and a significant Sequence × Interest interaction. To identify the locus of Sequence × TRP interaction, we examined the effect of sequence for per-day and per-year framings separately. At per-year framings, $F = 21.257$, $p < 0.001$, $\eta_p^2 = 0.23$. Pair comparisons show that the rising profile differs from the falling profile (mean difference = −1.542, $p < 0.001$) and the constant profile (mean difference = −1.625, $p < 0.001$). The difference between the falling and constant profile, however, is not statistically significant (mean difference = −0.08, $p > 0.05$). In contrast, at per-day framings, $F = 35.923$, $p < 0.001$, $\eta_p^2 = 0.336$. All three profiles are significantly different from each other. The rising profile differs from the falling profile (mean difference = −1.403, $p < 0.001$) and the constant profile (mean difference = −2.667, $p < 0.001$). The falling profile differs from the constant profile (mean difference = −1.264, $p < 0.001$). In general, the sequence effects are significant regardless of TRP involved in the profile. **Tables 1–4** show that the rising profile is with the least score in all conditions, indicating that participants are not financially rational and the DUM is violated. This result provides additional support for the improving sequence effect (Loewenstein and Prelec, 1993), and is consistent with the results of Hoelzl's (2011) study and Rambaud et al.'s (2019) study. However, individuals prefer the constant profile over the falling profile at per-day framings.

Table 3 shows a significant main effect of TRP, indicating that a per-day reframed price is generally perceived to be more complex than a per-year reframed price for falling and rising profiles. Thus, H3 is supported. There is also a significant interaction effect between Sequence and TRP for price complexity. We examined the sequence effect for per-day and per-year framings separately. At per-year framings, $F = 0.138$, $p > 0.05$, $\eta_p^2 = 0.002$. Pair comparisons suggest that individuals perceive all three profiles as equally complex. Neither the difference between the rising and falling profile (mean difference = 0.083, $p > 0.05$), the difference between the rising and constant profile (mean difference = 0.056, $p > 0.05$), nor the difference between the falling and constant profile (mean difference = −0.028, $p > 0.05$) is significant. In contrast, At per-day framings, $F = 32.399$, $p < 0.001$, $\eta_p^2 = 0.316$. Pair comparisons show that the constant profile differs from the falling profile (mean difference = −1.417, $p < 0.001$) and the rising profile (mean difference = −1.375, $p < 0.001$). But the difference between the falling and rising profile is not significant (mean difference = 0.042, $p > 0.05$). The result indicates that the constant profile is perceived to be less complex only when the loan profiles are expressed in a per-day form. This result is consistent with the result of **Table 2**, as the falling profile is preferred when the profiles are described in a per-year form.

Table 4 shows that using a per-day reframed price leads to a significantly more positive perception of price attractiveness than using a per-year reframed price, as the main effect of TRP is also significant. Therefore, H2 is supported. No significant interaction effect between Sequence and TRP is found.

As **Tables 2–4** show significant sequence x TRP interactions in score and price complexity, we examined the main effect of TRP for each sequence. **Table 5** shows that TRP affects score, price attractiveness, price complexity of falling and rising loan profiles. However, TRP does not significantly affect score and price complexity of constant profiles. This finding is consistent with the mean values in **Table 1**, in which the mean score of constant profiles in a per-day form is not significantly different from that in a per-year form. However, the mean score of constant profiles is significantly higher than the mean scores

TABLE 5 | TRP effect for each sequence.

Evaluations	DF1	DF2	F
Scores of falling profiles	1	142	10.488***
Scores of constant profiles	1	142	1.166
Scores of rising profiles	1	142	7.249**
PA of falling profiles	1	142	5.573*
PA of constant profiles	1	142	6.241*
PA of rising profiles	1	142	5.584*
PC of falling profiles	1	142	39.337***
PC of constant profiles	1	142	0.004
PC of rising profiles	1	142	23.438***

**p < 0.05, **p < 0.01, ***p < 0.001.*

TABLE 6 | Regression results.

	Falling		Constant		Rising	
	Coeff.	**VIF**	**Coeff.**	**VIF**	**Coeff.**	**VIF**
Model 1						
Attractiveness	0.687 (0.075)***	1.004	0.571 (0.058)***	1.005	0.602 (0.062)***	1.028
Complexity	−0.364 (0.074)***	1.004	−0.062 (0.062)	1.005	−0.271 (0.067)***	1.028
F(2,141)	51.246***		49.578***		50.387***	
R^2	0.421		0.413		0.417	
Model 2						
Attractiveness	0.677 (0.084)***	1.236	0.572 (0.058)***	1.005	0.640 (0.060)***	1.059
Complexity	−0.361 (0.075)***	1.029	−0.057 (0.062)	1.013	−0.337 (0.067)***	1.103
Attractiveness * Complexity	−0.017 (0.062)	1.248	0.034 (0.043)	1.009	0.139 (0.037)***	1.090
F(3,140)	33.967***		33.182***		41.505***	
ΔF	0.079		0.641		14.262***	
R^2	0.421		0.416		0.471	
ΔR^2	0.000		0.003		0.054	
Model 3						
Attractiveness	0.676 (0.114)***	2.543	0.489 (0.092)***	2.730	0.864 (0.089)***	2.769
Complexity	−0.210 (0.134)***	3.668	0.072 (0.077)	1.676	−0.481 (0.111)***	3.574
Attractiveness * Complexity	−0.017 (0.062)	1.405	−0.026 (0.050)	1.472	0.123 (0.036)***	1.211
TRP	−1.079 (0.255)***	1.356	−0.088 (0.172)	1.124	−0.884 (0.233)***	1.262
Interest	0.212 (0.222)	1.023	−0.457 (0.165)**	1.027	0.404 (0.222)	1.147
Complexity × TRP	0.028 (0.172)	3.378	−0.423 (0.154)**	2.346	0.352 (0.135)**	3.037
Attractiveness × Interest	0.110 (0.146)	2.097	0.031 (0.117)	2.411	−0.250 (0.115)*	2.525
F(7,136)	19.120***		17.615***		25.006***	
ΔF	4.052**		3.249**		9.079***	
R^2	0.496		0.476		0.563	
ΔR^2	0.075		0.063		0.146	

**p < 0.05, **p < 0.01, ***p < 0.001. Coeff. is unstandardized regression coefficient (standard error).*

of falling and rising profiles when all profiles are described in a per-day form. A possible explanation is that constant profiles are positively affected by TRP in terms of higher price attractiveness just like falling and rising profiles. But unlike other profiles, when switching from a per-year form to a per-day form, constant profiles are not perceived to be more complex, i.e., the falling and rising profiles are exposed to both positive and negative effects of TRP, while the constant profile only benefits from the positive effect of TRP. Therefore, H4 is supported.

Regression Analysis Between Scores of Loan Profiles, Price Attractiveness and Price Complexity

As the main focus of this study is to explore the interaction between the improving sequence effect and TRP effect, we treated TRP as a between-subjects factor in our experiment and ran hierarchical multiple regression analysis with one dependent variable (scores). In model 1, two independent variables were included: price complexity and price attractiveness. **Table 6** shows the results of regression tests (we run the tests in SPSS version 20.). Coefficients of price attractiveness are positive and coefficients of price complexity are negative for all loan profiles. All coefficients are statistically significant except for that of price complexity for constant profiles. The exception is possibly because the per-day and per-year framings have close mean price complexities (see **Table 1**). These variables accounted for a significant amount of variance in scores. For falling profiles, $R^2 = 0.421$, $F(2, 141) = 51.246$, $p < 0.001$; for constant profiles, $R^2 = 0.413$, $F(2, 141) = 49.578$, $p < 0.001$; for rising profiles, $R^2 = 0.417$, $F(2, 141) = 50.378$, $p < 0.001$.

In model 2, we centered price complexity and price attractiveness, and used the multiply as the third independent variable to examine the moderation. The interaction term between price complexity and price attractiveness was added to the regression model. For rising profiles, the interaction term is significant, and model 2 accounts for significantly more variance than model 1, $\Delta R^2 = 0.054$, $\Delta F = 14.262$, $p < 0.001$. This result shows that the effect of price attractiveness is higher when the perceived price complexity is high, relative to the effect when the perceived price complexity is low. However, the interaction term is not significant for falling or constant profiles.

In model 3, we included TRP and interest rate as independent variables, TRP = 0 for per-year reframings, TRP = 1 for per-day reframings, Interest = 0 for 0%, Interest = 1 for 10%. Complexity × TRP and Attractiveness × Interest interactions were also included because of the significant interaction effects (see **Tables 2–4**). Model 3 accounts for significantly more variance than model 1 for all profiles ($p < 0.01$ for falling and constant profiles, $p < 0.001$ for rising profiles). Although price attractiveness and price complexity captures most of the changes in scores, TRP and interest rate also influence evaluations of loan profiles.

In general, scores are positively correlated with the perceived price attractiveness, and negatively correlated with the perceived price complexity. The inclusion of covariates such as TRP and interest rate significantly increase the R^2, but price attractiveness and price complexity account for most of the variance in scores in all three models. No multicollinearity was detected. Therefore, H1 is supported.

DISCUSSION

In this study, we examined the preferences for sequences in context of a car loan when the loan repayment plans are expressed in temporally reframed prices. Our study is motivated by the fact that TRP tactic has been widely used as an effective pricing strategy to improve consumer's product evaluations. In general, our results show that TRP has positive effects through higher price attractiveness but negative effects through higher price complexity. The results also support the improving sequence effect. Also, we found an interaction effect between the improving sequence effect and TRP. Although TRP tactic improves price attractiveness for all loan profiles, it affects price complexity differently. Specifically, the introduction of TRP leads to higher price complexity for falling and rising loan profiles, but has no significant influence on constant profiles. Thus, individuals choosing among loan repayment profiles expressed in per-day forms will prefer constant profiles.

A number of research papers provided explanations for preferences in relation to money sequences (e.g., Loewenstein and Sicherman, 1991; Chapman, 1996, 2000; Read and Powell, 2002). Many studies believe that the violation of the DUM is caused by the misuse of exponential discount function. They explained the improving sequence effect by employing discount functions other than exponential discounting. For example, hyperbolic discounting (Loewenstein and Prelec, 1993; Overton and MacFadyen, 1998) and the q-exponential discounting (Rambaud et al., 2019) were used. Rambaud et al. (2019) stated that the falling profile is more appealing if participants discount future loan repayments using the q-exponential discounting instead of the traditional exponential function. The q-exponential discount function is known in the deformed algebra inspired in non-extensive thermodynamics (Tsallis, 1994), and was first utilized to study intertemporal choices, as proposed by Cajueiro (2006).

$$V(L) = \sum_{t=1}^{n} MP_t/[1 + (1 - q)r_q t)]^{1/(1-q)} \tag{2}$$

where t, L stay the same, MP_t is the monthly repayment, $V(L)$ is subject discounted value of the repayments, and r_q and q are discount parameters of the model, $t \in [1,2,...,n]$. For $q \rightarrow 1$, the q-exponential discount recovers the classical exponential discount. For $q \rightarrow 0$, it yields the simple hyperbolic discount (Cajueiro, 2006). Hence, with two free parameters, the q-exponential discount model is a general form of the exponential discount model and simple hyperbolic model, in which 1-q indicates the degree of inconsistency (Takahashi et al., 2007). If 1-$q > 0$, q-exponential discounting exhibits decreasing impatience, "the instantaneous discount rate is decreasing according to the value of q" (Rambaud and Torrecillas, 2013). Because the discount factor of the q-exponential discount function between adjacent periods is smaller than between similar periods that are further away, the discount rate of the q-exponential discount function is higher than that of the exponential discount function at the beginning of the loan term, but is lower in the long run.

The inconsistency level can be calculated as the coefficient of variation (CV) of the obtained average scores (see **Table 1**): 1-q (CV) for the four groups (Per-year, 0%, Per-day, 0%, Per-year, 10%, Per-day, 10%) are 0.1954, 0.3172, 0.2001, and 0.3419, and all greater than 0. Due to this time inconsistency, the falling profile is more appealing if participants discount MP_t using the q-exponential function instead of the exponential function, as the former function results in a small present value. This type of thinking was labeled as "optimization" by Read and Powell (2002), because individuals can always maximize their utilities by choosing the sequence with the highest present value of positive outcomes (Samuelson, 1937), or lowest present value in context of a loan.

However, some empirical results contradicts the "optimization" theory. For example, studies also found the improving sequence effect in the context of interest-free loans (Hirst et al., 1992; Wonder et al., 2008; Hoelzl et al., 2011). As Rambaud et al. (2019) also stated, no discount function can explain the improving sequence effect if the interest rate is zero. As the rising profile will always has the least subjective discounted value regardless of discount function, it should represent respondents' best choice. Moreover, individuals may have limited financial capability to discount future outcomes. Herrmann and Wricke (1998) found that when evaluating the attractiveness of auto loan offers, respondents did not even calculate the product of monthly payment and number of payments, not to mention using discounted values. "Optimization" cannot explain the preference pattern in our result either, as the introduction of TRP does not change the physical cash flows of the payments, the discounted values of the per-day and per-year reframed loan profiles are identical regardless of discount function.

A possible explanation is that consumers do not process price information completely but use simplifying heuristics (Anderson, 1971; Davis et al., 1986; Bambauer-Sachse and Mangold, 2009). Therefore, they may evaluate loan profiles on the basis of the reframed price and predict a lower total cost. Furthermore, consumers may compare the per-day loan cost to the cost of a petty cash expense. For example, an advertisement for smart phones stated "For the Cost of Your Morning Coffee, Never Be Un-Reachable!." Likewise, a per-day reframed constant loan profile can also be compared to a breakfast or a pack of cigarettes. TRP induces consumers to compare the per-day loan cost to a petty cash expense or daily budget, and thus influences their perceptions of product affordability. For example, the per-day expressed constant profile at 0% loan rate in our study is only ¥109.6 (≈$15) per day, very close to the expense of a good lunch or a pack of top brand cigarette, easily fitting into many respondents' daily budgets. Gourville (1999)'s result shows that an explicit petty cash comparison (e.g., one's morning coffee) can be as impactful as a per-day framing at influencing product purchase intention. Either an implicit comparison via per-day framing, or an explicit petty cash comparison will result in significantly higher perceived values. In the field of sequence preference, Read and Powell (2002) labeled this type of thinking as "Ideal consumption," as people tend to choose the sequence that they believe as appropriate (Chapman, 1996). Read and Powell (2002) also found a strong preference for constant sequences, mostly related to reasons of "convenience" or "the ease with which money can be managed." In our study, the per-day reframed rising or falling profile can only be expressed as a rising or falling sequence of per-day loan costs, i.e., there are three different per-day loan costs in three years, making the petty cash comparison less obvious. Therefore, rising and falling profiles are perceived as more difficult to manage than constant profiles.

CONCLUSION

Previous studies have shown a consistent preference for the falling sequence in loan repayment plans, suggesting that banks need to develop loan schemes in which the repayments are concentrated at the beginning of the loan term. However, our results show that consumers follow a comparison-based decision making process rather than optimization when evaluating temporally reframed loan offerings. Individuals preferred the falling over the constant profile only if the interest rate was 10% and the loan profiles were described in a per-year form. Otherwise, they preferred the constant profile. Therefore, regardless of the amply evidence supporting the improving sequence effect, borrowers may still prefer the level payment loans, especially when the loan profiles are expressed in a per-day form.

In general, we found that the improving sequence effect existed in a loan context and the DUM was violated. However, the violation of the DUM in the 0% interest condition cannot be explained by any discount function. Thus, we propose that future studies in sequence effect may also consider psychological reasons and comparison-based decision making process. However, there are limitations that need to be addressed in future studies. First, the study is limited in external validity in that respondents are not a representative sample from any particular population (all MBA students from the same university). Furthermore, the generalizability of the findings is limited in that the loan stimuli are entirely hypothetical based on a fictional job scenario provided to the students. Future research should design the experiment based on participants' real-life job and financial backgrounds.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YL designed research and wrote the manuscript. CL and HH collected data. JW and XZ analyzed data. All authors contributed to the article and approved the submitted version.

FUNDING

National Natural Science Foundation of China [Award number(s): 71571038 and 71971048], Fundamental Research Funds for the Central Universities [Award number: N2106002], Social Science Funds of Liaoning Province [Award number: L20BJY008], and Shenyang Science and Technology Innovation Support Plan [Award number: RC190458].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.532696/full#supplementary-material

REFERENCES

Adams, J. S. (1965). "Inequity in social exchange," in *Advances in Experimental Social Psychology*, ed. B. Leonard (New York, NY: Academic Press), 267–299. doi: 10.1016/s0065-2601(08)60108-2

Anderson, N. H. (1971). Integration theory and attitude change. *Psychol. Rev.* 78, 171–206. doi: 10.1037/h0030834

Ariely, D., and Loewenstein, G. (2000). When does duration matter in judgment and decision making. *J. Exp. Psychol. Gen.* 29, 508–523. doi: 10.1037//0096-3445.129.4.508

Ariely, D., and Zauberman, G. (2000). On the making of an experience: the effects of breaking and combining experiences on their overall evaluation. *J. Behav. Decis. Making* 13, 219–232. doi: 10.1002/(sici)1099-0771(200004/06)13:2, <219::aid-bdm331>3.0.co;2-p

Bambauer-Sachse, S., and Grewal, D. (2011). Temporal reframing of prices: when is it beneficial? *J. Retail.* 87, 156–165. doi: 10.1016/j.jretai.2011.02.002

Bambauer-Sachse, S., and Mangold, S. C. (2009). Are temporally reframed prices really advantageous? A more detailed look at the processes triggered by temporally reframed prices. *J. Retail. Consum. Serv.* 16, 451–457. doi: 10.1016/j.jretconser.2009.06.005

Bornemann, T., and Homburg, C. (2011). Psychological distance and the dual role of price. *J. Consum. Res.* 38, 490–504. doi: 10.1086/659874

Cajueiro, D. (2006). A note on the relevance of the q-exponential function in the context of intertemporal choices. *Phys. A* 364, 385–388. doi: 10.1016/j.physa.2005.08.056

Chapman, G. B. (1996). Expectations and preferences for sequences of health and money. *Organ. Behav. Hum. Decis. Process.* 67, 59–75. doi: 10.1006/obhd.1996.0065

Chapman, G. B. (2000). Preferences for improving and declining sequences of health outcomes. *J. Behav. Decis. Making* 13, 203–218.

Davis, H. L., Hoch, S. J., and Ragsdale, E. E. K. (1986). An anchoring and adjustment model of spousal predictions. *J. Consum. Res.* 13, 25–37. doi: 10.1086/209045

Dodds, W. B., Monroe, K. B., and Grewal, D. (1991). Effects of price, brand, and store information on buyers' product evaluations. *Journal of Marketing Research* 28, 307–319. doi: 10.1177/002224379102800305

Duffy, S., and Smith, J. (2013). Preference for increasing wages: how do people value various streams of income? *Judg. Decis. Making* 8, 74–90. doi: 10.2139/ssrn.1631845

Duxbury, D., Summers, B., Hudson, R., and Keasey, K. (2013). How people evaluate defined contribution, annuity-based pension arrangements: a behavioral exploration. *J. Econ. Psychol.* 34, 256–269. doi: 10.1016/j.joep.2012.10.008

Garcia, A., Torrecillas, M. J., and Rambaud, S. C. (2020). The improving sequence effect on monetary sequences. *Heliyon* 6:e05643. doi: 10.1016/j.heliyon.2020.e05643

Gigliotti, G., and Sopher, B. (1997). Violations of present-value maximization in income choice. *Theory Decis.* 43, 45–69. doi: 10.1023/A:1004950613488

Gourville, J. T. (1998). Pennies-a-Day: the effect of temporal reframing on transaction evaluation. *J. Consum. Res.* 24, 395–408. doi: 10.1086/209517

Gourville, J. T. (1999). The effect of implicit versus explicit comparisons on temporal pricing claim. *Mark. Lett.* 10, 113–124.

Gourville, J. T. (2003). The effects of monetary magnitude and level of aggregation on the temporal framing of price. *Mark. Lett.* 14, 125–135.

Grewal, D., Monroe, K. B., and Krishnan, R. (1998). The effects of price-comparison on buyers' perceptions of acquisition value, transaction value, and behavioral intentions. *J. Mark.* 62, 46–59. doi: 10.2307/1252160

Guyse, J., Keller, L., and Eppel, T. (2002). Valuing environmental outcomes: preferences for constant or improving sequences. *Organ. Behav. Hum. Decis. Process.* 87, 253–277. doi: 10.1006/obhd.2001.2965

Hassenzahl, M. (2005). Intertemporale Wahl: die präferenz für verbesserung bei der rückzahlung von Schulden. *Wirtschaftspsychologie* 7, 102–108.

Herrmann, A., and Wricke, M. (1998). Evaluating multidimensional prices. *J. Prod. Brand Manag.* 7, 161–169. doi: 10.1007/978-3-662-09119-7_12

Hirst, E., Joyce, E., and Schadewald, M. (1992). Mental accounting and outcome contiguity in consumer borrowing decisions. *Organ. Behav. Hum. Decis. Process.* 58, 136–152. doi: 10.1006/obhd.1994.1031

Hoelzl, E., Kamleitner, B., and Kirchler, E. (2011). Loan repayment plans as sequences of instalments. *J. Econ. Psychol.* 32, 621–631. doi: 10.1016/j.joep.2011.02.002

Langer, T., Sarin, R., and Weber, M. (2005). The retrospective evaluation of payment sequences: duration neglect and peak-and-end effects. *J. Econ. Behav. Organ.* 58, 157–175. doi: 10.1016/j.jebo.2004.01.001

Loewenstein, G., and Prelec, D. (1991). Negative time preference. *Am. Econ. Rev.* 81, 347–352.

Loewenstein, G., and Prelec, D. (1993). Preferences for sequences of outcomes. *Psychol. Rev.* 100, 91–108. doi: 10.1037/0033-295x.100.1.91

Loewenstein, G., and Sicherman, N. (1991). Do workers prefer increasing wage profiles. *J. Lab. Econ.* 9, 67–84. doi: 10.1086/298259

Martins, M., and Monroe, K. B. (1994). "Perceived price fairness: a new look at an old construct," in *Advances in Consumer Research*, eds C. T. Allen and D. Roedder-John (Provo, UT: Association for Consumer Research), 281–304.

Matsumoto, D., Peecher, M. E., and Rich, J. S. (2000). Evaluations of outcome sequences. *Organ. Behav. Hum. Decis. Process.* 83, 331–352. doi: 10.1006/obhd.2000.2913

Overton, A. A., and MacFadyen, A. J. (1998). Time discounting and the estimation of loan duration. *J. Econ. Psychol.* 19, 607–618. doi: 10.1016/s0167-4870(98)00027-0

Rambaud, S. C., Pascual, J. L., and Álvarez, M. (2019). Preferences over sequences of payments: a new validation of the q-exponential discounting. *Phys. A* 515, 332–345. doi: 10.1016/j.physa.2018.09.169

Rambaud, S. C., and Torrecillas, M. J. (2013). A generalization of the q-exponential discounting function. *Phys. A* 392, 3045–3050. doi: 10.1016/j.physa.2013.03.009

Rambaud, S. C., Torrecillas, M. J., and Garcia, A. (2018). A mathematical analysis of the improving sequence effect for monetary rewards. *Front. Appl. Math. Stat.* 4:55. doi: 10.3389/fams.2018.00055

Read, D., and Powell, M. (2002). Reasons for sequence preferences. *J. Behav. Decis. Making* 15, 433–460. doi: 10.1002/bdm.429

Samuelson, P. A. (1937). A note on measurement of utility. *Rev. Econ. Stud.* 4, 155–161. doi: 10.2307/2967612

Shirai, M. (2018). Consumer perceptions of price reframing in an in-store decision context. *J. Fin. Econ.* 7, 20–29. doi: 10.12735/jfe.v7n1p20

Takahashi, T., Oono, H., and Radford, M. H. B. (2007). Empirical estimation of consistency parameter in intertemporal choice based on Tsallis' statistics. *Phys. A* 381, 338–342. doi: 10.1016/j.physa.2007.03.038

Thaler, R. (1999). Mental accounting matters. *J. Behav. Decis. Making* 12, 183–206. doi: 10.1002/(sici)1099-0771(199909)12:3, <183::aid-bdm318>3.0.co;2-f

Tsallis, C. (1994). What are the numbers that experiments provide? *Quím. Nova* 17, 468–471.

Wonder, N., Wilhelm, W., and Fewings, D. (2008). The financial rationality of consumer loan choices: revealed preferences concerning interest rates, down payments, contract length and rebates. *J. Consum. Aff.* 42, 243–270. doi: 10.1111/j.1745-6606.2008.00107.x

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 July 2021
doi: 10.3389/fpsyg.2021.580015

Item-Weighted Likelihood Method for Measuring Growth in Longitudinal Study With Tests Composed of Both Dichotomous and Polytomous Items

Xuemei Xue[1], *Jing Lu*[2]* *and Jiwei Zhang*[3]*

[1] School of Mathematical Sciences, Xiamen University, Xiamen, China, [2] Key Laboratory of Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun, China, [3] Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, School of Mathematics and Statistics, Yunnan University, Kunming, China

In this paper, a new item-weighted scheme is proposed to assess examinees' growth in longitudinal analysis. A multidimensional Rasch model for measuring learning and change (MRMLC) and its polytomous extension is used to fit the longitudinal item response data. In fact, the new item-weighted likelihood estimation method is not only suitable for complex longitudinal IRT models, but also it can be used to estimate the unidimensional IRT models. For example, the combination of the two-parameter logistic (2PL) model and the partial credit model (PCM, Masters, 1982) with a varying number of categories. Two simulation studies are carried out to further illustrate the advantages of the item-weighted likelihood estimation method compared to the traditional Maximum a Posteriori (MAP) estimation method, Maximum likelihood estimation method (MLE), Warm's (1989) weighted likelihood estimation (WLE) method, and type-weighted maximum likelihood estimation (TWLE) method. Simulation results indicate that the improved item-weighted likelihood estimation method better recover examinees' true ability level for both complex longitudinal IRT models and unidimensional IRT models compared to the existing likelihood estimation (MLE, WLE and TWLE) methods and MAP estimation method, with smaller bias, root-mean-square errors, and root-mean-square difference especially at the low-and high-ability levels.

Keywords: longitudinal model, item-weighted likelihood, mixed-format test, dichotomous item response, polytomous item response

OPEN ACCESS

Edited by:
Salvador Cruz Rambaud,
University of Almería, Spain

Reviewed by:
Chester Chun Seng Kam,
University of Macau, China
Philipp Doebler,
Technical University Dortmund,
Germany

***Correspondence:**
Jing Lu
luj282@nenu.edu.cn
Jiwei Zhang
zhangjw713@nenu.edu.cn

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

Received: *04 July 2020*
Accepted: *21 June 2021*
Published: *27 July 2021*

Citation:
Xue X, Lu J and Zhang J (2021)
Item-Weighted Likelihood Method
for Measuring Growth in Longitudinal
Study With Tests Composed of Both
Dichotomous and Polytomous Items.
Front. Psychol. 12:580015.
doi: 10.3389/fpsyg.2021.580015

INTRODUCTION

The measurement of change has been a topic to both practitioners and methodologists (e.g., Dearborne, 1921; Woodrow, 1938; Lord, 1963; Fischer, 1973, 1976, 1995; Rasch, 1980; Andersen, 1985; Wilson, 1989; Embretson, 1991, 1997; von Davier and Xu, 2011; Barrett et al., 2015). Item response theory (IRT), particularly, a family of Rasch models (RM), provides a new perspective to modeling change. Andersen (1985) proposed the multidimensional Rasch model for modeling growth in the repeated administration of the same items at different occasions. Embretson (1991) presented a special multidimensional Rasch model for measuring learning and change (MRMLC) based on IRT. Embretson's model postulated the involvement of K abilities for K occasions. Specifically, the MRMLC assumes that on the first occasion ($k = 1$), performance depends on initial ability. The MRMLC further assumes that on later occasions ($k > 1$), performance also depends on $k-1$ additional abilities, termed "modifiabilities," as well as initial ability. Thus, the number of abilities increases at each time point. The same items are repeated over occasions in Andersen's model which may lead to practice effects or memory effects and result in local dependency among item responses (von Davier and Xu, 2011), whereas items in Embretson's MRMLC are not necessarily repeated. Fischer (2001) extended the MRMLC to polytomous items by extending the partial credit model (PCM, Masters, 1982). This paper extends Embretson's method to measure growth based on item responses from mixed-format tests composed of both dichotomous and polytomous items which are frequently used in large-scale educational assessments, such as the National Assessment of Educational Progress (NAEP) and the Program for International Student Assessment (PISA). For polytomous items, each response category provides information. If categories within an item are close together, the item information will be peaked near the center of the location parameter of category. However, if the categories are spread further apart, each can add information at a different location. Therefore, the item information for a polytomous item can have multiple peaks and can be spread over a broader extent of the ability range. Thus, polytomous items may contain more information than dichotomous items (e.g., Donoghue, 1994; Embretson and Reise, 2000, p. 95; Jodoin, 2003; Penfield and Bergeron, 2005; Yao, 2009; Christine, 2010; Tao et al., 2012). How to utilize the potential information difference hidden in different item types to improve estimates of the latent trait is the main concern in our study.

As mentioned above, it has been demonstrated that polytomous items can often provide more information than dichotomous items concerning the level of estimated latent trait (Tao et al., 2012). Meanwhile, different items of the same type may provide different amount of information about latent trait estimation. To improve the precision of ability estimation, the aim of this study is to develop an efficient item-weighting scheme by assigning different weights to different items in accordance with the amount of information for a certain latent trait level. As early as 40 years ago, Lord (1980) has considered to optimal item weights for dichotomously scored items. Tao et al. (2012) proposed a bias-reduced item-weighted likelihood estimation method, and Sun et al. (2012) proposed weighted maximum-a-posteriori estimation, which focused on differentiating the information gained from different item types. In their methods, the weights were pre-assigned and known or automatically selected such that the weights assigned to the polytomous items are larger than that assigned to the dichotomous items. They assign different weights to different item types, instead of assigning different weights to different items, and items of the same type all have the same weight. For convenience, we called these weighting methods type-weighted estimation. However, different items of the same type may have different information for a certain latent trait level; the same weights assigned to the same-type items may not be statistically optimal in terms of the precision and accuracy of ability estimation due to neglecting the difference in the individual item contribution. It is expected that assigning a weight for each item based on its own contribution may increase measurement precision.

The remainder of this paper is organized as follows. First, we present the MRMLC and its polytomous extension, and then the proposed item-weighted likelihood estimation (IWLE) method and the other two ability estimation methods: Warm's (1989) weighted likelihood estimation (WLE) and type-weighted maximum likelihood estimation (TWLE). Second, we show that the IWLE is consistent and asymptotically normal with mean zero and a variance-covariance matrix, and the bias of IWLE is of order n^{-1}. Third, a simulation study is conducted to compare the proposed IWLE method with MLE, MAP, WLE, and TWLE. Fourth, a simulation study is conducted to show IWLE can also be applied to general unidimensional item response models. Finally, we conclude this paper with discussion.

MATERIALS AND METHODS

MRMLC and Its Polytomous Extension

The MRMLC assumes that the probability of a correct response by person l on item i at occasion k can be written as:

$$P\left(U_{ilk}=1|\left(\theta_{l1},...,\theta_{lk}\right),b_i\right)=\frac{\exp\left(\sum_{v=1}^{k}\theta_{lv}-b_i\right)}{1+\exp\left(\sum_{v=1}^{k}\theta_{lv}-b_i\right)}, \quad (1)$$

where U_{ilk} is the response variable with values in {0, 1},θ_{l1} is the initial ability of person l on the first occasion $v = 1$,$\theta_{l2}, ..., \theta_{lk}$ are modifiabilities that correspond to occasion $k > 1$, and b_i is item difficulty Although the MRMLC may be applied to multiple occasions, for clarity, the model will be presented with only two occasions. To simplify the notation, the examinee subscript will not be shown in the following derivations. Using the abbreviated notations P_{i1} and P_{i2} for the probability of a correct item response for Occasions 1 and 2, respectively,

$$P_{i1}\left(\theta_1\right)=\frac{\exp\left(\theta_1-b_i\right)}{1+\exp\left(\theta_1-b_i\right)}, \quad (2)$$

and

$$P_{i2}\left(\theta_1,\theta_2\right)=\frac{\exp\left(\theta_1+\theta_2-b_i\right)}{1+\exp\left(\theta_1+\theta_2-b_i\right)}, \quad (3)$$

Regarding the polytomous items, we use the abbreviated notations P_{ij1} and P_{ij2} to denote the probability of selecting response category j (where $j = 1, ..., h$) of polytomous item i for Occasions 1 and 2, respectively,

$$P_{ij1}\left(\theta_1\right)=\frac{\exp\left(j\theta_1-\sum_{v=1}^{j}b_{iv}\right)}{\sum_{r=1}^{h}\exp\left(r\theta_1-\sum_{v=1}^{r}b_{iv}\right)}, \quad (4)$$

and

$$P_{ij2}\left(\theta_1,\theta_2\right)=\frac{\exp\left[j\left(\theta_1+\theta_2\right)-\sum_{v=1}^{j}b_{iv}\right]}{\sum_{r=1}^{h}\exp\left[r\left(\theta_1+\theta_2\right)-\sum_{v=1}^{r}b_{iv}\right]}, \quad (5)$$

To develop a conditional maximum likelihood estimation method for item parameters in the learning process model, Embretson (1991) constructed a data design structure for item calibration in which item blocks are counterbalanced in several occasions over groups. This data design matrix is needed to determine the occasion on which an item appears for an individual. Every item must be observed on every occasion, but to preserve local independence, an item should be administered only once to an individual across the two occasions. To incorporate Embretson's design structure, two groups of examinees are asked to respond to unique items on two occasions, k_{ig} is now defined as a binary variable to indicate the occasion on which item i is administered to group $g(g = 1, 2)$.

Specifically,

$$k_{ig} = \begin{cases} 1, \text{ if item } i \text{ is administered in group } g \text{ under Occasion } 1, \\ 0, \text{ if item } i \text{ is administered in group } g \text{ under Occasion } 2. \end{cases}$$

Thus, the probability of a response vector $\mathbf{u} = (u_1, ..., u_n)$ in group g, P_g for n items conditional on ability vector (θ_1, θ_2), item difficulty vector $\mathbf{b}$ and item occasion vector k_g, for $k_{1g}, ..., k_{ng}$ is given by:

$$\begin{aligned} &P_g\left(\boldsymbol{U} = \boldsymbol{u} | (\theta_1, \theta_2), \boldsymbol{b}, k_g\right) \\ &= \prod_{i=1}^{n} \left[P_{i1}(\theta_1)^{u_i}(1-P_{i1}(\theta_1))^{1-u_i}\right]^{k_{ig}} \cdot \\ &\qquad \left[P_{i2}(\theta_1, \theta_2)^{u_i}(1-P_{i2}(\theta_1, \theta_2))^{1-u_i}\right]^{1-k_{ig}}, \end{aligned}$$

where $\mathbf{b} = \left(b_1, ..., b_n\right)$.

First, suppose that person l is assigned to a test condition group g that receives items $\mathbf{I}$. For the following considerations, it is assumed that some of the items $\mathbf{I} = \{I_1, ..., I_n\}$ are presented at time point (Occasion) 1, called the "pretest," denoted $\mathbf{I}_1$, and some items are presented at point time 2, called the "posttest," denoted $\mathbf{I}_2$ according to Fischer (2001). The nonempty item subsets $\mathbf{I}_1$ and $\mathbf{I}_2$ may be completely different, may overlap, or may be identical. For convenience, however, a notation is adopted where $\mathbf{I}_1$ and $\mathbf{I}_2$ are considered disjoint subsets of $\mathbf{I}$, $\mathbf{I}_1 = \left\{I_1, ..., I_{n_1}\right\}$ and $\mathbf{I}_2 = \left\{I_{n_1+1}, ..., I_n\right\}$. However, the cases in which $\mathbf{I}_1$ and $\mathbf{I}_2$ overlap are implicitly covered; it suffices to let some $I_a \in \mathbf{I}_1$ have the same parameters as some $I_b \in \mathbf{I}_2$. Let us consider mixed-format tests; specifically, k items $I_1, ..., I_k$ are dichotomous and n_1-k items $I_{k+1}, ..., I_{n_1}$ are polytomous in the pretest; for the posttest, $m-n_1$ items $I_{n_1+1}, ..., I_m$ are dichotomous and $n-m$ items $I_{m+1}, ..., I_n$ are polytomous.

Maximum Likelihood Estimator

Now we consider the problem of likelihood estimation of ability $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The likelihood function of responses is the product of two types of likelihood functions given local independence:

$$L(\boldsymbol{\theta}|\mathbf{U}) = L_d(\boldsymbol{\theta}|\mathbf{U})L_p(\boldsymbol{\theta}|\mathbf{U}), \quad (6)$$

where

$$\begin{aligned} L_d(\boldsymbol{\theta}|\mathbf{U}) = &\left(\prod_{i=1}^{k} P_{i1}(\theta_1)^{u_i} Q_{i1}(\theta_1)^{1-u_i}\right) \cdot \\ &\left(\prod_{i=n_1+1}^{m} P_{i2}(\theta_1, \theta_2)^{v_i} Q_{i2}(\theta_1, \theta_2)^{1-v_i}\right), \end{aligned} \quad (7)$$

and

$$L_p(\boldsymbol{\theta}|\mathbf{U}) = \left(\prod_{i=k+1}^{n_1} \prod_{j=1}^{h} P_{ij1}(\theta_1)^{u_{ij}}\right) \cdot \left(\prod_{i=m+1}^{n} \prod_{j=1}^{h} P_{ij2}(\theta_1, \theta_2)^{v_{ij}}\right), \quad (8)$$

are the likelihood functions of the dichotomous model and the polytomous model of a mixed-format longitudinal test, respectively, in which,

$$Q_{i1}(\theta_1) = 1-P_{i1}(\theta_1), \quad Q_{i2}(\theta_1, \theta_2) = 1-P_{i2}(\theta_1, \theta_2).$$

The response matrix $\mathbf{U}$ contains the responses to dichotomous items u_i, v_i and the responses to polytomous items u_{ij}, v_{ij}. The conventional maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ can be obtained by maximizing the log-likelihood function $\log L(\boldsymbol{\theta}| U)$.

Weighted Likelihood Estimator

Warm (1989) proposed a weighted likelihood estimation (WLE) method for dichotomous IRT model. Compared with the maximum likelihood estimation, Warm's weighted likelihood estimation method can obtain less bias estimation. Penfield and Bergeron (2005) extended this method to the case of the generalized partial credit model (GPCM). The weighted likelihood function of a mixed-type model can be expressed as:

$$w(\theta)L(\boldsymbol{\theta}|\mathbf{U}) = w(\theta)L_d(\boldsymbol{\theta}|\mathbf{U})L_p(\boldsymbol{\theta}|\mathbf{U}),$$

where $w(\theta)$ is the weighting function, $w(\theta) = I^{\frac{1}{2}}$ in one or two parameter models of IRT. $w(\theta)$ is multiplied by the likelihood function $L(\boldsymbol{\theta}|\mathbf{U})$, and the product is maximized. WLE was proved to yield asymptotically normally distributed estimates, with finite variance, and with bias of only $o\left(n^{-1}\right)$.

Item-Weighted Maximum Likelihood Estimator

In this section, we consider the following item-weighted likelihood function:

$$IWL(\boldsymbol{\theta}|\mathbf{U}) = IWL_d(\boldsymbol{\theta}|\mathbf{U}) \cdot IWL_p(\boldsymbol{\theta}|\mathbf{U}), \quad (9)$$

where

$$\begin{aligned} IWL_d(\boldsymbol{\theta}|\mathrm{U}) = &\prod_{i=1}^{k} \left\{P_{i1}(\theta_1)^{u_i} \cdot Q_{i1}(\theta_1)^{1-u_i}\right\}^{w_i(\boldsymbol{\theta})} \cdot \\ &\prod_{i=n_1+1}^{m} \left\{P_{i2}(\theta_1, \theta_2)^{v_i} \cdot Q_{i2}(\theta_1, \theta_2)^{1-v_i}\right\}^{w_i(\boldsymbol{\theta})}, \end{aligned} \quad (10)$$

and

$$IWL_p(\boldsymbol{\theta} \mid \mathrm{U}) = \prod_{i=k+1}^{n_1} \left\{ \prod_{j=1}^{h} P_{ij1}(\theta_1)^{u_{ij}} \right\}^{w_i(\boldsymbol{\theta})} \cdot \prod_{i=m+1}^{n} \left\{ \prod_{j=1}^{h} P_{ij2}(\theta_1, \theta_2)^{v_{ij}} \right\}^{w_i(\boldsymbol{\theta})}, \quad (11)$$

are the item-weighted likelihood functions of the dichotomous model and the polytomous model of a mixed-format longitudinal test, respectively. Here the weight vector:

$(w_1(\boldsymbol{\theta}), \ldots, w_n(\boldsymbol{\theta}))$ satisfy $w_i(\boldsymbol{\theta}) > 0$ for each i and $\sum_{i=1}^{n} w_i(\boldsymbol{\theta}) = 1$.

Note that,

$$w_i(\boldsymbol{\theta}) = \frac{I_i(\boldsymbol{\theta})}{I(\boldsymbol{\theta})}, \quad \text{for all } i \in \{1, 2, \ldots, n\}, \quad (12)$$

where $I_i(\boldsymbol{\theta})$ is the information function of item i given as:

$$I_i(\boldsymbol{\theta}) = \begin{cases} P_i(\boldsymbol{\theta})Q_i(\boldsymbol{\theta}), & \text{for dichotomous item } i, \\ \sum_{j=1}^{h} j^2 P_{ij}(\boldsymbol{\theta}) - \left(\sum_{j=1}^{h} jP_{ij}(\boldsymbol{\theta})\right)^2, & \text{for polytomous item } i. \end{cases}$$

P_i is the probability of a correct response to item i, $Q_i = 1 - P_i$, P_{ij} is the probability of selecting response category j (where $j = 1, \ldots, h$) of polytomous item i, and $I(\boldsymbol{\theta}) = \sum_{i=1}^{n} I_i(\boldsymbol{\theta})$ is the test information function consisting both dichotomous and polytomous items (Muraki, 1993). Using the information ratio of each item to the test at a certain ability level, the weights of items are determined.

In IRT, the item and test information functions relate to how well an examinee's ability is being estimated over the whole ability scale; they are usually used to calculate the standard error of measurement and the reliability. Since the test information is a function of proficiency (or whatever trait or skill is measured) and the items on the test, the expression of the proposed weights involves the ability level $\boldsymbol{\theta}$ and item characteristic parameters. The weights may be "adaptive" in the sense that they are allowed to be estimated based on the ability level and individual test items. By using the information ratio of each item to the test to determine the weights, so the more information an item has at a certain ability level, the larger weight could be assigned to it. According to the proposed weighting method, the weight for the polytomous item is then larger than that for the dichotomous item and the weights for the same type item are different due to the difference between the amounts of item information. The weight assigned to each item just indicates its contribution to the precision for ability parameter estimation. This item weighting scheme maximizes the information obtained from both different types of items and different items of the same type and may lead to more accurate estimates of the latent trait than equally weighting all items. If each item with same scoring procedure has same item information at a certain latent trait level, the weights are equal between them. Hence, the proposed item-weighted likelihood method may be an extension of the method proposed by Tao et al. (2012). The item-weighted likelihood estimator (IWLE) can be obtained by maximizing the item-weighted log-likelihood function log $IWL(\boldsymbol{\theta}|\mathbf{U})$ (for derivation details, see **Supplementary Appendix A**). Maximum likelihood estimator (Lord, 1983) was shown to have bias of $O\left(n^{-1}\right)$. When the weights are determined at a certain ability level, with some assumptions made by Lord (1983), the bias of the item-weighted maximum likelihood estimation also has bias of $O\left(n^{-1}\right)$. The approach and techniques of this derivation were taken from, and parallel closely, the derivations in Lord (1983). The asymptotic properties of IWLM can be obtained by generalizing those of Bradley and Gart (1962) (for more details, see **Supplementary Appendix B**).

Type-Weighted Maximum Likelihood Estimator

In contrast to the MLE, the type-weighted maximum likelihood estimator (TWLE) yields usable ability estimator for mixed-type tests composed of both dichotomous and polytomous items (Sun et al., 2012). The type-weighted likelihood function of a mixed-type model can be expressed as:

$$TWL(\boldsymbol{\theta}|\mathbf{U}) = L_d(\boldsymbol{\theta}|\mathbf{U})^{\tilde{w}_1(\boldsymbol{\theta})} L_p(\boldsymbol{\theta}|\mathbf{U})^{\tilde{w}_2(\boldsymbol{\theta})},$$

where

$$\tilde{w}_1(\boldsymbol{\theta}) = \left(\frac{I_d(\boldsymbol{\theta})}{I(\boldsymbol{\theta})}\right)^{\alpha}, \quad \tilde{w}_2(\boldsymbol{\theta}) = \left(\frac{I_p(\boldsymbol{\theta})}{I(\boldsymbol{\theta})}\right)^{\beta},$$

$$I = I_d + I_p,$$

$I_d = \sum_{i=1}^{k} I_i + \sum_{i=n_1+1}^{m} I_i$, and $I_p = \sum_{i=k+1}^{n_1} I_i + \sum_{i=m+1}^{n} I_i$, are test information of the dichotomous and polytomous model based on the longitudinal model, respectively. According to the weighting scheme proposed by Sun et al. (2012), the ratio parameters α, β determined to make sure that the weight assigned to the polytomously scored item is larger than that assigned to the dichotomously scored item. Three steps are needed to determine the ratio parameters α, β and the two weights. First, we obtain the ML estimator $\hat{\theta}_0$ and take it as the initial estimator. Second, if $I_d\left(\hat{\theta}_0\right) < I_p\left(\hat{\theta}_0\right)$, the two ratio parameters are all equal to 1. Otherwise, we may set α and β to be a small value ε (such as ε < 0.4) to make sure $I_d\left(\hat{\theta}_0\right) < I_p\left(\hat{\theta}_0\right)$. Then, no change is needed for either α or β if $\tilde{w}_1\left(\hat{\theta}_0\right) < \tilde{w}_2\left(\hat{\theta}_0\right)$. Otherwise, we may increase α in increments of 0.05 or less, or decrease β in increments of 0.05 or less. We adjust α and β to ensure $\tilde{w}_1\left(\hat{\theta}_0\right) < \tilde{w}_2\left(\hat{\theta}_0\right)$. Third, we maximize the type-weighted log-likelihood function log $TWL(\boldsymbol{\theta}|\mathbf{U})$ to obtain $\hat{\theta}$ with the obtained α and β values from the above. If $\tilde{w}_1(\hat{\theta}) < \tilde{w}_2(\hat{\theta})$, the $\hat{\theta}$ is the TWLE. Otherwise, the ratio parameters should be adjusted continually basing on the above process until $\tilde{w}_1(\hat{\theta}) < \tilde{w}_2(\hat{\theta}\ )$.

The above three-weighted estimations TWLE, WLE, and IWLE have different weighting schemes. For TWLE, the larger weights are assigned to the polytomous items and the smaller weights are assigned to the dichotomous items. This method only assigns different weights to different item types, instead of assigning different weights to different items, thus items of the same type all have the same weight. However, different items of the same type may have different information about a certain latent trait level; the same weights assigned to the same-type items may not be statistically optimal in terms of the precision and accuracy of ability estimation due to neglecting the difference in the individual item contribution. The proposed IWLE assigns different weights to different items in accordance with the amount of the information an item provides at a certain latent trait level. Using the information ratio of each item to the test, the weights of items are determined. This improved IWLE procedure that incorporates item weights in likelihood functions for the ability parameter estimation may increase measurement precision. The WLE provides a bias correction to the maximum likelihood method. The weight function is multiplied by the likelihood function $L(\boldsymbol{\theta}|\mathbf{U})$ in the WLE method, which provides a correction to the maximum likelihood estimation method by solving an weighted, log-likelihood equation. The WLE and IWLE are both consistent and asymptotically normal with mean zero and a variance-covariance matrix, and the bias of the estimators is of order n^{-1}.

SIMULATION STUDY 1

Simulation Design

In this section, the performance of the three weighting methods, the WLE, the type-weighted likelihood estimation (TWLE), and IWLE are compared. To investigate the effects of the test-length and the proportion of dichotomous and polytomous items in a mixed-format test on the properties of the $\boldsymbol{\theta}$ estimators, nine artificial tests were constructed at each time point, three of them short (10 items with 7, 5, and 3 dichotomous items), three medium (30 items with 20, 15, and 10 dichotomous items), and three long (60 items with 40, 30, and 20 dichotomous items). In the simulation, the 3 levels of test length were representative of those encountered in measuring settings using fixed-length tests. The 3 levels of proportion of dichotomous and polytomous items ($\lambda = 2, 1, 0.5$) were selected, so that we may have a thorough investigation into the property of different weighting methods.

The item parameters and ability parameters are set as follows. The difficulty parameters of the dichotomous items were randomly generated from the standard normal distribution $N(0, 1)$. The polytomously scored items with four-category were constructed. The step parameters of each polytomous item were randomly generated from four normal distributions:

$\mathrm{b}_{i1} \sim N(-1.5, 0.2)$, $b_{i2} \sim N(-0.5, 0.2)$, $b_{i3} \sim N(0.5, 0.2)$, and $b_{i4} \sim N(1.5, 0.2)$.

This pattern of location parameters centers items on zero and thus centers the test on zero. In the simulation, 17 equally spaced θ_1 values were considered, ranging from -4.0 to 4.0 in increments of 0.5. We set 3 values of θ_2 (0.6, 0.8, and 1.0) for 3 different initial ability levels: high (value of θ_1 larger than 2), medium (value of θ_1 between -2 and 2), and low (value of θ_1 smaller than -2), respectively. Thus, a high initial ability will have low gain, a medium initial ability will have moderate gain, and a low initial ability will have high gain. At each level of (θ_1, θ_2), $N(N = 1000)$ replications were administered for all 9 tests. In each replication, the dichotomous item responses were simulated according to the MRMLC model as presented in Equations 2 and 3, and the polytomous item responses were simulated according to the PCM as presented in Equations 4 and 5. For the tests containing response patterns consisting of all correct responses for dichotomous items and all 4s for polytomous items or all incorrect responses for dichotomous items and all 4s, the Newton-Raphson algorithm cannot converge, and thus the likelihood estimators could not be obtained. These response patterns were removed from the analysis, and the same item responses were scored using the WLE, TWLE, and IWLE procedures. In the simulation, the $\boldsymbol{\theta}$ in the weight for each item is taken as $\hat{\boldsymbol{\theta}}$, the MLE of $\boldsymbol{\theta}$. All levels of the number of items, the proportion of dichotomous and polytomous items, and the number of examinee were crossed, resulting in 27 conditions of test properties at each time point. For each of the 27 conditions of test properties, the WLE, TWLE, and IWLE were obtained for each of the response patterns.

Evaluation Criteria

The bias, absolute bias, root mean squared error (RMSE) and root mean squared difference (RMSD) of the ability estimates were used as evaluation criteria to examine all estimation methods. The absolute bias is calculated using Equation 13. In Equation 13, θ denotes the true ability value and $\hat{\theta}_l$ the corresponding ability estimate for the l th replication.

$$|Bias| = |\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_l - \theta\right)| \tag{13}$$

RMSE and RMSD are calculated using Equation 14 and 15, respectively:

$$RMSE = \sqrt{\frac{1}{N}\sum_{l=1}^{N}\left(\hat{\theta}_l - \theta\right)^2}, \tag{14}$$

$$RMSD = \sqrt{\frac{1}{N}\sum_{l=1}^{N}\left(\hat{\theta}_l - \frac{1}{N}\sum_{l=1}^{N}\hat{\theta}_l\right)^2}. \tag{15}$$

N is the number of replications. In simulation studies, we fix the number of replications at 1000, that is, $N = 1000$.

Results of Simulation

The weights of IWLE for 6 dichotomous and 3 polytomous items are shown in **Figures 1**, **2** The purpose of these figures is to give more intuition in terms of our item weighting scheme. The weights are based on the individual test items and the ability level, with θ_1 ranging from -4.0 to 4.0 and 3 values of θ_2(0.6, 0.8, and 1.0). We can find that the different items are designed with

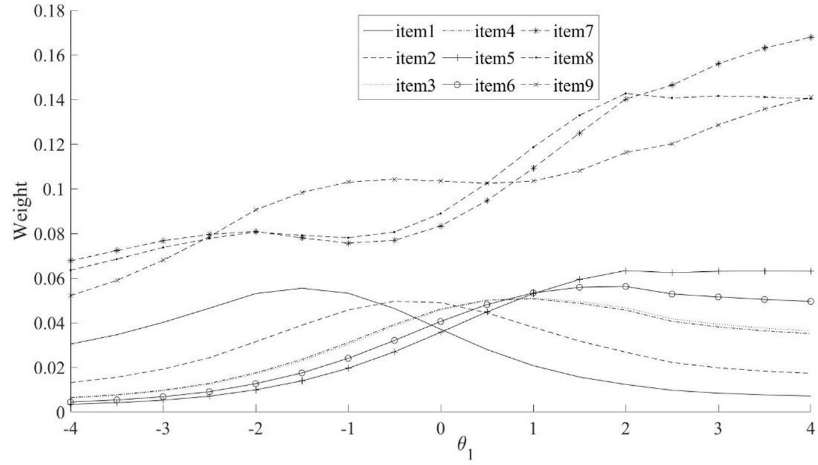

FIGURE 1 | The weights of IWLE based on θ_1 for dichotomous items (item 1 to 6) and polytomous items (items 7 to 9) in test 1.

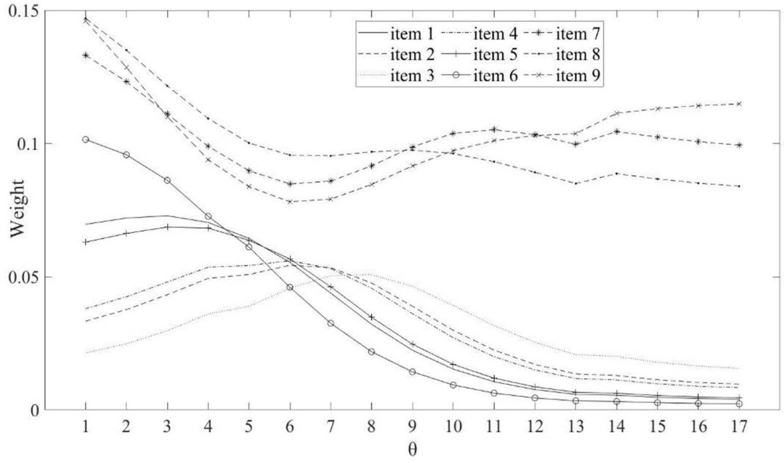

FIGURE 2 | Weights based on θ ($\theta = (\theta_1, \theta_2)$) at 17 ability levels for dichotomous items (item 1 to 6) and polytomous items (items 7 to 9) in test 2.

TABLE 1 | Correlation between the estimated abilities and the true abilities for all three weighting estimation methods under nine conditions.

		Test								
N	**Method**	**7d+3p**	**5d+5p**	**3p+7d**	**20d+10p**	**15d+15p**	**10d+20p**	**40d+20p**	**30d+30p**	**20d+40p**
1000	IWLE	0.8685	0.9068	0.9189	0.9478	0.9548	0.9593	0.9663	0.9609	0.9822
	WLE	0.8189	0.8378	0.8608	0.9246	0.9375	0.9470	0.9640	0.9606	0.9716
	TWLE	0.8001	0.8344	0.8542	0.9216	0.9360	0.9451	0.9612	0.9796	0.9711

n_1 d $+ n_2$p *means the* $(n_1 + n_2)$ *-item test with* n_1 *dichotomous items and* n_2 *polytomous items.*

different weights. In addition, the weights assigned to polytomous items are larger than that of dichotomous items.

Table 1 shows the correlation between the estimated abilities and the true abilities for all three weighting estimation methods under nine conditions. The higher degree of correlation obtained by the IWLE ability estimates indicates that the IWLE produces better quality ability estimates. The results in **Table 1** indicate that the longer tests provide higher correlation between the estimated abilities and the true abilities. In the tests with the same length, higher proportion of polytomous and dichotomous items also provide higher correlation between the estimated abilities and the true abilities.

The simulation results of 3 test lengths show similar trends for the three weighting estimators: WLE, TWLE, and IWLE. Due to page limitation, only those for the 30-item test are presented. The complete results can be obtained from the author.

Tables 2–7 displays the obtained values of absolute bias, and RMSD for WLE, TWLE, and IWLE at 17 different levels of initial ability $\theta_1(-4, -3.5,, 3.5, 4)$ and 3 different levels of growth $\theta_2(0.6, 0.8, 1)$ in the simulation scenarios.

Examining these results, the following general trends are observed. The absolute bias are all nearly to zero for three estimators when $|\theta_1| < 2$, or $\theta_2 = 0.8$, but IWLE has a considerably less absolute bias than the other two estimators when $|\theta_1| > 2$ or $\theta_2 = 0.6$ and 1. We note that in the 3 simulation scenarios the absolute bias of IWLE is slightly larger than that of WLE at some level of θ_1 when $|\theta_1| < 2$, but is considerably smaller than that of WLE at the low and the high levels of

TABLE 2 | Absolute bias and root mean squared difference for WLE, TWLE, and IWLE at 17 different levels of initial ability on 20d+10p.

20d+10p	**Methods**					
***N* = 1000**	**IWLE**		**WLE**		**TWLE**	
θ_1	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**
−4.0	0.5615	1.7379	1.1045	3.5101	1.1596	3.5370
−3.5	0.3221	1.3011	0.4678	2.3189	0.5249	2.3495
−3.0	0.1402	0.9134	0.1582	1.3797	0.1783	1.4074
−2.5	0.0342	0.5038	0.0361	0.5118	0.0494	0.5675
−2.0	0.0162	0.4809	0.0130	0.4401	0.0158	0.4931
−1.5	0.0047	0.4384	0.0005	0.4061	0.0054	0.4494
−1.0	0.0045	0.4020	0.0004	0.3821	0.0049	0.4237
−0.5	0.0047	0.3662	0.0119	0.3570	0.0053	0.3943
0.0	0.0092	0.3718	0.0114	0.3433	0.0107	0.3784
0.5	0.0071	0.3707	0.0041	0.3456	0.0083	0.3740
1.0	0.0076	0.3654	0.0039	0.3378	0.0078	0.3670
1.5	0.0179	0.3834	0.0164	0.3675	0.0189	0.4095
2.0	0.0112	0.4025	0.0039	0.3764	0.0169	0.4272
2.5	0.0205	0.4133	0.0187	0.4400	0.0455	0.4926
3.0	0.0190	0.5846	0.0282	0.6321	0.0508	0.6763
3.5	0.2811	1.1297	0.3414	2.0295	0.3876	2.0387
4.0	0.3805	1.3812	0.6268	2.8406	0.6900	2.8470

20 d + 10p *means the 30-item test with 20 dichotomous items and 10 polytomous items.*

TABLE 3 | Absolute bias and root mean squared difference for WLE, TWLE, and IWLE at 3 different levels of growth on 20d+10p.

20d+10p	**Methods**					
***N* = 1000**	**IWLE**		**WLE**		**TWLE**	
θ_2	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**
1	0.4219	1.9411	1.0146	3.7642	1.0438	3.7842
1	0.2559	1.5026	0.4351	2.5015	0.4801	2.5460
1	0.0907	1.0243	0.1419	1.4468	0.1536	1.4836
1	0.0215	0.6990	0.0296	0.6428	0.0399	0.7196
0.8	0.0145	0.6201	0.0050	0.5732	0.0225	0.6349
0.8	0.0078	0.5947	0.0071	0.5492	0.0101	0.6095
0.8	0.0188	0.5555	0.0144	0.5204	0.0194	0.5697
0.8	0.0055	0.5402	0.0073	0.4952	0.0075	0.5518
0.8	0.0054	0.5365	0.0042	0.5032	0.0100	0.5435
0.8	0.0283	0.5380	0.0236	0.4971	0.0276	0.5439
0.8	0.0056	0.5670	0.0023	0.5213	0.0059	0.5761
0.8	0.0301	0.6168	0.0197	0.5684	0.0330	0.6224
0.8	0.0726	0.7236	0.0504	0.7475	0.0779	0.8220
0.6	0.0782	0.9579	0.1165	1.2499	0.1224	1.3136
0.6	0.2395	1.4137	0.4164	2.3609	0.4538	2.3896
0.6	0.3946	2.2477	0.9091	4.3898	0.9462	4.3887
0.6	0.7397	2.8244	1.7386	5.9953	1.7629	5.9489

20 d + 10p *means the 30-item test with 20 dichotomous items and 10 polytomous items.*

TABLE 4 | Absolute bias and root mean squared difference for WLE, TWLE, and IWLE at 17 different levels of initial ability on 15d+15p.

15d+15p	Methods					
***N* = 1000**	**IWLE**		**WLE**		**TWLE**	
θ_1	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**
−4.0	0.4756	1.4673	0.8064	2.9131	0.8465	2.9195
−3.5	0.1589	0.8002	0.1603	1.2443	0.2009	1.2795
−3.0	0.0605	0.6537	0.0866	0.9311	0.0965	0.9520
−2.5	0.0104	0.4371	0.0167	0.4411	0.0152	0.4712
−2.0	0.0229	0.4076	0.0266	0.3855	0.0384	0.421
−1.5	0.0163	0.3677	0.0102	0.3535	0.0146	0.3791
−1.0	0.0035	0.3349	0.0057	0.3223	0.0039	0.3425
−0.5	0.0092	0.3433	0.0038	0.3295	0.0093	0.3509
0.0	0.0038	0.3336	0.0015	0.3168	0.0039	0.3375
0.5	0.0038	0.3334	0.0050	0.3199	0.0074	0.3365
1.0	0.0001	0.3306	0.0054	0.3111	0.0038	0.3398
1.5	0.0040	0.3578	0.0003	0.3333	0.0024	0.3553
2.0	0.0160	0.3776	0.0113	0.3611	0.0164	0.3796
2.5	0.0300	0.4917	0.0248	0.5686	0.0348	0.5867
3.0	0.1358	0.6881	0.1997	0.9829	0.1484	1.0048
3.5	0.2461	1.0279	0.2718	1.7932	0.3194	1.8233
4.0	0.4730	1.5333	0.8051	3.1026	0.8775	3.1494

15 d + 15p *means the 30-item test with 15 dichotomous items and 15 polytomous items.*

TABLE 5 | Absolute bias and root mean squared difference for WLE, TWLE, and IWLE at 3 different levels of growth on 15d+15p.

15d+15p	Methods					
***N* = 1000**	**IWLE**		**WLE**		**TWLE**	
θ_2	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**
1	0.3555	0.5993	0.7385	3.0097	0.7626	3.0257
1	0.0864	0.9536	0.1136	1.3520	0.1371	1.4041
1	0.0877	0.7725	0.1020	1.0090	0.1020	1.0338
1	0.0030	0.5857	0.0053	0.6040	0.0049	0.6358
0.8	0.0013	0.5067	0.0063	0.5133	0.0025	0.5528
0.8	0.0103	0.4933	0.0030	0.4738	0.0132	0.5085
0.8	0.0022	0.4669	0.0070	0.4513	0.0029	0.4735
0.8	0.0162	0.4728	0.0123	0.4462	0.0178	0.4820
0.8	0.0087	0.4572	0.0003	0.4324	0.0096	0.4603
0.8	0.0161	0.4787	0.0129	0.4531	0.0164	0.4796
0.8	0.0177	0.4941	0.0178	0.4640	0.0151	0.4906
0.8	0.0407	0.5626	0.0328	0.5328	0.0490	0.5632
0.8	0.0473	0.5864	0.0487	0.5805	0.0476	0.6206
0.6	0.0617	0.8184	0.0618	0.9583	0.0629	0.9855
0.6	0.1824	1.3755	0.3231	2.3313	0.3333	2.3686
0.6	0.3534	2.0572	0.8409	4.0178	0.8563	4.0486
0.6	0.5312	2.8113	1.4114	6.0129	1.4160	6.0353

15 d + 15p *means the 30-item test with 15 dichotomous items and 15 polytomous items.*

ability. IWLE consistently displays the level of absolute bias that is smaller than that of TWLE, especially substantially smaller than that of TWLE at the low and the high levels of ability. In addition, the absolute bias of WLE is less than that of TWLE at the extremes of ability level. However, the changes are observed when the proportion of the dichotomous and polytomous items in mixed-type test is changed. With the number of polytomous items increased, the absolute bias produced by TWLE and WLE are more similar, even TWLE produces a little larger absolute bias than WLE at the extremes of ability level. The similar

TABLE 6 | Absolute bias and root mean squared difference for WLE, TWLE, and IWLE at 17 different levels of initial ability for 10 d + 20p.

10d+20p	**Methods**					
***N* = 1000**	**IWLE**		**WLE**		**TWLE**	
θ_1	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**
−4.0	0.4139	1.4081	0.7042	2.7417	0.8002	2.9845
−3.5	0.1748	0.8937	0.2045	1.4292	0.2356	1.5412
−3.0	0.0547	0.5415	0.0580	0.7017	0.0750	0.7492
−2.5	0.0108	0.4168	0.0155	0.4233	0.0223	0.4351
−2.0	0.0035	0.3677	0.0073	0.3540	0.0047	0.3691
−1.5	0.0020	0.3567	0.0023	0.3459	0.0038	0.3579
−1.0	0.0275	0.3360	0.0237	0.3253	0.0241	0.3421
−0.5	0.0211	0.3281	0.0145	0.3196	0.0200	0.3295
0	0.0039	0.3087	0.0040	0.2968	0.0047	0.3130
0.5	0.0010	0.3030	0.0007	0.2883	0.0018	0.3053
1.0	0.0089	0.2886	0.0054	0.2798	0.0115	0.2903
1.5	0.0066	0.3048	0.0000	0.2963	0.0210	0.3024
2.0	0.0081	0.3392	0.0073	0.3282	0.0109	0.3391
2.5	0.0182	0.3904	0.0234	0.3912	0.0327	0.4157
3.0	0.0205	0.5022	0.0288	0.5837	0.0436	0.5952
3.5	0.1616	0.8334	0.1687	1.3778	0.2042	1.3954
4.0	0.3306	1.2024	0.4565	2.3132	0.5022	2.3207

10 d + 20p *means the 30-item test with 10 dichotomous items and 20 polytomous items.*

TABLE 7 | Absolute bias and root mean squared difference for WLE, TWLE, and IWLE at 3 different levels of growth for 10 d + 20p.

10d+20p	**Methods**					
***N* = 1000**	**IWLE**		**WLE**		**TWLE**	
θ_2	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**	**Abs.Bias**	**RMSD**
1	0.3451	1.5350	0.6859	2.8275	0.7519	3.0801
1	0.1355	0.9881	0.1899	1.4906	0.2130	1.6003
1	0.0359	0.6451	0.0560	0.7772	0.0650	0.8253
1	0.0111	0.5061	0.0188	0.5386	0.0253	0.5580
0.8	0.0057	0.5151	0.0109	0.4972	0.0078	0.5181
0.8	0.0056	0.4737	0.0002	0.4611	0.0095	0.4753
0.8	0.0198	0.4510	0.0196	0.4408	0.0199	0.4593
0.8	0.0162	0.4532	0.0116	0.4398	0.0168	0.4536
0.8	0.0168	0.4385	0.0167	0.4201	0.0169	0.4428
0.8	0.0008	0.4142	0.0028	0.3979	0.0032	0.4152
0.8	0.0276	0.4301	0.0204	0.4293	0.0277	0.4389
0.8	0.0089	0.4325	0.0140	0.4250	0.0119	0.4349
0.8	0.0408	0.5412	0.0375	0.5347	0.0466	0.5510
0.6	0.0123	0.6007	0.0151	0.6038	0.0270	0.6342
0.6	0.1462	1.0715	0.2162	1.7098	0.2350	1.7319
0.6	0.3919	1.8900	0.8098	3.5369	0.8265	3.5700
0.6	0.5514	2.4670	1.3889	5.0994	1.4169	5.1159

10 d + 20p *means the 30-item test with 10 dichotomous items and 20 polytomous items.*

change patterns are also observed for RMSD produced by three estimators. The RMSD of IWLE is slightly larger than that of WLE at some level of θ_1 when $|\theta_1| < 2$, but is considerably smaller than that of WLE and TWLE at the low and the high levels of ability.

To investigate the performance of the proposed IWLE method, an simulation study was conducted for the comparison of the five estimators: MLE, MAP [with a non-informative prior distribution $U(4, 4)$] WLE, TWLE, and IWLE under the above simulation condition. **Figures 3–8** show the results of

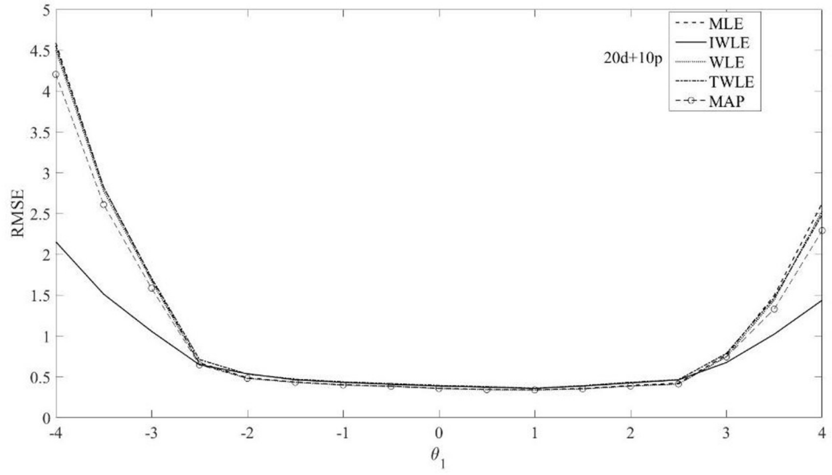

FIGURE 3 | RMSE of the five θ_1 estimation methods MLE, MAP, IWLE, WLE, and TWLE for 20d+10p.

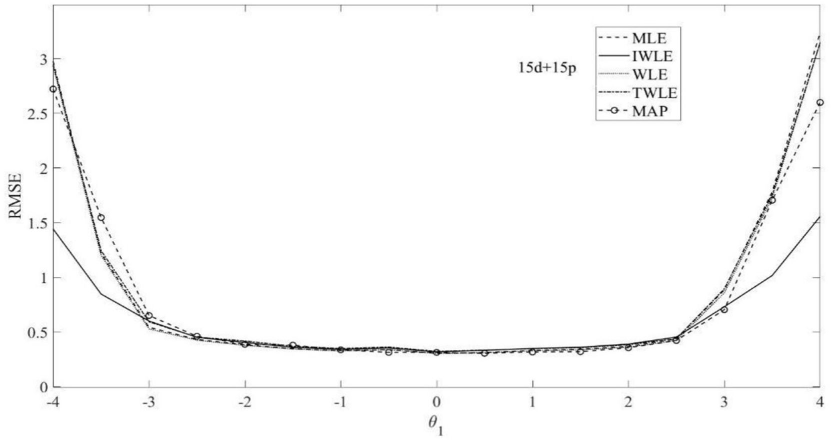

FIGURE 4 | RMSE of the five θ_1 estimation methods MLE, MAP, IWLE, WLE, and TWLE for 15d+15p.

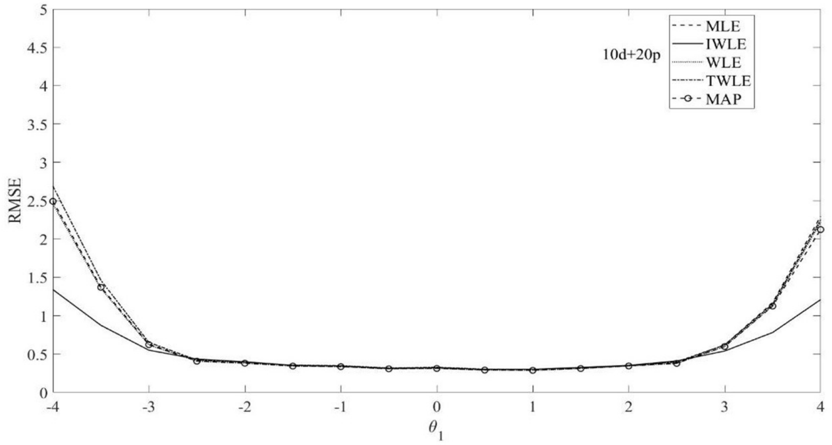

FIGURE 5 | RMSE of the five θ_1 estimation methods MLE, MAP, IWLE, WLE, and TWLE for 10d+20p.

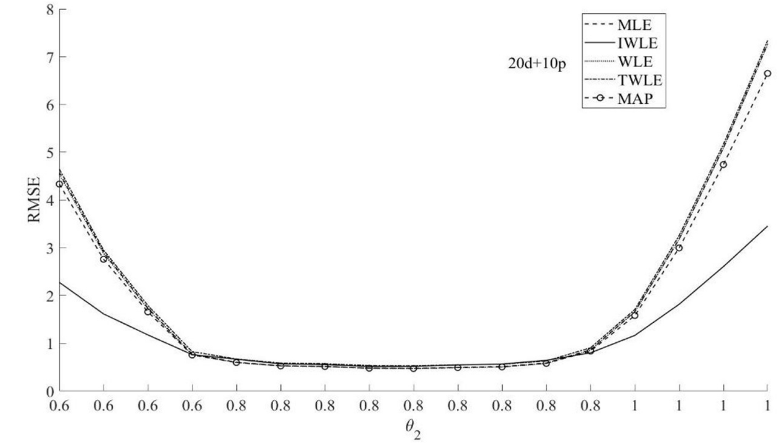

FIGURE 6 | RMSE of the five θ_2 estimation methods MLE, MAP, IWLE, WLE, and TWLE for 20d+10p.

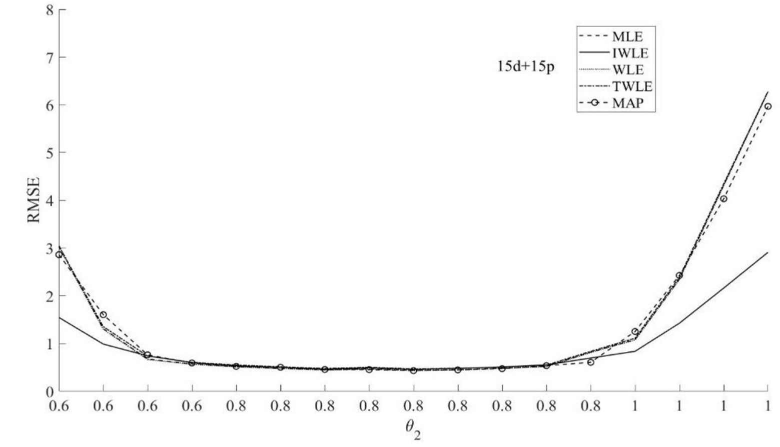

FIGURE 7 | RMSE of the five θ_2 estimation methods MLE, MAP, IWLE, WLE, and TWLE for 15d+15p.

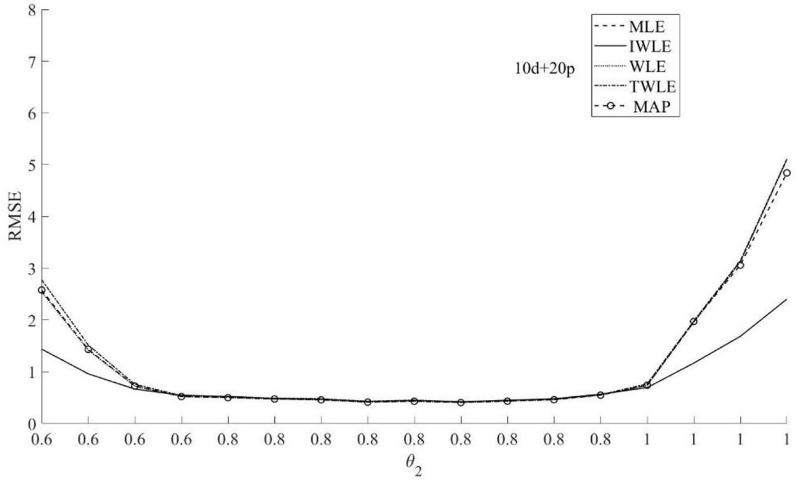

FIGURE 8 | RMSE of the five θ_2 estimation methods MLE, MAP, IWLE, WLE, and TWLE for 10d+20p.

RMSE calculated from 30-item test in the following simulation scenarios:

(1). 30-item test includes 20 dichotomous items and 10 polytomous items (20 d + 10p).
(2). 30-item test includes 15 dichotomous items and 15 polytomous items (15 d + 15p).
(3). 30-item test includes 10 dichotomous items and 20 polytomous items (10 d + 20p).

The RMSE presented in **Figures 3–5** show that among the five θ_1 estimation methods, IWLE has a slight large RMSE when $|\theta_1| < 2$, but is considerably smaller than that of MLE, MAP, WLE and TWLE at extreme levels of the latent trait. The RMSE of WLE is very similar to that of MLE and TWLE. EAP has lower RMSE than MLE, WLE, TWLE, and IWLE in the middle of the ability range because of the shrinkage. The RMSE plotted in **Figures 6–8** shows the similar change patterns for θ_2.

The proposed IWLE method outperforms the MLE, MAP, WLE and TWLE in terms of controlling the absolute bias, RMSE, and RMSD at the low and the high levels of ability, but has a slight large RMSE and RMSD in the middle range of the ability scale.

In general, test length had a dramatic impact on the relative performance of the five estimators. We can observe the strongest differences between the five estimators are obtained when the test length is short. The absolute bias, RMSE, and RMSD of five estimation methods have a slightly decrease with the length of test increased. The proportion of dichotomous and polytomous items in a mixed-format test appears to affect the absolute bias, RMSE, and RMSD of five estimation methods.

SIMULATION STUDY 2

When we only care about the ability of the examinee without considering the ability growth at multiple time points, the

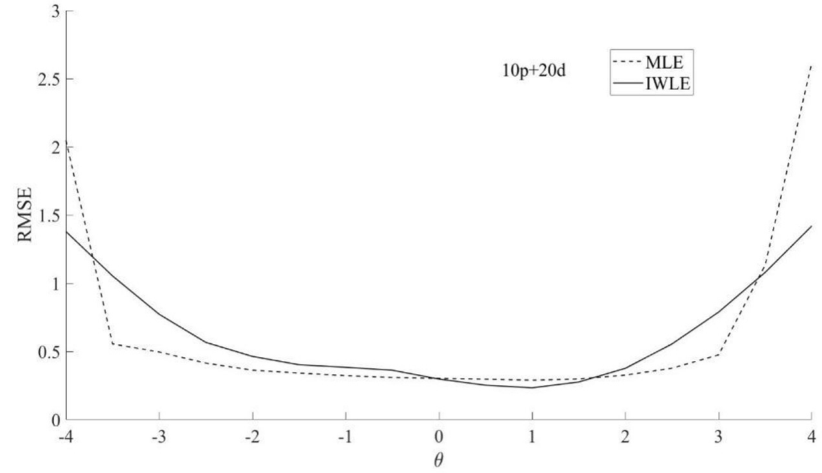

FIGURE 9 | RMSE of the two θ estimation methods MLE and IWLE for 10p+20d.

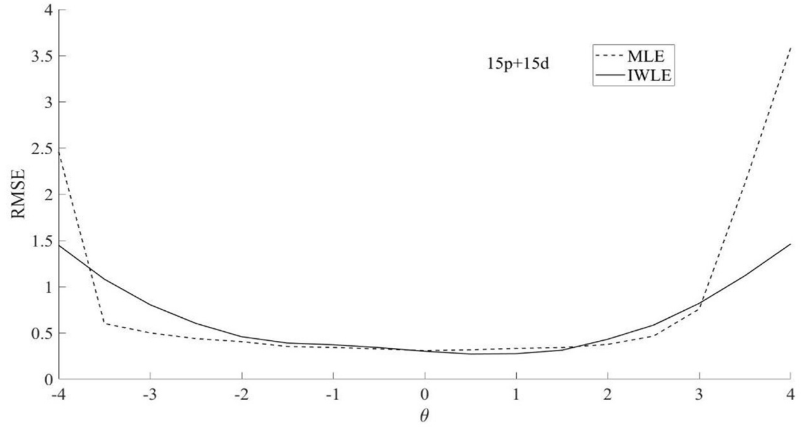

FIGURE 10 | RMSE of the two θ estimation methods MLE and IWLE for 15p+15d.

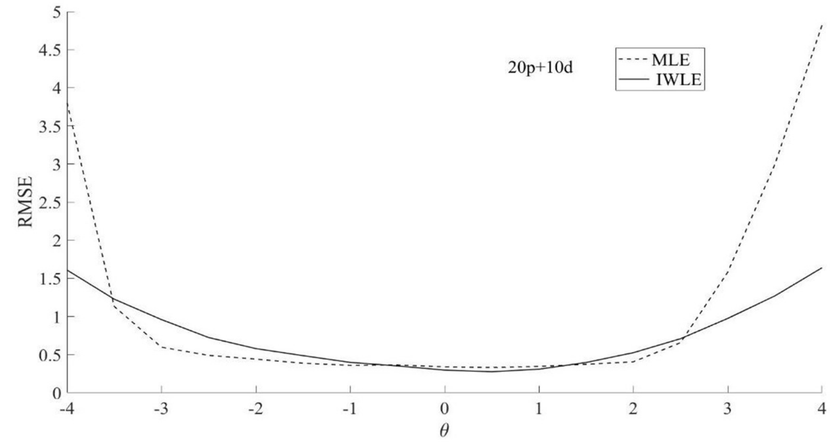

FIGURE 11 | RMSE of the two θ estimation methods MLE and IWLE for 20p+10d.

unidimensional IRT models are the focus of many educational psychometrists. In fact, our IWLE method can't only be used to analyze multidimensional IRT models, but also can be implemented for unidimensional IRT models. In this simulation study, we evaluate the accuracy of the IWLE method in the unidimensional models.

The proposed IWLE method is applied to the unidimensional IRT models for mixed-format test that is the combination of the two-parameter logistic model and the partial-credit model. We consider the following item-weighted likelihood function:

$$IWL(\theta|\mathbf{U}) = IWL_d(\theta|\mathbf{U}) \cdot IWL_p(\theta|\mathbf{U}),$$

where

$$IWL_d(\theta|\mathbf{U}) = \prod_{i=1}^{k} \left\{P_i(\theta)^{u_i} \cdot Q_i(\theta)^{1-u_i}\right\}^{w_i(\theta)},$$

and

$$IWL_p(\theta|\mathbf{U}) = \prod_{i=k+1}^{24^n} \prod_{j=1}^{h}; \left\{P_{ij}(\theta)^{u_{ij}}\right\}^{w_i(\theta)},$$

$P_i(\theta)$ is determined by dichotomously scored items; $P_{ik}(\theta)$ is determined by polytomously scored items. Here the weight $w_i(\theta)$ assigned to item i is defined as equation 4, and $\sum_{i=1}^{n}\ w_i(\theta) = 1$. The 3 levels of test length (10 items, 30 items and 60 items) and the 3 levels of proportion of dichotomous and polytomous items ($\lambda = 2, 1, 0.5$) were selected. The item parameters were generated similar to simulation 1, and 17 equally spaced θ_1 values were considered, ranging from −4.0 to 4.0 in increments of 0.5.

The simulation results of three test lengths show similar trends. The proposed IWLE method outperforms the MLE in terms of the absolute bias, RMSE and RMSD at the low and high levels of ability. However, the IWLE has a slight large absolute bias, RMSE and RMSD in the middle range of the ability scale compared with the MLE. **Figures 9–11** show the results of RMSE calculated from 30-item test. According to the simulation results, we find that the IWLE can also be applied to the general unidimensional IRT models for tests composed of both dichotomous and polytomous items.

DISCUSSION AND CONCLUSION

In this study, an improved IWLE procedure that incorporates item weights in likelihood functions for the ability parameter estimation is proposed. The weights may be "adaptive" in the sense that they are allowed to be estimated with the ability level and individual test items. We assign different weights to different items in accordance with the amount of the information an item provides at a certain latent trait level. Using the information ratio of each item to the test, the weights of items are determined. We also give the rigorous derivations for asymptotic properties and the bias of IWL estimators. The results from the simulation study clearly demonstrate that the proposed IWLE method outperforms the usual, MLE, MAP, WLE and TWLE in terms of controlling absolute bias, RMSE, and RMSD especially at low and high ability levels. Latent trait estimation is one of the most important components in IRT, but when an examinee scores high (or low) in a test, we known that the examinee is high (or low) on the trait but we do not have a very precise estimate of how high (or low). It could be considerably higher (or lower) than the test instrument' scale reaches. In the case, improving latent trait estimation especially at extreme levels of ability scale is worthy of attention.

Improving latent trait estimation is always important in longitudinal survey assessments, such as the Early Childhood Longitudinal Study (ECLS) and the PISA (von Davier and Xu, 2011), which aims at tracking growth of a representative sample of the target population over time. The proposed weighting scheme also can be applied in the general unidimensional item response models. Other issues should be further explored. First, the proposed weighting scheme could be generalized to other application settings where latent ability needs to be estimated for each person such as computerized adaptive testing (CAT). Second, although the Rasch model and the PCM are commonly used in practical tests, there are other more general item response models, for instance the three-parameter logistic (3PL) model and the generalized partial credit model. Therefore, it is worth studying the extension of the IWLE to these more complex models, with different test lengths and sample sizes. Third, more than two occasions can be considered in longitudinal study, so the proposed weighting method can be generalized to deal with more general situations. Finally, the proposed IWLE method can be extended to multidimensional longitudinal IRT model.

From a practical point of view, we would not use a test that is way too difficult or way too easy items. This is because each item should have a certain discrimination to distinguish the examinees with different ability levels. In fact, the reliability and validity of the test items are pre-calibrated before the actual assessment. When the examinees answer the pre-calibrated test, some examinees answer all items correctly while others do not answer all items correctly. In this case, the extreme ability estimator will occur. Thus, the extreme ability occur because there are large differences between examinees' abilities rather than items being too difficult or too easy (the test items are pre-calibrated, reliable and valid). In addition, the examinees were obtained through a multistage stratified sample in the actual assessment. In the first stage, the sampling population is classified according to district, and schools are selected at random. In the second stage, students are selected at random from each school. Therefore, in this case, there are some extreme cases of the examinees' ability. For example, some examinees with high abilities answer all the items correctly, or some examinees with low abilities answered all the items incorrectly. Traditional methods (WLE and TWLE) fail to estimate these extreme abilities. However, our IWLE method is more accurate in estimating these extreme abilities. This is the main advantage of our item-weighted scheme.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 12001091), China Postdoctoral Science Foundations (Grant Nos. 2021M690587 and 2021T140108). In addition, this work was also supported by the Fundamental Research Funds for the Central Universities of China (Grant No. 2412020QD025).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.580015/full#supplementary-material

REFERENCES

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika* 50, 3–16. doi: 10.1007/bf02294143

Barrett, J., Diggle, P., Henderson, R., and Taylor-Robinson, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *J. R. Stat. Soc. Ser. B* 77, 131–148. doi: 10.1111/rssb.12060

Bradley, R. A., and Gart, J. J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika* 49, 205–214. doi: 10.2307/2333482

Chanda, K. C. (1954). A note on the consistency and maxima of the roots of likelihood equations. *Biometrika* 41, 56–61. doi: 10.2307/2333005

Chiang, C. L. (1956). On best regular asymptotically normal estimates. *Ann. Math. Statist.* 27, 336–351. doi: 10.1214/aoms/1177728262

Christine, D. (2010). *Item Response Theorem*. Oxford: Oxford University press.

Cramér, H. (1962). *Random Variables and Probability Distributions*. Cambridge, MA: Cambridge University Press.

Dearborne, D. F. (1921). Intelligence and its measurement: a symposium. *J. Educ. Psychol.* 12, 271–275.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *J. Educ. Meas.* 41, 295–311. doi: 10.1111/j.1745-3984.1994.tb00448.x

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika* 56, 495–515. doi: 10.1007/bf02294487

Embretson, S. E. (1997). "Multicomponent response models," in *Handbook of Modern Item Response Theory*, eds W. van der Linden and R. Hambleton (New York, NY: Springer-Verlag), 305–321. doi: 10.1007/978-1-4757-2691-6_18

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

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80, 27–38. doi: 10.1093/biomet/80.1.27

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychol.* 37, 359–374. doi: 10.1016/0001-6918(73)90003-6

Fischer, G. H. (1976). "Some probabilistic models for measuring change," in *Advances in Psychological and Educational Measurement*, eds D. N. M. de Gruijter and L. J. T. van der Kamp (New York, NY: Wiley), 97–110.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika* 60, 459–487. doi: 10.1007/bf02294324

Fischer, G. H. (2001). "Gain scores revisited under an IRT perspective," in *Essays on Item Response Theory*, eds A. Boomsma, M. A. J. Van Duijn, and T. A. B. Snijders (New York, NY: Springer), 43–68. doi: 10.1007/978-1-4613-0169-1_3

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *J. Educ. Meas.* 40, 1–15. doi: 10.1111/j.1745-3984.2003.tb01093.x

Lord, F. M. (1963). "Elementary models for measuring change," in *Problems in Measuring Change*, ed. C. W. Harris (Madison: University of Wisconsin press), 21–38.

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

Lord, F. M. (1983). Unbiased estimators of ability parameters of their variance, and of their parallel-forms reliability. *Psychometrika* 48, 233–245. doi: 10.1007/bf02294018

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika* 47, 149–174. doi: 10.1007/bf02296272

Muraki, E. (1993). Information functions of the generalized partial credit model. *Appl. Psychol. Meas.* 17, 351–363. doi: 10.1177/014662169301700403

Penfield, R. D., and Bergeron, J. M. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Appl. Psychol. Meas.* 29, 218–233. doi: 10.1177/0146621604270412

Rasch, G. (1980). *Probabilistic Model for Some Intelligence and Achievement Tests, Expanded Edition*. Chicago, IL: University of Chicago Press.

Sun, S.-S., Tao, J., Chang, H.-H., and Shi, N.-Z. (2012). Weighted maximum-a-posteriori estimation in tests composed of dichotomous and polytomous items. *Appl. Psychol. Meas.* 36, 271–290. doi: 10.1177/0146621612446937

Tao, J., Shi, N.-Z., and Chang, H.-H. (2012). Item-weighted likelihood method for ability estimation in tests composed of both dichotomous and polytomous items. *J. Educ. Behav. Stat.* 37, 298–315. doi: 10.3102/1076998610393969

von Davier, M., and Xu, X. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika* 76, 318–336. doi: 10.1007/s11336-011-9202-z

Wang, S., and Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Appl. Psychol. Meas.* 25, 317–331. doi: 10.1177/01466210122032163

Wang, T., Hanson, B. A., and Lau, C. A. (1999). Reducing bias in CAT ability estimation: a comparison of approaches. *Appl. Psychol. Meas.* 23, 263–278. doi: 10.1177/01466219922031383

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika* 54, 427–450. doi: 10.1007/bf02294627

Wilson, M. (1989). Saltus: a psychometric model for discontinuity in cognitive development. *Psychol. Bull.* 105, 276–289. doi: 10.1037/0033-2909.105.2.276

Woodrow, H. (1938). The relationship between abilities and improvement with practice. *J. Psychol.* 29, 215–230. doi: 10.1037/h0058249

Yao, L. (2009). Multidimensional linking for tests with mixed item types. *J. Educ. Mea.* 46, 197–197.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership