the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2020), 73, 51–81

www.wileyonlinelibrary.com

Marginalized maximum a posteriori estimation for the four-parameter logistic model under a mixture modelling framework

Xiangbin Meng[1]*, Gongjun Xu[2], Jiwei Zhang[3] and Jian Tao[1]

[1]School of Mathematics and Statistics, KLAS, Northeast Normal University, Changchun, Jilin, China
[2]Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA
[3]Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, School of Mathematics and Statistics, Yunnan University, Kunming, Yunnan, China

The four-parameter logistic model (4PLM) has recently attracted much interest in various applications. Motivated by recent studies that re-express the four-parameter model as a mixture model with two levels of latent variables, this paper develops a new expectation-maximization (EM) algorithm for marginalized maximum a posteriori estimation of the 4PLM parameters. The mixture modelling framework of the 4PLM not only makes the proposed EM algorithm easier to implement in practice, but also provides a natural connection with popular cognitive diagnosis models. Simulation studies were conducted to show the good performance of the proposed estimation method and to investigate the impact of the additional upper asymptote parameter on the estimation of other parameters. Moreover, a real data set was analysed using the 4PLM to show its improved performance over the three-parameter logistic model.

1. Introduction

The four-parameter logistic model (4PLM) was proposed by Barton and Lord (1981), who introduced an upper asymptote parameter, d, that is slightly < 1, to model the uncertainty of a high-ability examinee missing an easy item. The limitation of Barton and Lord's modelling approach is that all items in a test share a common upper asymptote parameter, and Barton and Lord did not estimate the fourth parameter but rather fitted the model with some fixed values for d. Recent studies (Linacre, 2004; Rouse, Finger, & Butcher, 1999; Rupp, 2003; Tavares, de Andrade, & Pereira, 2004; Waller & Reise, 2010) have demonstrated that, in most cases, the upper asymptote varies across items in a test. The following formulation of the 4PLM, which allows the upper asymptote parameter to be item-specific, is therefore considered more appropriate:

$$p_j(\theta_i) = P\left(U_{ij} = 1 \middle| \theta_i, \xi_j\right) = c_j + \left(d_j - c_j\right) \frac{e^{a_j\left(\theta_i - b_j\right)}}{1 + e^{a_j\left(\theta_i - b_j\right)}}, \tag{1}$$

where U_{ij} denotes the observed dichotomous response of examinee i ($i = 1, \ldots, N$) to item j ($j = 1 = 1, \ldots, M$), with $U_{ij} = 1$ denoting a correct response and $U_{ij} = 0$ otherwise;

*Correspondence should be addressed to Xiangbin Meng, Northeast Normal University, 5268 Renmin Street, Changchun 130024, China (email: mengxb600@nenu.edu.cn).

DOI:10.1111/bmsp.12185

$\theta_i \in (-\infty, +\infty)$ is the ability parameter; and $\xi_j = \{a_j, b_j, c_j, d_j\}$ is the item parameter set for the jth item, with $a_j \in (0, +\infty)$, $b_j \in (-\infty, +\infty)$, $c_j \in [0, 1]$, and $d_j \in (c_j, 1]$ being the discrimination, difficulty, guessing, and upper asymptote parameters, respectively. The parameter d_j is the maximum probability of endorsing item j, and so $1–d_j$ can be considered as the slipping probability of a student who can answer correctly but missing the item. Here, N and M are used to denote the number of the examinees (sample size) and the number of the items (test length).

Difficulties in parameter estimation and a lack of evidence supporting the need for it are the probable reasons why the 4PLM was not widely applied for a long time (Loken & Rulison, 2010). In recent years, however, researchers have shown renewed interest in the 4PLM. For instance, Liao, Ho, Yen, and Cheng (2012) and Rulison and Loken (2009) argued that the 4PLM can improve the accuracy of ability estimation by taking into account examinees' early careless errors in computerized adaptive testing. Reise and Waller (2003) and Waller and Reise (2010) demonstrated that the item response model with an upper asymptote parameter may be more appropriate for measuring psychopathology traits than the logistic model with three (3PLM) or two parameters (2PLM), since the situation of a high-trait subject who is reluctant to self-report attitudes is very common in psychopathology measurement. Ogasawara (2012) gave the asymptotic distribution of the ability estimate under the 4PLM, and Magis (2013) derived the maximum value of the information function. Furthermore, several methods for the estimation of the parameters in the four-parameter model have been proposed. For instance, Loken and Rulison (2010) employed a Bayesian approach with the Markov chain Monte Carlo (MCMC) sampler to estimate the 4PLM parameters. Feuerstahler and Waller (2014) employed the marginal maximum likelihood (MML) method to recover the 4PLM using the R package *mirt*. In comparison to the Bayesian estimation method calculated with the MCMC sampler algorithm, the MML method requires less computation time, but it may not be stable and may produce deviant values in many cases (Baker & Kim, 2004). To overcome this disadvantage of MML estimation, Mislevy (1986) proposed Bayes modal (BM) estimation for the 3PLM. This can be considered as a form of marginalized maximum a posteriori (MMAP) estimation; it employs an augmented optimization objective that includes the likelihood and some prior beliefs on the item parameters, and these priors were used to prevent deviant parameter estimates from occurring. In fact, BM estimation can be seen as a regulation of MML estimation, while MML estimation is a special case of BM estimation that assumes uniform prior distributions of parameters. Waller and Feuerstahler (2017) recently applied BM estimation as implemented in *mirt* for the 4PLM.

In addition to the above research on estimating the 4PLM, mixture modelling approaches have been developed by introducing latent variables to deal with the response process. For instance, Béguin and Glas (2001), San Martin, del Pino, and DeBoeck (2006), and von Davier (2009) interpreted the 3PLM from the perspective of a two-response (guessing and non-guessing) strategy, by revising the 3PLM as a mixture model. Recently, Culpepper (2016, 2017) further developed a mixture modelling approach to reformulate the four-parameter normal ogive model (4PNOM) and multidimensional 4PNOM. To estimate the model parameters, the existing works mostly focused on Bayesian estimation with an MCMC sampling procedure and may be computationally time-consuming, especially for large data sets. Motivated by the mixture modelling specification in these researches, this paper proposes a computationally efficient expectation–maximization (EM) algorithm to compute the MMAP estimates of the 4PLM parameters.

The rest of the paper is organized as follows. Section 2 reviews the mixture modelling reformulation of the 4PLM and discusses the relationship between the 4PLM and cognitive diagnosis model. Section 3 presents the derivation of the EM algorithm for MMAP estimation of the 4PLM under the mixture modelling framework. Section 4 reports three simulation studies conducted to evaluate the performance of the proposed method. Section 5 presents an application of the 4PLM to an empirical dataset. Finally, Section 6 provides further discussion on future research directions.

2. An alternative expression of the 4PLM from the two response processes: Guessing versus slipping

From equation (1), the probability of a correct response in the 4PLM is equivalent to

$$P(U_{ij} = 1|\theta_i, \xi_j) = c_j\Big(1 - p_j^*(\theta_i)\Big) + d_j p_j^*(\theta_i), \tag{2}$$

where

$$p_j^*(\theta_i) = \frac{\exp\big[a_j(\theta_i - b_j)\big]}{1 + \exp\big[a_j(\theta_i - b_j)\big]} \tag{3}$$

is the 2PLM.

Following the mixture framework for conceptualizing the process of ability-based responding and guessing behaviors for the 3PLM in von Davier (2009) and the study of the 4PNOM in Culpepper (2016), we present an alternative expression for the 4PLM using a mixture model. Specifically, we introduce an unobserved latent variable $W_{ij} \in \{0,1\}$ to characterize the two random response status of an examinee: $W = 1$ indicates that the examinee is 'capable' of answer the item based on his/her ability and $W = 0$ otherwise. Following the 4PLM representation in (2) and (3), we let W_{ij} follow a Bernoulli distribution,

$$W_{ij}|\theta_i, \xi_j \sim \text{Bernoulli}(p_j^*(\theta_i)), \tag{4}$$

where $p_j^*(\theta_i)$ is specified in (3), indicating that a higher ability θ_i leads to a higher chance of having $W_{ij} = 1$. When $W_{ij} = 1$, the conditional probability of the response U_{ij} is specified as

$$U_{ij}|W_{ij} = 1, \xi_j \sim \text{Bernoulli}(d_j), \tag{5}$$

where $1–d_j$ corresponds to the slipping probability of making a mistake even though the examinee is 'capable' of answering item j. On the other hand, when $W_{ij} = 0$ (i.e., the ith examinee does not know the correct answer to the jth item), the conditional distribution of U_{ij} is

$$U_{ij}|W_{ij} = 0, \xi_j \sim \text{Bernoulli}(c_j), \tag{6}$$

where c_j is the probability of guessing a correct response.

We next show that the mixture model specification in (4–6) is equivalent to the 4PLM given in (2). Based on the above distributions in (4–6), the joint probability distribution of U_{ij} and W_{ij} (conditionally on θ_i and ξ_j) can be given as

$$
\begin{aligned}
p_{(U_{ij},W_{ij})}(u_{ij},w_{ij}|\theta_i,\xi_j) &= p_{U_{ij}|W_{ij},\theta_i,\xi_j}(u_{ij}|w_{ij})p_{W_{ij}|\theta_i,\xi_j}(w_{ij}|\theta_i,\xi_j) \\
&= d_j^{w_{ij}u_{ij}}(1-d_j)^{w_{ij}(1-u_{ij})}c_j^{(1-w_{ij})u_{ij}}(1-c_j)^{(1-w_{ij})(1-u_{ij})} \\
&\quad \times p_j^*(\theta_i)^{w_{ij}}\left[1-p_j^*(\theta_i)\right]^{1-w_{ij}}.
\end{aligned} \tag{7}
$$

Hence, the marginal probability distribution of U_{ij} over W_{ij} can be given by

$$
\begin{aligned}
p_{U_{ij}}(u_{ij}|\theta_i,\xi_j) &= \sum_{w_{ij}=1,0} p_{(U_{ij},W_{ij})}(u_{ij},w_{ij}|\theta_i,\xi_j) \\
&= d_j^{u_{ij}}(1-d_j)^{(1-u_{ij})}p_j^*(\theta_i) + c_j^{u_{ij}}(1-c_j)^{(1-u_{ij})}(1-p_j^*(\theta_i)),
\end{aligned} \tag{8}
$$

which is a two-class mixture Bernoulli distribution. From equation (8), we have the marginal probability of $U_{ij} = 1$,

$$
p_{U_{ij}}(u_{ij}=1|\theta_i,\xi_j) = p_j^*(\theta_i)d_j + (1-p_j^*(\theta_i))c_j, \tag{9}
$$

which is the same as the 4PLM given in (2).

The above derivations demonstrate that the 4PLM can be considered as a two-strategy mixture model. What is more, the mixture model framework offers new insight into the 4PLM and naturally connects it with the cognitive diagnosis models (CDMs) as shown in Remark (1).

Remark 1. **(Connection to CDMs).** From the CDM literature, W_{ij} can also be interpreted as the ideal response variable, where $W_{ij} = 1$ indicates that the ith examinee is capable of answering item j and $W_{ij} = 0$ otherwise. Then the distribution of U_{ij} specified in (5) and (6) is the same as the deterministic input, noisy AND gate (DINA) model specification, where c_j corresponds to the guessing parameter and $1–d_j$ corresponds to the slipping parameter.

Moreover, we show that the 4PLM can also be viewed as a generalization of the higher-order DINA model (de la Torre & Douglas, 2004) with only one latent attribute. In particular, consider a cognitive diagnosis test with only one latent attribute $A \in \{0,1\}$. Then the Q-matrix is $J \times 1$ and we set $Q = (1, \ldots, 1)'_{J\times 1}$, that is, all items require the attribute A. Note that in this special case, the ideal responses of an examinee to all items are the same. Let A_i be the ith examinee's latent attribute and the common ideal responses to all items are $I(A_i = 1) = A_i$. The higher-order DINA model assumes that the probability of $A_i = 1$ is from a 2PLM given by

$$
P(A_i = 1|\theta_i,\lambda) = \frac{\exp[\lambda_0(\theta_i - \lambda_1)]}{1+\exp[\lambda_0(\theta_i - \lambda_1)]}, \tag{10}
$$

where θ_i denotes a latent variable representing general ability in the studied domain and the λ are regression parameters. Furthermore, given $I(A_i = 1) = A_i$, the ith examinee's response U_{ij} to the jth item follows the same models in (5) and (6) under the higher-order DINA model. Therefore, the only difference between the 4PLM and the one-attribute

higher-order DINA model lies in how they model the ideal responses (W_{ij} and A_i, respectively). Comparing the model set-up of the ideal responses between the higher-order DINA model in (10) and the 4PLM in (2), we can see that (10) can be considered as a special case of (2) with all the a_j replaced by a common parameter λ_0, the b_j replaced by λ_1, and W_{ij} replaced by a common variable A_i not depending on j. From this perspective, the the one-attribute higher-order DINA model can be viewed as a special case of the 4PLM. More generally, we may consider the multi-attribute higher-order DINA model as a sub-model of the multidimensional 4PLM.

3. MMAP estimation for the 4PLM with an EM algorithm

Under the mixture model framework, we develop an EM algorithm for MMAP estimation of the item parameters in the 4PLM. In the following, we first specify the prior distributions on the 4PLM parameters and then derive the EM algorithm formula to calculate the MMAP estimators of the 4PLM item parameters.

We first introduce some notation. Let $\mathbf{u}_{i\cdot} = (u_{i1}, \ldots, u_{iM})$ denote the observed response vector of examinee i, $\mathbf{u}_{\cdot j} = \left(u_{1j}, \ldots, u_n\right)'$ denote the observed response vector of item j, and $\mathbf{u} = (\mathbf{u}_{\cdot 1}, \ldots, \mathbf{u}_{\cdot M})$ denote the realized response matrix. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ be the ability parameter vector of all N examinees, $\xi_j = (a_j, b_j, c_j, d_j)$ be the item parameter vector of item j, and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_M)$ be all the item parameters of all M items.

The prior distribution for the ability variable θ_i, is specified to be normal, $\theta_i \sim N\left(\mu_\theta, \sigma_\theta^2\right)$. This is the standard choice in calculating the MML or MMAP estimates of the parameters in IRT models. For the discrimination parameter a_j, we first transform $a_j = e^{\alpha_j}$, then assign a normal prior for $\alpha_j, \alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$. The prior for b_j is a normal distribution, $b_j \sim N\left(\mu_b, \sigma_b^2\right)$. The prior for c_j is a beta prior, $c_j \sim Beta(s_c, t_c)$. These prior distributions are commonly used in applications of the IRT models. Finally, we assign a truncated Beta prior for $d_j, d_j | c_j \sim Beta(s_d, t_d) I(c_j < d_j)$, since $d_j > c_j$. Such a truncated prior is used in Culpepper (2016) to enforce the monotonicity condition. Here $\Omega := \{\mu_\alpha, \sigma_\alpha^2, \mu_b, \sigma_b^2, s_c, t_c, s_d, t_d\}$are hyperparameters to be prespecified in practice.

According to Bayes' theorem, the joint posterior density of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ is $p(\boldsymbol{\xi}, \boldsymbol{\theta} | \mathbf{u}, \Omega, \boldsymbol{\tau}) \propto L(\mathbf{u} | \boldsymbol{\xi}, \boldsymbol{\theta}) f(\boldsymbol{\theta} | \boldsymbol{\tau}) f(\boldsymbol{\xi} | \Omega)$, where

$$L(\mathbf{u} | \boldsymbol{\theta}, \boldsymbol{\xi}) = \prod_{i=1}^{N} \prod_{j=1}^{M} p_j(\theta_i)^{u_{ij}} \left(1 - p_j(\theta_i)\right)^{1 - u_{ij}},$$

is the likelihood of the observed response data $\mathbf{u}$, and

$$f(\boldsymbol{\theta} | \boldsymbol{\tau}) = \prod_{i=1}^{N} f(\theta_i | \tau), \quad f(\xi | \Omega) = \prod_{j=1}^{N} f(\xi_j | \Omega),$$

are the prior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, respectively.

As known in the literature (Baker & Kim, 2004; Neyman & Scott, 1948), direct joint estimation of person ability parameters θ_i and item parameters often leads to inconsistent estimators, therefore it is generally necessary to integrate over the θ_i in order to estimate the item parameters. Then we have the corresponding marginal distribution,

$$p(\boldsymbol{\xi}|\mathbf{u},\boldsymbol{\Omega},\boldsymbol{\tau}) = \int p(\boldsymbol{\xi},\boldsymbol{\theta}|\mathbf{u},\boldsymbol{\Omega},\boldsymbol{\tau})d\boldsymbol{\theta}, \tag{11}$$

and the modes of the marginal posterior $p(\boldsymbol{\xi}|\mathbf{u},\boldsymbol{\Omega},\boldsymbol{\tau})$,

$$\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi}\in\Theta_{\boldsymbol{\xi}}} p(\boldsymbol{\xi}|\mathbf{u},\boldsymbol{\Omega},\boldsymbol{\tau}), \tag{12}$$

are defined as the MMAP estimates of $\boldsymbol{\xi}$.

From equation (7), if the latent variables $\mathbf{W} = \{W_{ij}, i = 1,\ldots,N; j = 1,\ldots,M\}$ were observed, the 4PLM could be divided into two Bernoulli models, and the calculation of the estimators of $\boldsymbol{\xi}$ would be straightforward. Specifically, let $\mathbf{z} = (\mathbf{u},\mathbf{W},\boldsymbol{\theta})$ be the complete data. The likelihood of $\mathbf{z}$ is

$$\begin{aligned} L(\mathbf{z}|\boldsymbol{\xi}) = \prod_{i=1}^{N}\prod_{j=1}^{M} & d_j^{W_{ij}u_{ij}}(1-d_j)^{W_{ij}(1-u_{ij})}c_j^{(1-W_{ij})u_{ij}}(1-c_j)^{(1-W_{ij})(1-u_{ij})} \\ & \times p_j^*(\theta_i)^{W_{ij}}(1-p_j^*(\theta_i))^{1-W_{ij}}f(\theta_i|\boldsymbol{\tau}). \end{aligned} \tag{13}$$

The marginal posterior distribution $p(\boldsymbol{\xi}|\mathbf{u},\boldsymbol{\Omega},\boldsymbol{\tau})$ in (11) can be calculated by

$$p(\boldsymbol{\xi}|\mathbf{u},\boldsymbol{\Omega},\boldsymbol{\tau}) = \int\int p(\boldsymbol{\xi},\mathbf{z}|\mathbf{u},\boldsymbol{\Omega},\boldsymbol{\tau})d\mathbf{W}d\boldsymbol{\theta},$$

where

$$p(\boldsymbol{\xi},\mathbf{z}|\mathbf{u},\boldsymbol{\Omega},\boldsymbol{\tau}) \propto L(\mathbf{z}|\boldsymbol{\xi})f(\boldsymbol{\xi}|\boldsymbol{\Omega}). \tag{14}$$

With the $\mathbf{W}$ unobserved in practice, we propose an EM interaction procedure under the complete data ($\mathbf{z}$) for calculating the MMAP estimators of $\boldsymbol{\xi}$ in equation (12). Let $\boldsymbol{\xi}^{(t)}$ be the current values for $\boldsymbol{\xi}$ at the tth iteration. The EM algorithm consists of the following two steps:

E-step. Given $\boldsymbol{\xi}^{(t)}$ and $\mathbf{u}$, calculate the conditional distribution of the latent variables $\mathbf{W}$ and $\boldsymbol{\theta}$, denoted by $p(\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\boldsymbol{\xi}^{(t)})$, and then use $p(\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\boldsymbol{\xi}^{(t)})$ to calculate the corresponding expectation of $\ln p(\boldsymbol{\xi},\mathbf{z}|\mathbf{u},\boldsymbol{\Omega},\boldsymbol{\tau})$, that is,

$$Q\left(\boldsymbol{\xi},\boldsymbol{\xi}^{(t)}\right) = E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\boldsymbol{\xi}^{(t)}}\{\ln p(\mathbf{z},\boldsymbol{\xi}|\mathbf{u},\boldsymbol{\Omega},\boldsymbol{\tau})\}. \tag{15}$$

M-step. Update the parameter estimate $\boldsymbol{\xi}^{(t+1)}$ by maximizing $Q(\boldsymbol{\xi},\boldsymbol{\xi}^{(t)})$, that is,

$$\boldsymbol{\xi}^{(t+1)} = \arg\max Q\left(\boldsymbol{\xi},\boldsymbol{\xi}^{(t)}\right).$$

We next describe the details in the E- and M-steps. From equations (13) and (14),

$$\begin{aligned} \ln p(\boldsymbol{\xi},\mathbf{z}|\mathbf{u},\boldsymbol{\Omega},\boldsymbol{\tau}|) &= \ln L(\mathbf{z}|\boldsymbol{\xi}) + \sum_{j=1}^{M}\ln f(\xi_j|\boldsymbol{\Omega}) \\ &= L_1(\boldsymbol{c},\boldsymbol{d}) + L_2(\boldsymbol{\alpha},\boldsymbol{b}) + \sum_{i=1}^{N}\ln f(\theta_i|\boldsymbol{\tau}) + \sum_{j=1}^{M}\ln f(\xi_j|\boldsymbol{\Omega}), \end{aligned} \tag{16}$$

where

$$L_1(c,d) = \sum_{i=1}^{N}\sum_{j=1}^{M}\{W_{ij}u_{ij}\ln d_j + W_{ij}(1-u_{ij})\ln(1-d_j) + (1-W_{ij})u_{ij}\ln c_j$$
$$+(1-W_{ij})(1-u_{ij})\ln(1-c_j)\},$$
$$L_2(\alpha,b) = \sum_{i=1}^{N}\sum_{j=1}^{M} W_{ij}\ln p_j^*(\theta_i) + (1-W_{ij})\ln(1-p_j^*(\theta_i)).$$

From equation (16), we note that the estimators of (c_j, d_j) and (α_j, b_j) can be calculated separately with respect to $L_1(c,d)$ and $L_2(\alpha,b)$ in the E- and M-steps. Since $L_1(c,d)$ is a linear function of W_{ij}, the E-step is done by simply replacing W_{ij} with $E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\xi^{(t)}}(W_{ij})$. In the M-step, the estimators of c_j and d_j can then be calculated as

$$c_j^{(t+1)} = \frac{\sum_{i=1}^{N}\left(1-E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\xi^{(t)}}(W_{ij})\right)u_{ij}+s_c-1}{\sum_{i=1}^{N}\left(1-E_{\mathbf{W},\boldsymbol{\theta}\left|\mathbf{u},\xi^{(t)}\right.}(W_{ij})\right)+s_c+t_c-2},$$
$$d_j^{(t+1)} = d_j^*\mathbf{I}(d_j^* > c_j^{(t+1)}) + (c_j+\delta)\left[1-\mathbf{I}(d_j^* > c_j^{(t+1)})\right], \tag{17}$$

where

$$d_j^* = \frac{\sum_{i=1}^{N}\left(E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\xi^{(t)}}(W_{ij})\right)u_{ij} + s_d - 1}{\sum_{i=1}^{N}\left(E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\xi^{(t)}}(W_{ij})\right) + s_d + t_d - 2}, \tag{18}$$

and $\mathbf{I}(d_j^* > c_j^{(t+1)})$ is the indicative function of $d_j^* > c_j^{(t+1)}$. Note that to impose the restriction that $d_j > c_j$, $d_j^{(t+1)}$ is assigned to be $c_j^{(t+1)} + \delta$ for a small $\delta > 0$ when $d_j^* \le c_j^{(t+1)}$.

Based on equations (7) and (8), we have

$$E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\xi^t}\left[W_{ij}\right] = \int\left[\frac{d_j p_j^*(\theta_i)}{p_j(\theta_i)}\right]^{u_{ij}}\left[\frac{(1-d_j)p_j^*(\theta_i)}{1-p_j(\theta_i)}\right]^{1-u_{ij}} p\left(\theta_i\left|\mathbf{u}_i,\xi^{(t)}\right.\right)d\theta_i,$$

where $p_j^*(\cdot)$ is defined in (3). A quadrature approximation method is used to compute the integrals in the E-step. In particular, define a grid of K equally spaced points, x_k $(k = 1,\ldots,K)$, specified for θ, and the associated weights $A(x_k)$ are assigned by $f(x_k|\boldsymbol{\tau}) \times (x_{k+1} - x_k)$. The posterior probability of x_k can be given by

$$p\left(x_k\left|\mathbf{u}_i,\xi^{(t)}\right.\right) \cong \frac{\prod_{j=1}^{M} p_j^{(t)}(x_k)^{u_{ij}} q_j^{(t)}(x_k)^{1-u_{ij}} A(x_k)}{\sum_{k=1}^{K}\prod_{j=1}^{M} p_j^{(t)}(x_k)^{u_{ij}} q_j^{(t)}(x_k)^{1-u_{ij}} A(x_k)}, \tag{19}$$

where

$$p_j^{(t)}(x_k) = c_j^{(t)} - (d_j^{(t)} - c_j^{(t)})\frac{\exp(e^{\alpha_j^{(t)}}(x_k - b_j^{(k)}))}{1+\exp(e^{\alpha_j^{(t)}}(x_k - b_j^{(k)}))}$$

and $q_j^{(t)}(x_k) = 1 - p_j^{(t)}(x_k)$. Then $E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\xi^t}\left[W_{ij}\right]$ can be approximately calculated by

$$E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\xi^t}\left[W_{ij}\right] \cong \sum_{k=1}^{K}\left[\frac{d_j^{(t)}p_j^{*(t)}(x_k)}{p_j^{(t)}(x_k)}\right]^{u_{ij}}\left[\frac{(1-d_j^{(t)})p_j^{*(t)}(x_k)}{1-p_j^{(t)}(x_k)}\right]^{1-u_{ij}} p(x_k|\mathbf{u}_i,\xi_j^{(t)}),$$

where $i = 1, \ldots, N, j = 1, \ldots, M$. Finally, plugging these into the equations (17) and (18), the revised estimators, $c_j^{(t+1)}$ and $d_j^{(t+1)}$, can be approximately calculated.

In the M-step, the estimation equations for α_j and b_j can be approximated by

$$\frac{\partial E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\xi^t}(\ln p(\boldsymbol{\xi},\mathbf{z}|\mathbf{u},\Omega,\boldsymbol{\tau}))}{\partial\alpha_j} \cong \sum_{k=1}^{K}(x_k-b_j)(\hat{N}(x_k)-\hat{R}(x_k)p_j^*(x_k)) - \frac{\alpha_j-\mu_\alpha}{\sigma_\alpha} = 0, \quad (20)$$

$$\frac{\partial E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\xi^t}(\ln p(\boldsymbol{\xi},\mathbf{z}|\mathbf{u},\Omega,\boldsymbol{\tau}))}{\partial b_j} \cong -e^{e^{\alpha_j}}\sum_{k=1}^{K}(\hat{N}(x_k)-\hat{R}(x_k)p_j^*(x_k)) - \frac{b_j-\mu_b}{\sigma_b} = 0, \quad (21)$$

where

$$\hat{N}(x_k) = \sum_{i=1}^{N}\left[\frac{d_j^{(t)}p_j^{*(t)}(x_k)}{p_j^{(t)}(x_k)}\right]^{u_{ij}}\left[\frac{(1-d_j^{(t)})p_j^{*(t)}(x_k)}{1-p_j^{(t)}(x_k)}\right]^{1-u_{ij}} p(x_k|\mathbf{u}_i,\xi_j^{(t)}),$$
$$\hat{R}(x_k) = \sum_{i=1}^{N} p\left(x_k\middle|\mathbf{u}_i,\boldsymbol{\xi}^{(t)}\right),$$

and $p(x_k|\mathbf{u}_i,\boldsymbol{\xi}^{(t)})$ is calculated as in (19). A Newton–Raphson algorithm is used to solve the nonlinear equations (20) and (21); the detailed calculation procedure and the corresponding MATLAB code are presented in the Appendices A and B.

4. Monte Carlo simulation

This section reports three simulation studies in order to show the performance of the proposed MMAP estimation procedure. Specifically, the aim of the first simulation study is to investigate the influences of the prior distributions on the performance of the MMAP estimation. The second simulation was conducted to study the relationship between the d parameter and the properties of the MMAP estimation. The third simulation was performed to compare the performances of the proposed MMAP\EM method with the existing BM estimation procedure implemented in the R package *mirt* (Waller & Feuerstahler, 2017).

4.1. Simulation study 1

In this simulation, the test length was $M = 20$ and the true values of a_j, b_j and c_j $(j = 1, \ldots, M)$ were randomly drawn from a large-scale achievement test that was analysed in Wang, Chang, and Douglas (2013). Following a similar set-up to that of Loken and Rulison (2010), the parameters $d_j (j = 1, \ldots, M)$ were randomly generated from a truncated beta distribution, $d_j \sim Beta(8, 2)$, with the constraint $d_j > c_j$. The true values of these item parameters are shown in the leftmost four columns of Table 2. The examinees' ability variables, $\theta_i (i = 1, \ldots, N)$, were randomly drawn from the standard normal distribution, $\theta_i \sim N(0, 1)$. As the sample size is an important data characteristic

determining the properties of the item parameter estimation, we generated response data with three sample sizes of $N = 1{,}000$, 5,000, 10,000.

To investigate the influence of the prior distributions of the parameters a, b, c, and d, the MMAP estimation was implemented under three groups of priors (see Table 1). Specifically, among the three groups of priors, those in the first row (denoted by MMAP1) provide the strongest prior information. The distributions shown in the third row (denoted by MMAP3) are the weakest informative priors, where *Beta*(1,1) is the uniform distribution on [0, 1], and $N(0, 10^2)$ is a close to non-informative prior. That is, the MMAP estimators calculated under this group of priors can be considered as an approximation of the MML estimators. The prior distributions shown in the middle row (denoted by MMAP2) are weaker than MMAP1 but stronger than MMAP3.

To reduce the Monte Carlo error, 500 replications of the response data sets were randomly generated, and the MMAP estimates were calculated for each of the 500 data sets. The number of quadrature points in the MMAP estimation was set to 20, and both the convergence criteria for the EM algorithm and the Newton–Raphson iterations were specified to be 0.001. Finally, the root mean squared error (RMSE) and mean error (ME) were calculated across the 500 replications to evaluate the accuracy and bias of the MMAP estimators. The RMSE is defined as

$$\mathrm{RMSE}(\delta_j) = \sqrt{G^{-1}\sum_{g=1}^{G}\left(\hat{\delta}_{gj} - \delta_j\right)^2}, \tag{22}$$

and the ME is defined as

$$\mathrm{ME}(\delta_j) = G^{-1}\sum_{g=1}^{G}\left(\hat{\delta}_{gj} - \delta_j\right), \tag{23}$$

where δ_j is the item parameter (any one of α_j, b_j, c_j, d_j) of interest, $\hat{\delta}_{gj}$ denotes the estimate of δ_j in the gth repetition, and G is the number of replications ($G = 500$ in this study).

In this simulation, there were no deviant parameter estimates or unsuccessful iterations, even in the case of the weakly informative priors given in MMAP3. We consider that the proposed estimation method based on the mixture model interpretation is helpful for improving the convergence rate of the EM algorithm. Furthermore, the implementation of the EM procedure was generally fast. For instance, the average calculation time (on a PC with an Intel Core i5-8200 1.6 GHz processor and 8 GB RAM) did not exceed 0.8, 2.5 and 10.0 s under the three sample sizes $N = 1{,}000$, 5,000, 100,000, respectively. Tables 2–4 show the RMSE values obtained for the MMAP estimators with the three prior specifications (MMAP1, MMAP2 and MMAP3) across the three sample sizes. Based on these results, the following trends can be observed.

Table 1. Prior distributions of item parameters in the 4PLM

	Prior (α)	Prior (b)	Prior (c)	Prior (d)
MMAP 1	$(\mu_\alpha = 0, \sigma_\alpha^2 = 1^2)$	$(\mu_b = 0, \sigma_b^2 = 1^2)$	$(s_c = 5, t_c = 17)$	$(s_d = 17, t_d = 5)$
MMAP 2	$(\mu_\alpha = 0, \sigma_\alpha^2 = 5^2)$	$(\mu_b = 0, \sigma_b^2 = 5^2)$	$(s_c = 3, t_c = 9)$	$(s_d = 9, t_d = 3)$
MMAP 3	$(\mu_\alpha = 0, \sigma_\alpha^2 = 10^2)$	$(\mu_b = 0, \sigma_b^2 = 10^2)$	$(s_c = 1, t_c = 1)$	$(s_d = 1, t_d = 1)$

Table 2. RMSE values for the MMAP estimators of the 4PLM item parameters, sample size $N = 1{,}000$

	True values				MMAP1				MMAP2				MMAP3			
Item	*a*	*b*	*c*	*d*	*a*	*b*	*c*	*d*	*a*	*b*	*c*	*d*	*a*	*b*	*c*	*d*
1	0.92	−0.48	0.16	0.85	0.28	0.13	0.04	0.03	0.30	0.15	0.04	0.03	0.35	0.26	0.09	0.05
2	0.93	0.75	0.18	0.82	0.20	0.19	0.03	0.04	0.23	0.20	0.04	0.04	0.23	0.23	0.04	0.08
3	1.22	0.23	0.16	0.95	0.36	0.14	0.03	0.06	0.33	0.14	0.04	0.05	0.27	0.13	0.04	0.04
4	0.65	1.77	0.18	0.87	0.23	0.45	0.03	0.11	0.23	0.41	0.03	0.10	0.25	0.54	0.03	0.16
5	1.35	2.16	0.24	0.77	0.33	0.25	0.02	0.02	0.40	0.27	0.02	0.03	0.44	0.38	0.02	0.20
6	1.09	1.64	0.12	0.89	0.39	0.26	0.02	0.10	0.41	0.23	0.02	0.10	0.39	0.25	0.02	0.12
7	0.49	1.60	0.17	0.90	0.18	0.60	0.02	0.14	0.17	0.55	0.03	0.14	0.24	0.65	0.04	0.18
8	0.74	1.46	0.11	0.91	0.25	0.41	0.03	0.13	0.27	0.39	0.03	0.13	0.29	0.42	0.03	0.14
9	0.86	0.16	0.13	0.92	0.39	0.14	0.06	0.08	0.39	0.16	0.06	0.07	0.36	0.17	0.07	0.06
10	0.72	0.45	0.18	0.88	0.21	0.25	0.03	0.07	0.21	0.25	0.04	0.07	0.24	0.27	0.05	0.07
11	1.31	1.23	0.16	0.93	0.32	0.24	0.02	0.11	0.33	0.21	0.03	0.10	0.31	0.17	0.03	0.07
12	1.09	1.69	0.14	0.91	0.40	0.28	0.02	0.12	0.39	0.26	0.02	0.12	0.38	0.27	0.02	0.14
13	1.07	0.61	0.05	0.86	0.48	0.12	0.05	0.07	0.45	0.13	0.05	0.07	0.26	0.17	0.04	0.08
14	1.09	0.78	0.19	0.88	0.26	0.22	0.03	0.06	0.29	0.19	0.04	0.05	0.26	0.20	0.04	0.06
15	1.23	0.89	0.20	0.84	0.25	0.17	0.03	0.04	0.28	0.16	0.03	0.04	0.27	0.17	0.03	0.07
16	0.97	1.88	0.08	0.81	0.40	0.22	0.02	0.04	0.39	0.21	0.02	0.04	0.40	0.30	0.02	0.14
17	0.61	0.17	0.05	0.87	0.37	0.18	0.12	0.08	0.37	0.21	0.11	0.08	0.40	0.28	0.13	0.09
18	0.60	1.14	0.10	0.86	0.29	0.26	0.05	0.09	0.28	0.28	0.05	0.09	0.32	0.32	0.05	0.12
19	0.79	1.89	0.25	0.91	0.27	0.64	0.07	0.16	0.29	0.54	0.07	0.14	0.23	0.62	0.05	0.19
20	0.68	0.56	0.18	0.92	0.23	0.32	0.03	0.10	0.24	0.32	0.04	0.10	0.25	0.27	0.05	0.09

Table 3. RMSE values for the MMAP estimators of the 4PLM item parameters, sample size $N = 5{,}000$

	True values				MMAP1				MMAP2				MMAP3			
Item	*a*	*b*	*c*	*d*	*a*	*b*	*c*	*d*	*a*	*b*	*c*	*d*	*a*	*b*	*c*	*d*
1	0.92	−0.48	0.16	0.85	0.11	0.09	0.04	0.02	0.11	0.10	0.03	0.02	0.12	0.13	0.04	0.02
2	0.93	0.75	0.18	0.82	0.10	0.13	0.03	0.02	0.12	0.13	0.03	0.03	0.12	0.12	0.03	0.03
3	1.22	0.23	0.16	0.95	0.16	0.08	0.02	0.03	0.13	0.08	0.02	0.02	0.12	0.07	0.02	0.02
4	0.65	1.77	0.18	0.87	0.08	0.42	0.03	0.14	0.09	0.47	0.03	0.14	0.08	0.52	0.01	0.16
5	1.35	2.16	0.24	0.77	0.30	0.12	0.01	0.02	0.32	0.12	0.01	0.04	0.29	0.19	0.01	0.14
6	1.09	1.64	0.12	0.89	0.23	0.18	0.01	0.11	0.23	0.19	0.01	0.11	0.23	0.21	0.01	0.12
7	0.49	1.60	0.17	0.90	0.07	0.62	0.02	0.15	0.07	0.62	0.02	0.15	0.09	0.67	0.02	0.17
8	0.74	1.46	0.11	0.91	0.10	0.37	0.02	0.13	0.09	0.35	0.02	0.13	0.10	0.37	0.02	0.13
9	0.86	0.16	0.13	0.92	0.18	0.09	0.04	0.04	0.14	0.08	0.03	0.04	0.13	0.09	0.03	0.03
10	0.72	0.45	0.18	0.88	0.08	0.18	0.02	0.05	0.08	0.17	0.02	0.04	0.08	0.17	0.02	0.04
11	1.31	1.23	0.16	0.93	0.23	0.14	0.01	0.07	0.22	0.11	0.01	0.06	0.19	0.09	0.01	0.05
12	1.09	1.69	0.14	0.91	0.20	0.22	0.01	0.12	0.20	0.22	0.01	0.12	0.21	0.24	0.01	0.14
13	1.07	0.61	0.05	0.86	0.23	0.08	0.03	0.04	0.22	0.08	0.03	0.04	0.19	0.08	0.02	0.04
14	1.09	0.78	0.19	0.88	0.12	0.13	0.02	0.03	0.13	0.12	0.02	0.03	0.15	0.11	0.03	0.03
15	1.23	0.89	0.20	0.84	0.18	0.09	0.02	0.02	0.19	0.09	0.03	0.03	0.20	0.09	0.03	0.04
16	0.97	1.88	0.08	0.81	0.18	0.16	0.01	0.07	0.19	0.18	0.01	0.09	0.22	0.27	0.01	0.14
17	0.61	0.17	0.05	0.87	0.24	0.13	0.09	0.06	0.22	0.13	0.09	0.06	0.22	0.15	0.09	0.06
18	0.60	1.14	0.10	0.86	0.16	0.24	0.03	0.10	0.15	0.25	0.03	0.10	0.15	0.28	0.03	0.11
19	0.79	1.89	0.25	0.91	0.26	0.60	0.07	0.17	0.27	0.57	0.07	0.17	0.25	0.62	0.06	0.19
20	0.68	0.56	0.18	0.92	0.11	0.24	0.02	0.07	0.10	0.24	0.02	0.07	0.10	0.22	0.02	0.07

Table 4. RMSE values for the MMAP estimators of the 4PLM item parameters, sample size $N = 10{,}000$

	True values				MMAP1				MMAP2				MMAP3			
Item	*a*	*b*	*c*	*d*	*a*	*b*	*c*	*d*	*a*	*b*	*c*	*d*	*a*	*b*	*c*	*d*
1	0.92	−0.48	0.16	0.85	0.08	0.08	0.03	0.02	0.08	0.09	0.03	0.02	0.08	0.10	0.04	0.02
2	0.93	0.75	0.18	0.82	0.10	0.09	0.02	0.02	0.11	0.09	0.03	0.02	0.11	0.09	0.03	0.02
3	1.22	0.23	0.16	0.95	0.08	0.06	0.02	0.02	0.08	0.06	0.02	0.01	0.09	0.05	0.02	0.01
4	0.65	1.77	0.18	0.87	0.07	0.43	0.03	0.13	0.07	0.49	0.03	0.14	0.06	0.52	0.03	0.15
5	1.35	2.16	0.24	0.77	0.30	0.09	0.01	0.02	0.30	0.10	0.01	0.04	0.27	0.13	0.01	0.10
6	1.09	1.64	0.12	0.89	0.16	0.18	0.01	0.10	0.16	0.19	0.01	0.10	0.17	0.19	0.01	0.10
7	0.49	1.60	0.17	0.90	0.06	0.65	0.02	0.15	0.05	0.63	0.02	0.16	0.06	0.64	0.02	0.16
8	0.74	1.46	0.11	0.91	0.06	0.35	0.02	0.12	0.06	0.34	0.02	0.13	0.07	0.35	0.02	0.12
9	0.86	0.16	0.13	0.92	0.12	0.06	0.03	0.03	0.09	0.07	0.03	0.03	0.09	0.07	0.03	0.02
10	0.72	0.45	0.18	0.88	0.06	0.15	0.02	0.04	0.05	0.16	0.02	0.03	0.06	0.13	0.02	0.03
11	1.31	1.23	0.16	0.93	0.16	0.10	0.01	0.05	0.15	0.08	0.01	0.04	0.15	0.07	0.01	0.03
12	1.09	1.69	0.14	0.91	0.14	0.22	0.01	0.12	0.14	0.23	0.01	0.12	0.17	0.23	0.01	0.12
13	1.07	0.61	0.05	0.86	0.17	0.05	0.02	0.03	0.15	0.06	0.02	0.03	0.12	0.06	0.02	0.03
14	1.09	0.78	0.19	0.88	0.10	0.08	0.02	0.02	0.11	0.07	0.02	0.02	0.12	0.07	0.02	0.02
15	1.23	0.89	0.20	0.84	0.17	0.06	0.02	0.02	0.19	0.06	0.02	0.03	0.20	0.06	0.02	0.03
16	0.97	1.88	0.08	0.81	0.14	0.17	0.01	0.08	0.15	0.20	0.01	0.11	0.19	0.25	0.01	0.13
17	0.61	0.17	0.05	0.87	0.20	0.11	0.08	0.06	0.19	0.10	0.08	0.05	0.19	0.12	0.08	0.05
18	0.60	1.14	0.10	0.86	0.13	0.23	0.02	0.09	0.12	0.24	0.02	0.09	0.13	0.24	0.02	0.09
19	0.79	1.89	0.25	0.91	0.27	0.60	0.07	0.17	0.26	0.59	0.07	0.17	0.24	0.62	0.07	0.18
20	0.68	0.56	0.18	0.92	0.08	0.20	0.01	0.06	0.08	0.20	0.02	0.06	0.07	0.18	0.02	0.05

For a sample size of $N = 1{,}000$, there are slight differences in the RMSE values of the MMAP estimators under the three groups of priors (MMAP1, MMAP2, MMAP3). Overall, the MMAP3 estimators displayed larger RMSE values than the MMAP1 and MMAP2 estimators. However, as the sample size increased, the differences in the RMSEs of the three estimators become much smaller. For instance, under sample sizes of $N = 5{,}000$ and 10,000, the differences in RMSEs of the three MMAP estimators were negligible for most item parameters. The same phenomenon was observed for the ME values (not reported here due to space limitations). This suggests that when the number of examinees is large, the MMAP estimators are mainly determined by the response data and the specification of the prior distributions is not less crucial. On the other hand, when the sample size is small, the prior information will have a larger impact on the performance of the MMAP estimation, so in order to avoid the subjective error from the misspecification of prior distributions, weakly informative or non-informative priors may be recommended in practice. Additionally, we calculated the BM estimates of the 4PLM using the *mirt* package. The results showed that BM estimators with informative priors perform similarly to our method, while BM estimators with non-informative priors not only displayed lower accuracy but also suffered frequently from unsuccessful convergence. It can be considered that the mixture strategies framework of the 4PLM is helpful for the convergence of the EM algorithm. The BM estimation results are not reported here as they are not the main focus of this simulation study, and more comparisons between our method and BM estimation are provided in simulation study 3.

It can be observed that the RMSE(d) values of items $j = \{4, 7, 8, 12, 19\}$ are much larger than those of the other items. The common characteristics of these items are that their a parameters were much lower than those of the other items, and their b and d parameters were relatively larger. This phenomenon was also observed in Culpepper (2016). Inspired by the research of Lord (1975) and Mislevy (1986), which verified under the 3PLM that the estimation accuracy of c_j and $b_j - 2/a_j$ are positively correlated, we may explain this phenomenon by a negative correlation between the estimation accuracy of d_j and the value of $b_j + 2/a_j$ under the 4PLM. Heuristically, a larger value of $b_j + 2/a_j$ implies fewer examinees satisfying $a_j(\theta_i - b_j) > 2$, and therefore less information on d_j is provided by the responses, which then reduces the estimation accuracy of d_j. Scatter plots with Pearson correlation coefficients were created to display the influence of $b_j + 2/a_j$ on the estimation errors and biases of the MMAP estimators of d (see Figure 1). It can be seen that across the three sample sizes, both the RMSE(d) and absolute ME(d) were positively correlated with $b_j + 2/a_j$, and these correlations increase with the sample size. These results demonstrated that the higher the difficulty and the lower the discrimination, the poorer the estimation accuracy for the d parameter in terms of both root mean squared error and bias.

4.2. Simulation study 2

The main purpose of this simulation is to investigate the impact of the d parameter on the performance of the MMAP estimation. An artificial test with four levels of d, $d \in \{.65, .75, .85, .95\}$, was conducted, where each level of d included five items and the test length was $M = 20$. To produce a controlled experiment, the values of a, b and c were identical for all items, with $a = 1$, $b = 0$, and $c = .2$. Following simulation study 1, the sample sizes were set to $N = \{1{,}000, 5{,}000, 10{,}000\}$, and the examinees' ability parameters θ were randomly drawn from $N(0,1)$. Additionally, 500 response data sets were randomly generated, and the MMAP estimates were calculated with the three groups of priors in Table 1. Finally, the

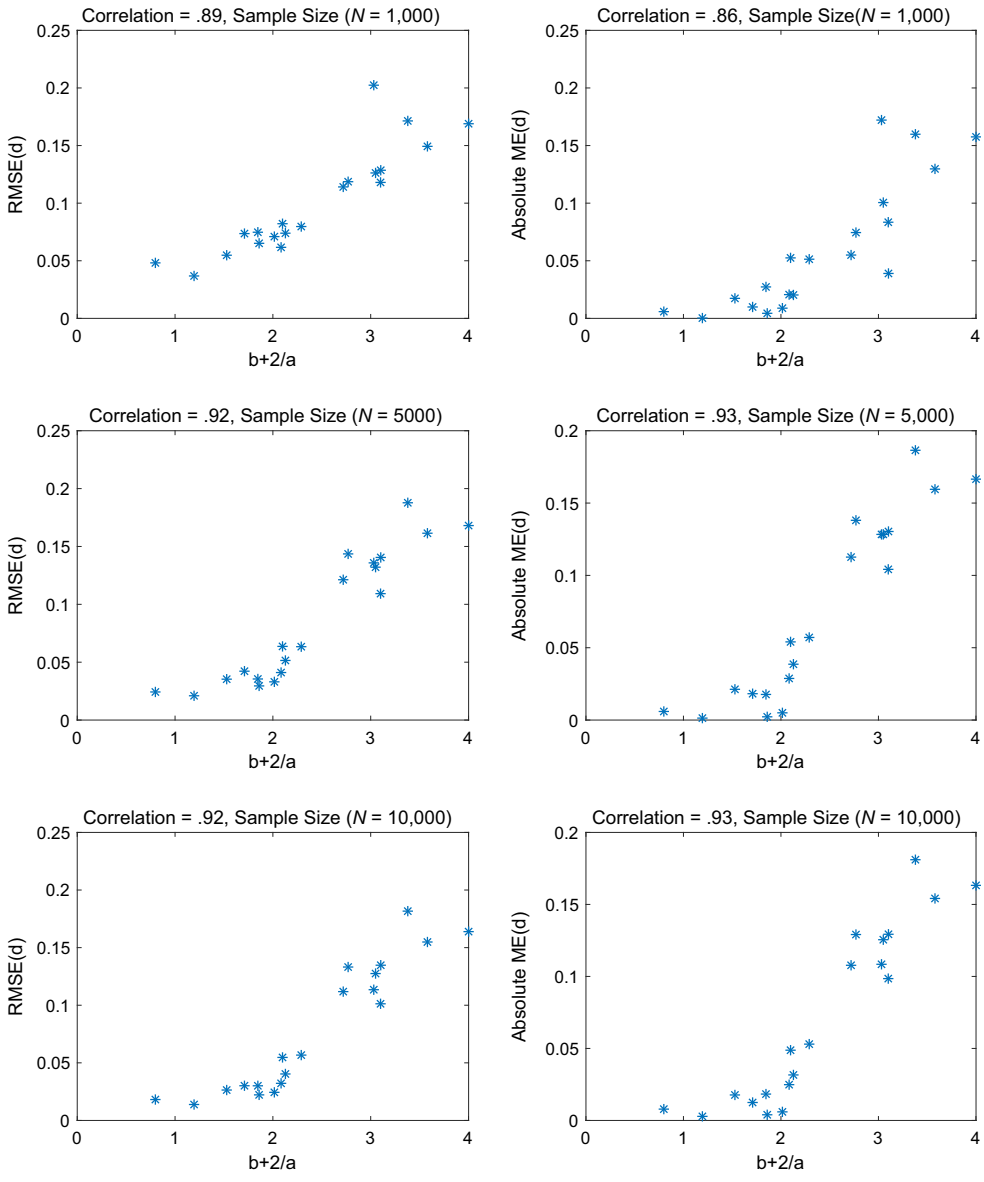

Figure 1. Scatter plots of (left) the RMSE and (right) the absolute ME of the MMAP estimators for the d parameter against $b + 2/a$ for sample sizes $N = \{1{,}000, 5{,}000, 10{,}000\}$.

RMSE and ME of the MMAP estimates were calculated to display the properties (efficiency and bias) of the estimator. Because the trends on the MMAP estimators with the three groups of priors were consistent, we only report the results under the priors of MMAP1 here.

Figures 2 and 3 show the RMSE and ME values for the MMAP estimators of a, b, c and d at the four different levels of d. For the a and b parameters, it can be seen that the values of RMSE(a) and RMSE(b) at $d = \{.75, .85\}$ were smaller than at $d = \{.65, .95\}$. Similarly, the values of ME(a) were closer to 0 (smaller biases) for $d = \{.75, .85\}$ than for $d = \{.65, .95\}$.

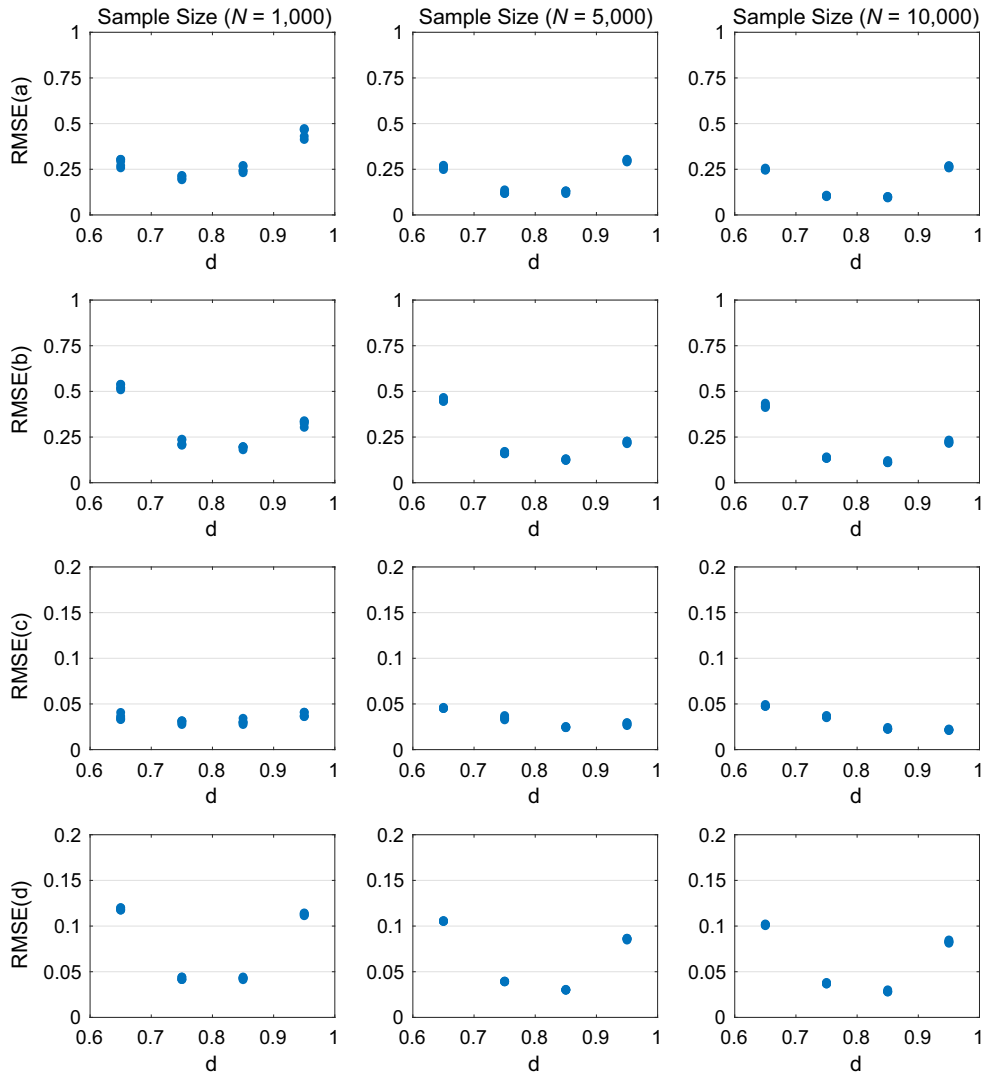

Figure 2. RMSE values of the MMAP estimators for the 4PLM item parameters for $d = \{.65; .75; .85; .95\}$ and sample sizes $N = \{1{,}000; 5{,}000; 10{,}000\}$.

This indicates that a and b are more difficult to estimate when d takes more extreme values.

For the c parameters. it can be seen that the relationships between d and RMSE(c) were the weakest among the four types of item parameters, and the highest values were not larger than .05. The values of ME(c) were very close to 0. These results demonstrate that the d parameter has the smallest impact on the MMAP estimator of c.

For the d parameters, RMSE(d) displays substantial differences under the four levels of d: for the two middle levels of d, $d = \{.75, .85\}$, RMSE(d) was smaller than for $d = \{.65, .95\}$ and had smaller biases. This suggests that the estimators of the middle d values are more accurate than those of the extreme d values.

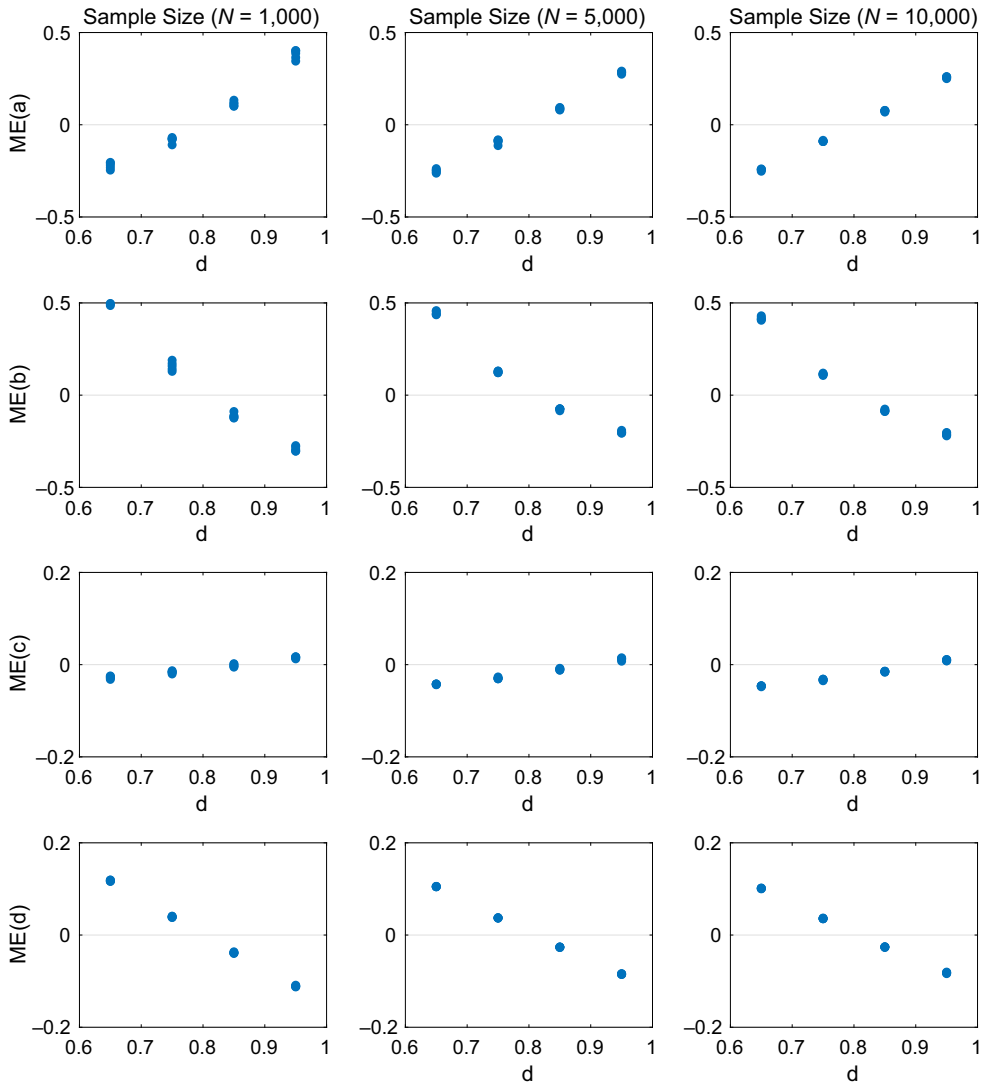

Figure 3. ME values of the MMAP estimators for the 4PLM item parameters for d = {.65; .75; .85; .95} and sample sizes N = {1,000; 5,000; 10,000}.

4.3. Simulation study 3

Many researchers have studied the application of the 4PLM to psychopathology testing (Culpepper, 2016; Reise & Waller, 2003; Waller & Reise, 2010), where subjects with higher levels of psychopathology may be reluctant to self-report attitudes, behaviours, and/or experiences. Therefore, in this simulation, we compared the performance of the proposed MMAP estimation with that of BM estimation in estimating the 4PLM with a set of psychopathology items. Following Culpepper (2016) and Waller and Feuerstahler (2017), this study generated responses based on the 4PLM with the M = 23 psychopathology item parameters from Waller and Reise (2010) as the true values (see Table 5). As in simulation studies 1 and 2, the examinees' abilities (θ) were randomly drawn from $N(0,1)$, and three sample sizes N = {1,000, 5,000, 10,000} were considered.

The MMAP estimates were calculated with the informative prior distributions that were given for MMAP1 in Table 1. In the *mirt* R library, the logistic model was design by a slope-threshold parameterizations, that is, $1.7a_i$ and $1.7a_ib_i$ were estimated instead of directly estimating a_i and b_i. According to Waller and Feuerstahler (2017), the priors for $1.7a$ and $1.7ab$ were set to $1.7a \sim LN(1,1^2)$ and $1.7ab \sim N(0,2^2)$. In addition, the prior distributions for c and d were set to logistic (c) $\sim N(-1.2, 0.5^2)$ and logistic(d) $\sim N(1.2, 0.5^2)$, which are approximately equal to $Beta(5,17)$ and $Beta(17,5)$ (see Figure 4). To sum up, the prior distributions for the two estimation methods were very close. The MMAP and BM estimations of the 4PLM were calculated across 500 replications, and the RMSE were calculated to evaluate the properties of the estimators (see Figures 5–7).

From these plots, it can be observed that, for most of the 23 items, the MMAP estimators of the item parameters (*a, b, c, d*) provided lower RMSE values than did the BM estimators across the three sample sizes. It is ecident that the accuracy of the MMAP estimators was superior to that of the BM estimators. It is obvious that the RMSEs of the MMAP and BM estimators both display decreasing trends as the sample size increases. That is, increasing the sample size can improve the estimation accuracy, which is expected. Finally, there are still differences between the RMSEs of the MMAP and BM estimators under a sample size of $N = 10{,}000$, but the superiority of the MMAP estimator is weaker, especially for the *b* and *c* parameters, and the two estimators were extremely close.

5. Empirical study

This section demonstrates an application of the 4PLM with an empirical example. The data set is from a state reading assessment test that was previously analysed in Tao, Shi, and Chang (2012). The data set includes 50 dichotomous items and the sample size is = 2,000. In our study, the 4PLM was fitted to the response data of the 50 dichotomous items. The item parameters were estimated using the MMAP method, and the examinees' abilities were estimated using Warm's weighted maximum likelihood estimation (WMLE). Warm's WMLE has been proved to be superior to the ML and expected a posteriori estimates by many studies (Meng, Tao, & Chen, 2016; Peneld & Bergeron, 2005; Wang & Wang, 2001;

Table 5. Item parameter values for the psychopathology item in Waller and Reise (2010)

	Item parameters					Item parameters			
Item	*a*	*b*	*c*	*d*	Item	*a*	*b*	*c*	*d*
1	1.91	−0.28	0.04	0.52	14	0.84	0.72	0.04	0.75
2	1.95	−0.16	0.02	0.48	15	1.13	0.15	0.03	0.61
3	1.50	0.05	0.02	0.60	16	0.79	1.19	0.04	0.73
4	1.12	0.06	0.02	0.63	17	1.27	0.48	0.01	0.84
5	0.89	0.45	0.04	0.82	18	0.94	1.37	0.09	0.94
6	1.08	−0.50	0.06	0.83	19	0.84	1.44	0.02	0.82
7	1.16	−0.47	0.07	0.71	20	1.14	1.52	0.00	0.82
8	1.10	0.01	0.04	0.73	21	1.10	0.25	0.02	0.93
9	0.78	0.45	0.05	0.57	22	0.72	0.53	0.24	0.95
10	1.23	0.19	0.01	0.90	23	0.88	1.56	0.06	0.91
11	1.34	0.41	0.02	0.85					
12	1.54	−0.48	0.06	0.59					
13	1.16	0.18	0.02	0.40					

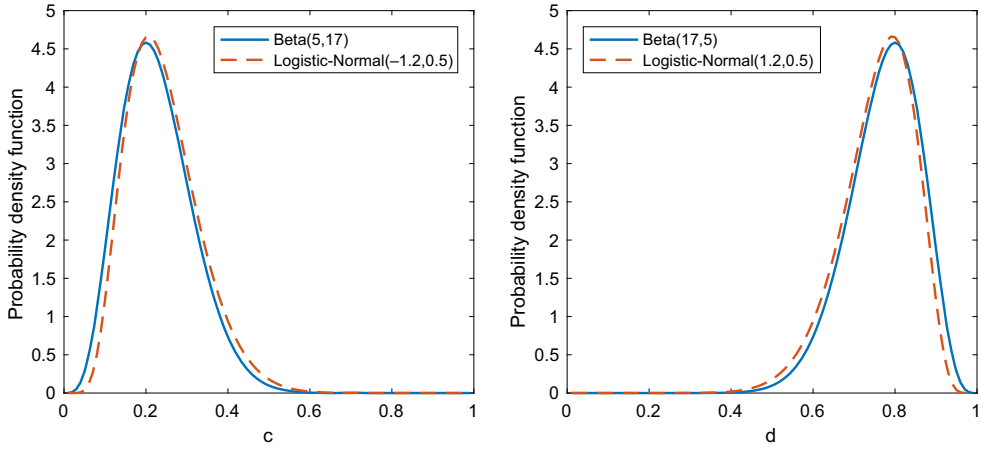

Figure 4. Probability density function curves for the Beta(5, 17), $LN(-1.2, 0.5^2)$, Beta(17, 5), and $LN(1.2, 0.5^2)$ distributions.

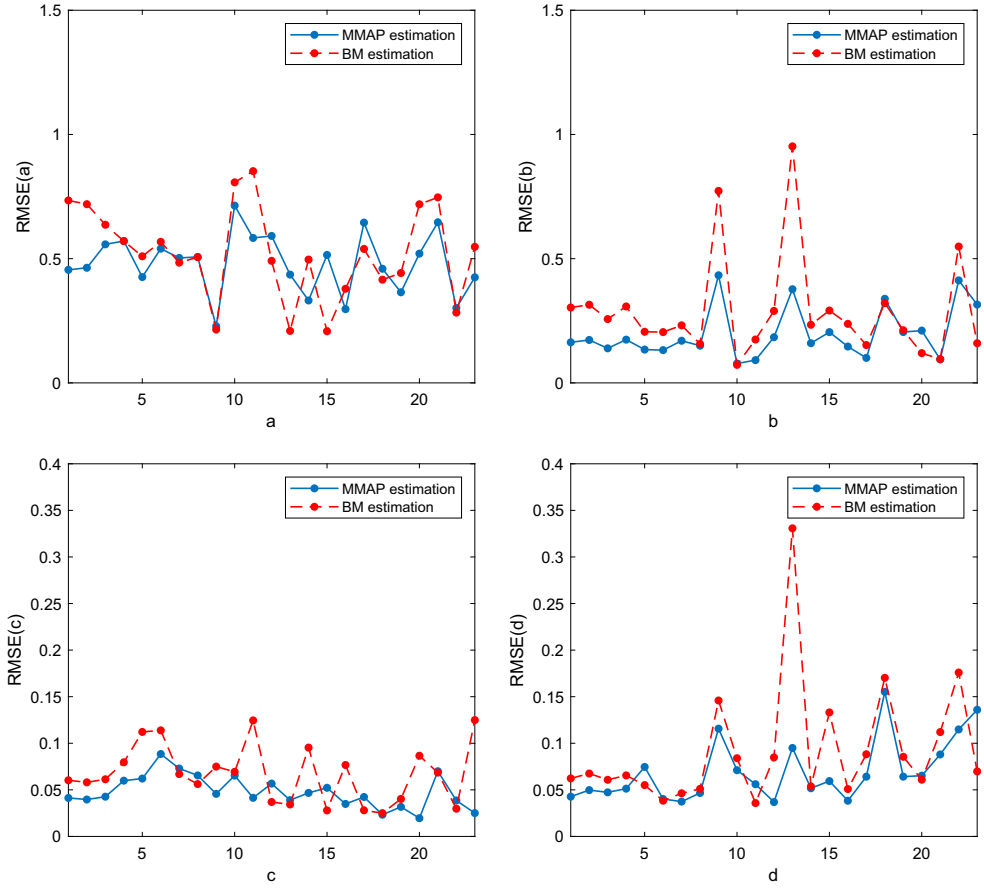

Figure 5. RMSE values for the MMAP and BM estimators of the 4PLM item parameters for sample size $N = 1{,}000$.

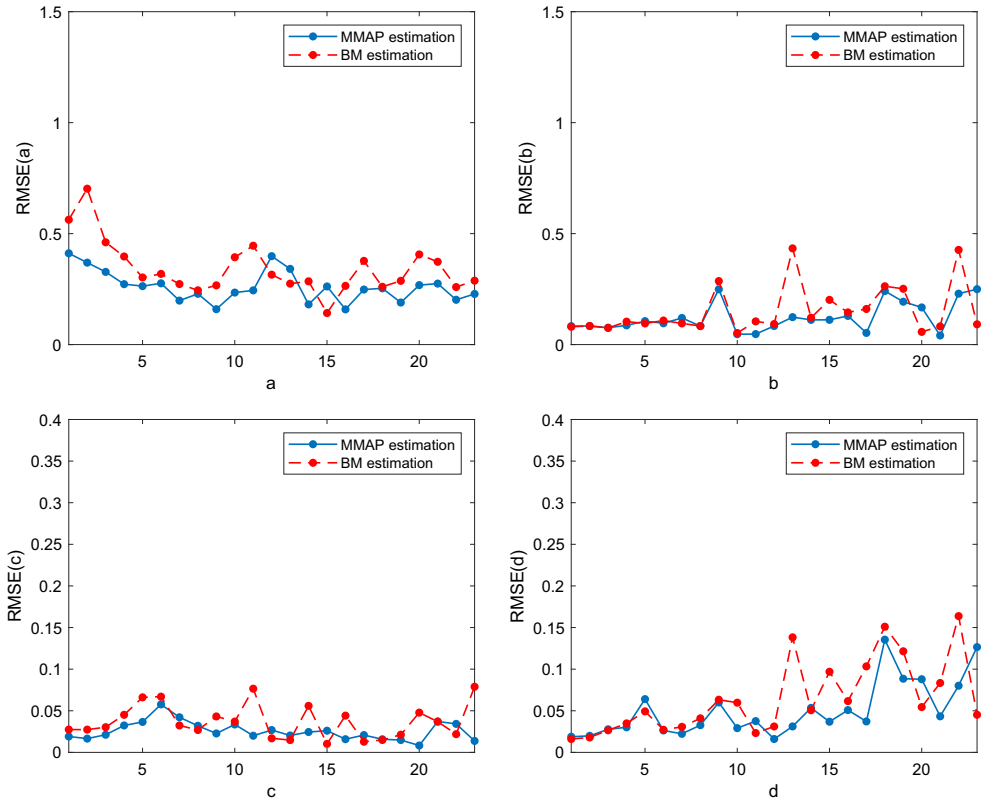

Figure 6. RMSE values for the MMAP and BM estimators of the 4PLM item parameters for sample size $N = 5{,}000$.

Warm, 1989). In what follows, the item parameter estimation results and model fitting evaluation are reported.

5.1. Item parameter estimation results

The item parameter estimates from the 3PLM and 4PLM are presented in Table 6. It can be observed that the estimates of the parameters (*a, b, c*) in the two models (3PLM and 4PLM) are close for most items, while for the items with lower values of *d*, the differences between the estimates are more substantial. For instance, for items $j = 5, 9, 18, 50$, the *a* parameters estimated from the 3PLM are extremely small, while the estimates from the 4PLM are much larger. This may be because a large proportion of examinees slipped in their responses to these items, resulting in the 3PLM underestimating their discrimination (see also the model fitting evaluation results given in Table 6 to be discussed in the next subsection).

The Pearson correlation coefficients between the parameter estimates of the 3PLM and 4PLM are $r_{a^{(3PL)},a^{(4PL)}} = .68$, $r_{b^{(3PL)},b^{(4PL)}} = .94$, and $r_{c^{(3PL)},c^{(4PL)}} = .88$, and the corresponding scatter plots are shown in the left-hand column of Figure 8. We also illustrate the differences of the distributions of *a*, *b*, and *c* between the 3PLM and 4PLM by estimating their kernel density curves across the test (see the right-hand

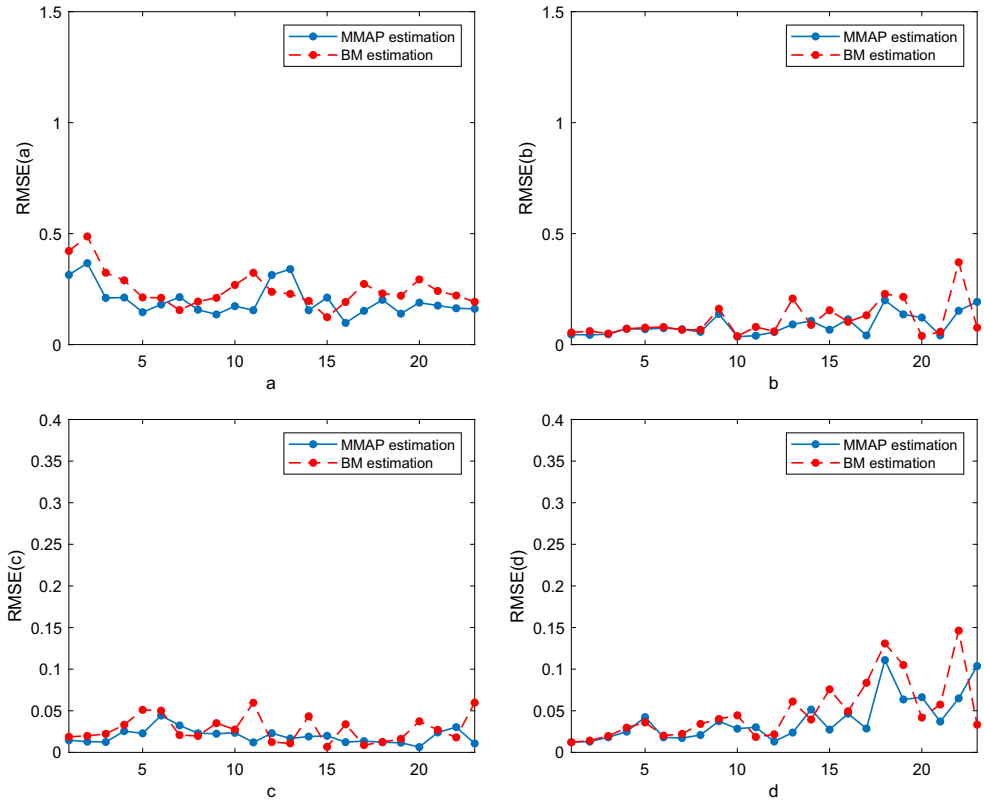

Figure 7. RMSE values for the MMAP and BM estimators of the 4PLM item parameters for sample size N = 10,000.

column in Figure 8). The estimates of *a*, *b* and *c* in the 4PLM are highly correlated with those in the 3PLM. Furthermore, it can be observed that the *a* parameter of the 4PLM was consistently higher than that of the 3PLM for each item, but the *b* parameter presented the opposite trend. This phenomenon has also been found in Loken and Rulison (2010). The reason for this may be that an upper asymptote < 1 results in the response function not having to flatten out to accommodate the poorly fitting responses (Loken & Rulison, 2010).

Finally, we compare the performances of the 4PLM and the 3PLM in estimating the examinees' abilities θ. The scatter plot between the estimates of θ from the 3PLM and 4PLM and their kernel probability density function curves are presented in Figure 9. It can be seen that the estimates of θ from the two models are highly correlated with their Pearson correlation, $r_{\theta^{(3PL)},\theta^{(4PL)}} = .98$. However, when θ > 1.0, the estimates of θ from the 4PLM are a little larger than those from the 3PLM. This indicates that the 3PLM is likely to underestimate the high-ability examinees. Furthermore, from the kernel density curves, it can be observed that the θ curves mostly overlap, except for the right tail, where the 3PLM may fail to capture the behaviours of the high-ability students. It would be interesting to further investigate whether the result obtained in the empirical study still holds in general and how it would impact test-taking strategies if 4PLM were known to be the scoring model beforehand. We leave this interesting topic for future study.

Table 6. Item parameter estimates of the 4PLM and 3PLM for the empirical data

	4PLM				3PLM				4PLM				3PLM		
Item	*a*	*b*	*c*	*d*	*a*	*b*	*c*	Item	*a*	*b*	*c*	*d*	*a*	*b*	*c*
1	0.95	−0.96	0.07	0.97	0.88	−0.98	0.05	26	1.58	0.07	0.07	0.96	1.39	0.17	0.05
2	1.20	−0.88	0.07	0.91	0.75	−0.81	0.04	27	0.99	−0.23	0.15	0.95	0.84	−0.16	0.13
3	0.91	−0.38	0.08	0.93	0.74	−0.28	0.05	28	1.15	−0.63	0.09	0.94	0.84	−0.58	0.05
4	1.01	−0.79	0.07	0.96	0.83	−0.76	0.05	29	1.31	0.47	0.23	0.93	1.21	0.64	0.23
5	1.79	−1.31	0.08	0.84	0.47	−1.49	0.05	30	1.25	0.12	0.15	0.94	1.16	0.28	0.15
6	1.50	−1.29	0.07	0.98	1.17	−1.38	0.04	31	1.41	−0.31	0.16	0.94	1.06	−0.23	0.12
7	1.97	−1.04	0.08	0.98	1.43	−1.12	0.04	32	0.68	0.04	0.11	0.87	0.55	0.31	0.08
8	0.71	−0.89	0.11	0.92	0.53	−0.81	0.06	33	1.01	−0.96	0.13	0.98	0.85	−1.05	0.07
9	0.87	0.31	0.13	0.76	0.57	0.83	0.08	34	1.08	−0.61	0.06	0.91	0.76	−0.45	0.04
10	1.13	−0.28	0.10	0.93	0.88	−0.19	0.06	35	1.58	−0.26	0.20	0.92	1.00	−0.19	0.13
11	1.44	−0.66	0.11	0.97	1.11	−0.67	0.06	36	1.54	−0.35	0.14	0.87	0.79	−0.23	0.05
12	1.09	0.14	0.08	0.88	0.80	0.36	0.06	37	1.25	−0.39	0.11	0.96	1.02	−0.36	0.07
13	1.91	−0.43	0.19	0.97	1.45	−0.41	0.16	38	0.97	−0.29	0.11	0.92	0.73	−0.19	0.07
14	1.28	−0.78	0.07	0.90	0.77	−0.68	0.03	39	0.70	0.52	0.11	0.85	0.61	0.84	0.08
15	0.97	−0.84	0.10	0.97	0.81	−0.86	0.06	40	1.07	0.14	0.13	0.94	0.99	0.29	0.13
16	1.27	−0.88	0.08	0.94	0.87	−0.87	0.04	41	1.31	0.26	0.09	0.93	1.17	0.39	0.08
17	0.85	−0.11	0.07	0.83	0.62	0.24	0.04	42	1.33	−0.89	0.09	0.96	0.96	−0.92	0.04
18	1.17	−0.96	0.09	0.70	0.37	−0.24	0.05	43	0.79	−0.06	0.07	0.87	0.63	0.19	0.04
19	1.54	−0.32	0.12	0.96	1.22	−0.25	0.05	44	1.64	−0.45	0.14	0.97	1.32	−0.39	0.11
20	1.31	−0.17	0.13	0.94	1.11	−0.02	0.09	45	1.23	−0.58	0.13	0.93	0.81	−0.57	0.05
21	0.84	0.09	0.15	0.94	0.81	0.28	0.12	46	1.82	−0.53	0.09	0.94	1.15	−0.49	0.04
22	1.30	−0.55	0.09	0.97	1.08	−0.52	0.16	47	2.22	−0.57	0.09	0.98	1.81	−0.56	0.06
23	2.36	−0.38	0.18	0.98	1.76	−0.32	0.06	48	1.38	−0.67	0.11	0.98	1.21	−0.67	0.08
24	1.98	−0.90	0.07	0.94	1.01	−0.93	0.15	49	1.12	−0.39	0.09	0.91	0.81	−0.26	0.04
25	1.17	−0.67	0.12	0.96	0.91	−0.69	0.07	50	0.89	−0.60	0.10	0.70	0.39	0.22	0.05

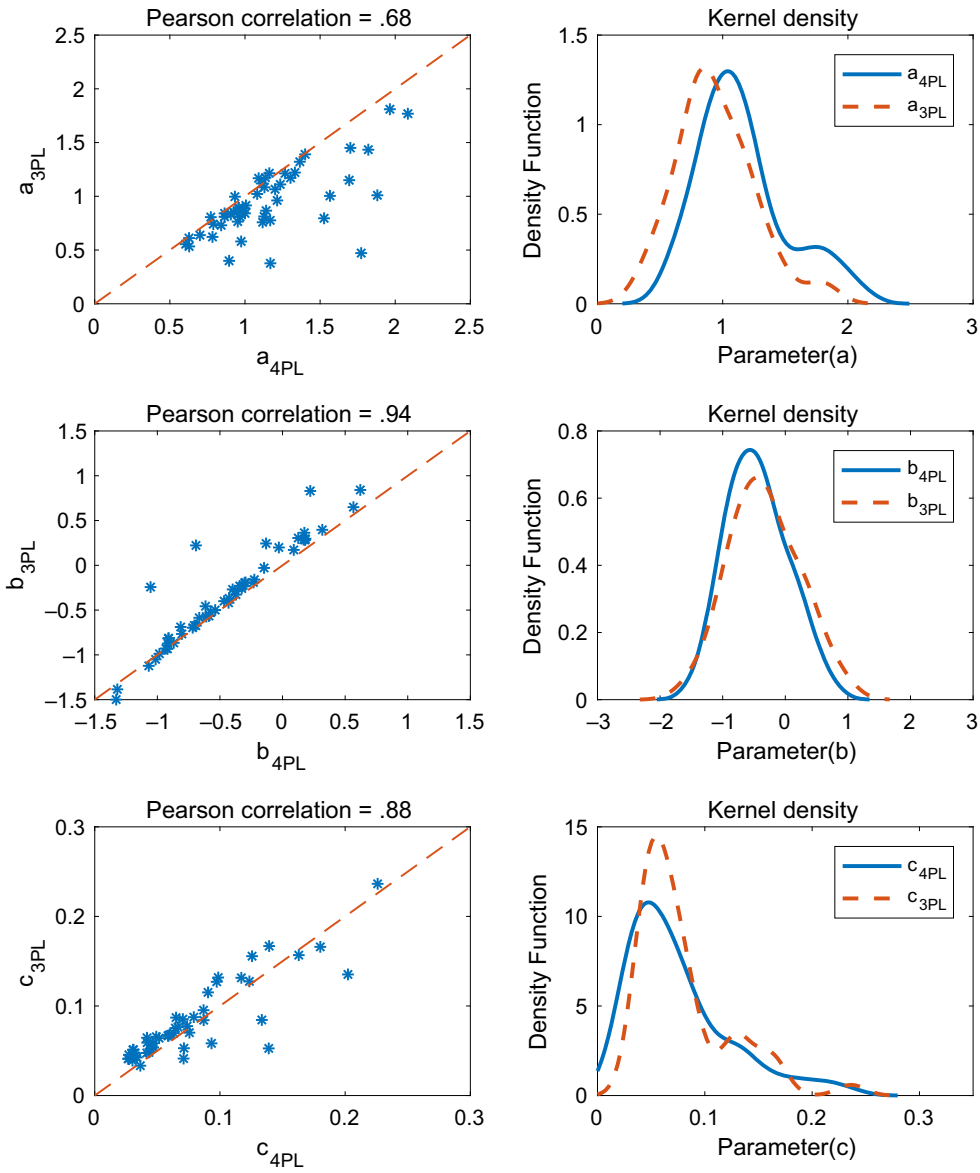

Figure 8. (Left) scatter plots of 3PL item parameter estimates (*a, b, c*) against 4PL estimates. (Right) kernel probability density function curves of *a*, *b*, *c* under the 4PLM and 3PLM.

5.2. Assessing model data fit

Assessing model fit is a routine and important procedure in the item response theory (IRT) domain. IRT models can be implemented effectively for analysing educational and psychological test data only when the model fit is reasonably good. In this study, the fit of the model to data was evaluated at the test and item levels.

At the test level, the chi-square statistic, minus twice the log-likelihood (−2logL) and Akaike's information criterion (AIC; Akaike, 1973) were calculated. The test chi-square statistic is defined as

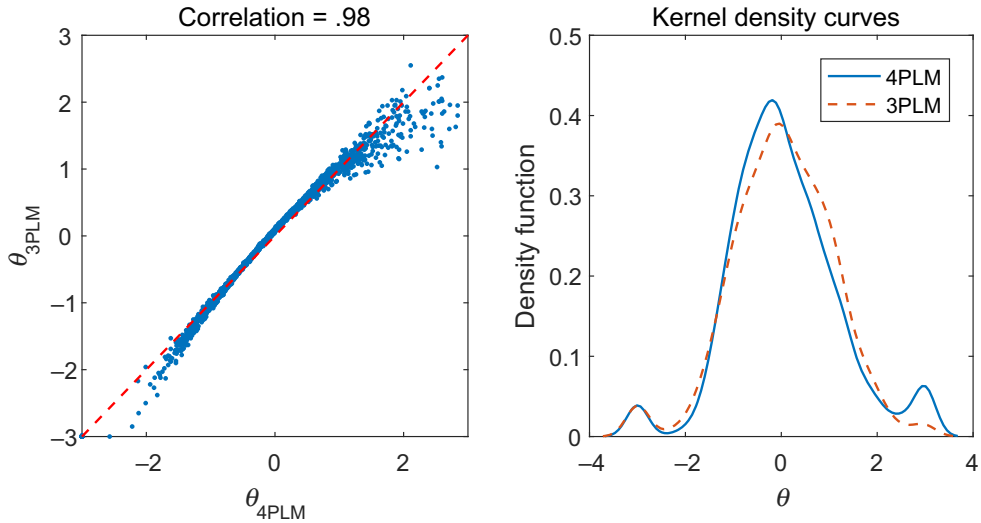

Figure 9. (Left) scatter plot of 4PL estimates of θ against 3PLM estimates. (Right) kernel density function curves of θ estimates in the 4PLM and 3PLM.

$$\chi^2_{\text{test}} = \sum_{b=1}^{H} \frac{(f_{ob} - f_{eb})^2}{f_{eb}},$$

where f_{ob} and f_{eb} is the observed and expected frequency of score b (b = 0, 1, , . . . , 50). The results obtained are displayed in Table 7. It can be seen that the three test model fitting indexes consistently show that the 4PLM fits the data better than the 3PLM.

Moreover, to display the difference between the observed and the model predicted number-correct score distributions, the test fitting plot (Hambleton & Traub, 1973; Swaminathan, Hambleton, & Rogers, 2006) is reported in Figure 10. It can be observed that the differences of the lines between the two models are very small for test takers with test scores up to 40, but when the test scores exceed 40 the fitting frequency curve of the 4PLM is much closer to the observed score distribution than that of the 3PLM. That is, the 4PLM can better describe the data of the high scores by modelling the slipping behaviours.

Following one reviewer's suggestion, we also fitted the 4PLM with several fixed upper asymptotes <1. We calculated the fitting indexes of the 4PLM under fixed parameters d = .98, .95, .90. The results of the model–data fitting assessment are given in the bottom panel of Table 7. All the model indexes consistently show that the fitting of the 4PLM

Table 7. Test model fit indices for the 4PLM, 3PLM, and 4PLM with three constrained upper asymptotes (d = .98, .95, .90)

	χ^2_{item}	−2LogL	AIC
4PLM	99.87	104,631	105,031
3PLM	112.25	104,896	105,196
4PLM −.98	101.20	104,944	105,244
4PLM −.95	103.20	105,124	105,424
4PLM −.90	301.57	105,850	106,150

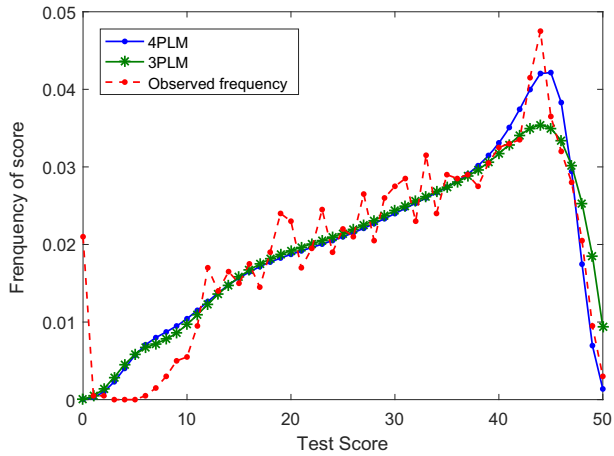

Figure 10. Observed and expected test score distributions based on the 4PLM and 3PLM.

(without specifying *d*) is the best among all the models considered. This suggests that the 4PLM is a better choice in practice than the 4PLM with a fixed upper asymptote.

At the item level, the Pearson chi-square fit statistic (Hambleton & Han, 2005; Hambleton, Swaminathan, & Rogers, 1991; Rogers & Hattie, 1987),

$$\chi^2_{\text{item}} = \sum_{t=1}^{T} N_t \frac{(O_t - E_t)^2}{E_t(1 - E_t)},$$

and the likelihood ratio statistic (McKinley & Mills, 1985; Mislevy & Bock, 1990) provided in BILOG-MG,

$$G^2 = 2\sum_{t=1}^{T} N_t \left(O_t \ln \frac{O_t}{E_t} + (1 - O_t) \ln \frac{1 - O_t}{1 - E_t} \right),$$

were calculated in order to assess the model fit. Here O_t denotes the observed proportion correct in trait interval t, E_t denotes the expected proportion correct in the interval under the given model, N_t is the number of persons in the interval, and T is the number of the trait intervals. In this study, $T = 15$ equal size intervals between -2.5 and 2.5 were chosen and the mean of the probabilities of a correct response was calculated to give the expected values. The results are shown in Table 8. It can be seen that the χ^2_{item} and G^2 values for the 4PLM are smaller than those for the 3PLM for most items, and there are fewer significant χ^2_{item} and G^2 statistics for the 4PLM, indicating that the 4PLM fits the data better than the 3PLM.

To further illustrate, we use a graphical display to examine the discrepancy between observed and expected proportions (Swaminathan *et al.*, 2006). For illustration purposes, the fitting plot of item 5 is displayed in Figure 11. It shows that the upper asymptote of the probability of correct response gets close to .85 rather than approaching 1, as the ability level increases. Hence, the fitting of the 3PLM for this item shows serious deviation, while the 4PLM better captures the response behaviour on this item.

Table 8. Item model fit indices for the 4PLM and 3PLM

	χ^2_{item}		G^2			χ^2_{item}		G^2	
Item	4PL	3PL	4PL	3PL	Item	4PL	3PL	4PL	3PL
1	21.56	23.34[a]	19.51	25.79[a]	26	14.81	6.98	20.17[a]	8.34
2	5.58	11.27	7.97	12.70	27	20.12[a]	16.00	19.96[a]	16.13
3	4.24	6.85	4.39	5.24	28	13.14	10.96	12.38	12.52
4	18.15	28.04[a]	22.11[a]	27.72[a]	29	14.34	13.91	17.99	15.78
5	5.14	48.02[a]	7.81	45.26[a]	30	17.79	9.26	21.07[a]	10.06
6	6.94	6.78	10.19	7.61	31	9.93	13.11	9.25	15.76
7	10.61	15.19	14.44	15.85	32	24.91[a]	13.46	22.22[a]	12.85
8	14.79	10.66	16.65	11.76	33	19.39	18.74	22.06[a]	22.13[a]
9	16.51	21.47[a]	15.69	19.12	34	15.58	13.68	15.06	15.71
10	24.07[a]	24.49[a]	27.83[a]	23.28[a]	35	11.48	25.55[a]	12.43	21.88[a]
11	6.57	7.66	8.98	7.58	36	14.52	28.88[a]	17.21	25.99[a]
12	11.74	9.07	12.20	10.05	37	11.13	14.85	14.14	13.36
13	11.03	14.39	14.03	13.81	38	9.37	16.36	9.80	17.22
14	7.81	15.75	7.54	18.61	39	32.99[a]	17.97	31.28[a]	16.69
15	5.40	9.42	13.17	12.03	40	16.32	14.27	18.04	15.49
16	7.10	12.50	8.01	14.12	41	9.30	17.08	12.78	20.84
17	9.28	9.28	10.38	9.75	42	12.50	22.64[a]	11.42	19.81[a]
18	21.59[a]	31.20[a]	23.35[a]	31.74[a]	43	18.41	7.75	20.29[a]	8.72
19	11.13	11.88	12.09	11.15	44	8.84	8.59	12.59	7.98
20	13.60	12.98	14.62	14.31	45	9.54	15.06	11.00	13.97
21	19.02	24.35[a]	19.01	29.91[a]	46	13.82	29.57[a]	15.33	19.78[a]
22	6.01	8.72	6.88	9.62	47	16.58	8.97	18.56	10.68
23	19.56	19.98	18.25	17.60	48	9.06	9.81	11.30	13.85
24	13.55	60.45[a]	20.77[a]	51.16[a]	49	4.32	10.38	5.87	10.39
25	8.73	6.26	9.77	8.17	50	8.97	21.56[a]	8.78	24.45[a]

[a]Value of χ^2_{item} or G^2 greater than the critical value at the 5% significance level.

6. Discussion

In this paper, we utilize a mixture model representation of the 4PLM and propose an MMAP approach for estimating the 4PLM with an EM algorithm. The mixture modelling revision of the 4PLM not only made the EM algorithm easier to implement but also provided a natural connection with the popular cognitive diagnosis models. Three simulation studies were conducted to investigate the properties of the MMAP/EM estimation under various conditions. The first simulation study was designed to investigate the impacts of prior distributions on the accuracy of the MMAP estimation. The simulation results demonstrated that the accuracy of the MMAP estimators under different prior specifications was almost equivalent when the sample size is as large as $N = 5{,}000$ and 10,000. For a smaller sample size, $N = 1{,}000$, the prior information has a larger impact on the MMAP estimation. Thus uninformative priors are recommended when the sample size is small and accurate prior information can not be obtained beforehand. The aim of the second simulation was to study the influences of the upper asymptote parameter d on the MMAP estimation. The results of this simulation demonstrated that the parameter d displayed substantial impacts on the MMAP estimates of a, b, and c, where extreme values of d led to a decrease in the accuracy of MMAP estimators, but the influences of d on c were weaker. The goal of the third simulation was

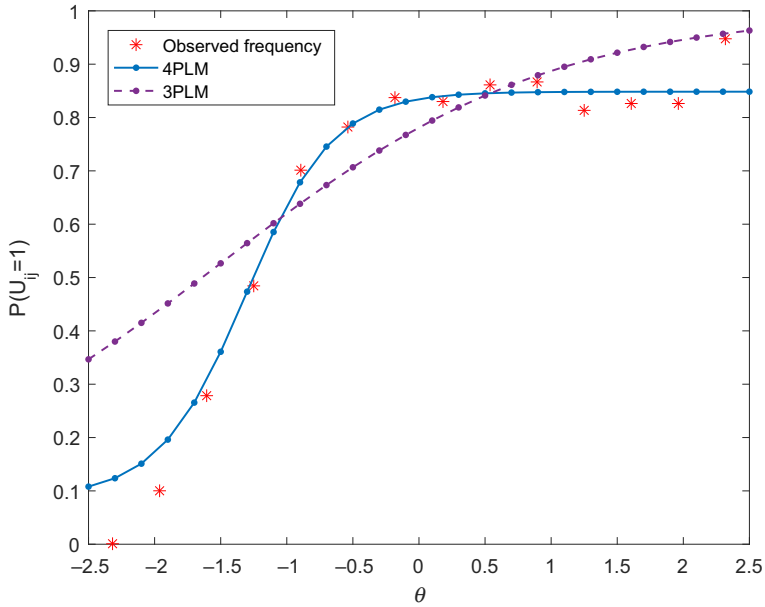

Figure 11. Observed and expected proportion of correct response on item 5 based on the 4PLM and 3PLM.

to compare the performance of the MMAP estimation with the BM estimation in Waller and Feuerstahler (2017). The results suggested that our MMAP estimators are more accurate than the BM estimators across different sample sizes. Finally, a real data from a state reading assessment test was analysed using the 4PLM. The results suggested that the upper asymptote parameter was needed, and in comparison with the 3PLM, the 4PLM can better fit this data set. Additionally, the relationships of the common parameter estimators of the two models (3PLM and 4PLM) were investigated in this empirical study, which further illustrates that the 4PLM outperforms the 3PLM.

There are several issues to be pursued in the future. First, it would be interesting to study MMAP estimation based on a hierarchical prior distribution that jointly models all the item parameters. The more flexible priors would allow the subjective error to be reduced when specifying the prior distributions. On the other hand, this is also likely to increase the computational complexity which may result in a decrease in the accuracy of the parameter estimation. Second, the results of the empirical study demonstrated that scaling the high-ability examinees based on the 4PLM is more accurate than based on the 3PLM. Further study of the estimation performance under different simulation conditions is needed. Furthermore, it would be interesting to study how it would impact test takers' strategies to answer items if the scoring model (such as 4PLM or 3PLM) is known beforehand. This is an important issue in practice and will be studied in the future. Third, the distribution of the ability parameter θ is specified as standard normal in this study, as is common in IRT. However, this assumption is likely to fail in practice, as suggested by the kernel density curves in Figure 9. It would be interesting to apply the joint likelihood estimation approach to estimate item parameters and θ simultaneously, relaxing the normality assumption for θ. On the other hand, it is known in the literature that joint estimation may be dogged by inconsistency issues when the number of items is not large enough. We leave this interesting topic for future study.

Acknowledgements

The authors are greatly indebted to the editor and two anonymous reviewers for their valuable comments and suggestions. This research was supported by the National Natural Science Foundation of China (11571069), the Fundamental Research Funds for the Central Universities, Research Program Funds of the Collaborative Innovation Center of Assessment toward Basic Education Quality (1804047) and National Science Foundation (1659328, 1712717).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akadémiai Kiadó.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model (Technical Report No. 80-20)*. Princeton, NJ: Educational Testing Service.

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–561. https://doi.org/10.1007/BF02296195

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

Culpepper, S. A. (2017). The prevalence and implications of slipping on low-stakes, large-scale assessments. *Journal of Educational and Behavioral Statistics*, *42*, 706–725. https://doi.org/10.3102/1076998617705653

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

Feuerstahler, L. M., & Waller, N. G. (2014). Estimation of the 4-parameter model with marginal maximum likelihood. *Multivariate Behavioral Research*, *49*, 285. https://doi.org/10.1080/00273171.2014.912889

Hambleton, R. K., & Han, N. (2005). Assessing the t of IRT models to educational and psychological test data: Ave step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications*. Washington, DC: Degnon Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hambleton, R. K., & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology*, *24*, 273–281. https://doi.org/10.1111/j.2044-8317.1973.tb00517.x

Liao, W., Ho, R., Yen, Y., & Cheng, H. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality: An International Journal*, *40*, 1679–1694. https://doi.org/10.2224/sbp.2012.40.10.1679

Linacre, J. M. (2004). Discrimination, guessing and carelessness: Estimating IRT parameters with Rasch. *Rasch Measurement Transactions*, *18*, 959–960. https://www.rasch.org/rmt/rmt181b.htm

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*, 509–525. https://doi.org/10.1348/000711009X474502

Lord, F. M. (1975). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (Research Memorandum RB-75-33)*. Princeton, NJ: Educational Testing Service.

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, *37*, 304–315. https://doi.org/10.1177/0146621613475471

McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*, 49–57. https://doi.org/10.1177/014662168500900105

Meng, X. B., Tao, J., & Chen, S. L. (2016). Warm's weighted maximum likelihood estimation of latent trait in the four-parameter logistic model. *Acta Psychologica Sinica*, *48*, 1047–1056. https://doi.org/10.3724/SP.J.1041.2016.01047

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177–195. https://doi.org/10.1007/BF02293979

Mislevy, R. J., & Bock, R. D. (1990). *BILOG: Item analysis and test scoring with binary logistic models [Computer program]*. Chicago, IL: Scientific Software.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32. https://doi.org/10.2307/1914288

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

Peneld, R. D., & Bergeron, J. M. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Applied Psychological Measurement*, *29*, 218–233. https://doi.org/10.1177/0146621604270412

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

Rogers, H., & Hattie, J. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement*, *11*, 47–57. https://doi.org/10.1177/014662168701100103

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment*, *72*, 282–307. https://doi.org/10.1207/S15327752JP720212

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

Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing*, *3*, 365–384. https://doi.org/10.1207/S15327574IJT0304_5

San Martin, E., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*, 183–203. https://doi.org/10.1177/0146621605282773

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2006). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, Volume 26: Psychometrics* (pp. 683–715). Amsterdam, Netherlands: North-Holland.

Tao, J., Shi, N. Z., & Chang, H. H. (2012). Item-weighted likelihood method for ability estimation in tests composed of both dichotomous and polytomous items. *Journal of Educational and Behavioral Statistics*, *37*, 298–315. https://doi.org/10.3102/1076998610393969

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

Von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement: Interdisciplinary Research and Perspective*, *7*, 110–114. https://doi.org/10.1080/15366360903117079

Waller, N. G., & Feuerstahler, S. (2017). Bayesian modal estimation of the four-parameter item response model in real, realistic, and idealized data sets. *Multivariate Behavioral Research*, *52*, 350–370. https://doi.org/10.1080/00273171.2017.1292893

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with nonstandard item response theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality

Inventory. In S. Embretson (Ed.), *Measuring psychological constructs: Advances in model based approaches*. Washington, DC: American Psychological Association.

Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*, 144–168. https://doi.org/10.1111/j.2044-8317.2012.02045.x

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement*, *25*, 317–331. https://doi.org/10.1177/01466210122032163

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. https://doi.org/10.1007/BF02294627

Received 28 November 2017; revised version received 17 May 2019

Appendix A:

The Newton–Raphson interaction for solving equations (20) and (21)

Let $\alpha_j^{(r)}$ and $b_j^{(r)}$ be the current estimates. Then the next estimates are given by

$$\begin{pmatrix} \alpha_j^{(r+1)} \\ b_j^{(r+1)} \end{pmatrix} = \begin{pmatrix} \alpha_j^{(r)} \\ b_j^{(r)} \end{pmatrix} - \begin{pmatrix} L_{\alpha_j\alpha_j}(\alpha_j^{(r)}, b_j^{(r)}) & L_{\alpha_j b_j}(\alpha_j^{(r)}, b_j^{(r)}) \\ L_{\alpha_j b_j}(\alpha_j^{(r)}, b_j^{(r)}) & L_{b_j b_j}(\alpha_j^{(r)}, b_j^{(r)}) \end{pmatrix}^{-1} \begin{pmatrix} L_{\alpha_j}(\alpha_j^{(r)}, b_j^{(r)}) \\ L_{b_j}(\alpha_j^{(r)}, b_j^{(r)}) \end{pmatrix}, \quad \text{(A1)}$$

where

$$L_{\alpha_j}(\alpha_j, b_j) = \frac{\partial E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\xi^t}(\ln p(\boldsymbol{\xi}, \mathbf{z}|\mathbf{u}, \Omega, \boldsymbol{\tau}))}{\partial_{\alpha_j}},$$

$$L_{b_j}(\alpha_j, b_j) = \frac{\partial E_{\mathbf{W},\boldsymbol{\theta}|\mathbf{u},\xi^t}(\ln p(\boldsymbol{\xi}, \mathbf{z}|\mathbf{u}, \Omega, \boldsymbol{\tau}))}{\partial_{b_j}},$$

are given in equations (20) and (21), and

$$L_{\alpha_j\alpha_j}(\alpha_j, b_j) = \frac{\partial L_{\alpha_j}(\alpha_j, b_j)}{\partial \alpha_j} = -e^{2\alpha_j} \sum_{i=k}^{K} \left[\hat{R}(x_k)(x_k - b_j)^2 p_j^*(x_k)(1 - p_j^*(x_k))\right] - \frac{1}{\sigma_\alpha}, \quad \text{(A2)}$$

$$L_{b_j b_j}(\alpha_j, b_j) = \frac{\partial L_{b_j}(\alpha_j, b_j)}{\partial b_j} = -e^{2\alpha_j} \sum_{i=k}^{K} \left[\hat{R}(x_k) p_j^*(x_k)(1 - p_j^*(x_k))\right] - \frac{1}{\sigma_b}, \quad \text{(A3)}$$

$$\begin{aligned} L_{b_j\alpha_j}(\alpha_j, b_j) &= L_{\alpha_j b_j}(\alpha_j, b_j) \\ &= \tfrac{\partial L_{\alpha_j}(\alpha_j, b_j)}{\partial b_j} = e^{2\alpha_j} \sum_{i=k}^{K} \left[\hat{R}(x_k)(x_k - b_j) p_j^*(x_k)(1 - p_j^*(x_k))\right], \end{aligned} \quad \text{(A4)}$$

where $p_j^*(\cdot)$ is defined in (3).

Appendix B:

Matlab code for MMAP\EM for 4PLM

```
function [Ra, Rb, Rc, Rd]=MAEM(u, n, priora, priorb, priorc, priord)
%   u: is the response matrix
%   priora: is the prior of a
%   priorb: is the prior of b
%   priorc: is prior of c
%   priord: is prior of d
%   M: is the number of test takers
%   N: is the number of items
%   ntime: is number of the Fisher-Scoring iteration
%   NTIME: is number of the EM algorithm
%   a0: is initial value of a parameter
%   b0: is initial value of b parameter
%   c0: is initial value of c parameter
%   d0: is initial value of d parameter
%   Note: The initial values should are specified by yourself
%   n: is the number of the quadrature points
%   x: is quadrature points
indice=1;
INDICE=1;
ntime=0;
NTIME=0;
[M,N]=size(u);
x=linspace(-4,4,n);
x1=x';
d=x1(2)-x1(1);
Ak=normpdf(x1,0,1)*d;
% intial value a b c d
r0=identify(u);
a0=r0./sqrt(1-r0.^2);
a0=log(a0);
b0=sum(u)./M;
b0=-norminv(b0,0,1)./r0;
c0=0*a0+0.25;
d0=0*a0+0.75;
delta=0.1;
%Note:The intial values can be given by yourself.
P=@(a,b,c,d,x)c+(d-c)./(1+exp(-a.*(x-b)));
% ------------------------------------------
MK=ones(M,1);
a1=MK*a0;
b1=MK*b0;
c1=MK*c0;
d1=MK*d0;
amu=priora(1);
asigma2=priora(2);
bmu=priora(1);
bsigma2=priora(2);
Niteration=100;
for k=1:n
```

```
    p=P(exp(a1),b1,c1,d1,x(k));
    L=p.^u.*((1-p).^(1-u));
    LL(:,k)=prod(L,2)*Ak(k);
end
LL0=sum(LL,2);
LH=sum(log(LL0));
% E-step and M-step
NK=ones(1,n);
nn=ones(1,n);
while INDICE==1 && NTIME<Niteration
    LL1=LL0*nn;
    h=LL./LL1;
    f=sum(h);
    for i=1:N
        U=(u(:,i))*NK;
        p=MK*P(exp(a0(i)),b0(i),c0(i),d0(i),x);
        ppp=(p-c0(i))/(d0(i)-c0(i));
        pz=(d0(i)*ppp./p).*U+((1-d0(i))*ppp./(1-p)).*(1-U);
        PZ=pz.*h;
        r=sum(PZ);
        c0(i)=(sum(u(:,i).*(1-sum(PZ,2)))+priorc(1)-1)/(sum(1-sum(PZ,2))+
        sum(priorc)-2);
        S=(sum(sum(PZ,2).*u(:,i))+priord(1)-1)/(sum(sum(PZ,2))+sum(priord
        )-2);
        if S>c0(i)
            d0(i)=S;
        else
            d0(i)=c0(i)+delta;
        end
        at=a0(i);
        bt=b0(i);
        while indice==1 && ntime<50
            Pi=P(exp(at),bt,0,1,x);
            w=Pi.*(1-Pi);
            la1=exp(at)*sum((x-bt).*(r-f.*Pi))-(at-amu)/asigma2;
            lb1=-exp(at)*sum(r-f.*Pi)-(bt-bmu)/bsigma2;
            laa=-exp(2*at)*sum((f.*(x-bt).^2.*w))-1/asigma2;
            lbb=-exp(2*at)*sum((f.*w))-1/bsigma2;
            lab=exp(2*at)*sum((x-bt).*f.*w) ;
            res=[at;bt]-[laa,lab;lab,lbb]^(-1)*[la1;lb1];
            at1=res(1);
            bt1=res(2);
            if norm([at1-at;bt1-bt],2)<10^(-3)
                indice=0;
            else
                at=at1;
                bt=bt1;
                ntime=ntime+1;
            end
        end
```

```
            ntime=1;
            indice=1;
            a0(i)=at1;
            b0(i)=bt1;
        end
        a1=MK*a0;
        b1=MK*b0;
        c1=MK*c0;
        d1=MK*d0;
        for k=1:n
            p=P(exp(a1),b1,c1,d1,x(k));
            L=p.^u.*((1-p).^(1-u));
            LL(:,k)=prod(L,2)*Ak(k);
        end
        LL0=sum(LL,2);
        LH1=sum(log(LL0));
        if abs(LH-LH1)<10^(-3)
            INDICE=0;
        else
            NTIME=NTIME+1;
            LH=LH1;
        end
        RL(NTIME)=LH1;
    end
    % --------Final Results--------
    Ra=exp(a0);
    % If considering scaling constant D=1.7, Ra=Ra/D.
    Rb=b0;
    Rc=c0;
    Rd=d0;
```